

**“Like, comment, subscribe” –
Perception of acoustic-prosodic features of
content creators’ charismatic speech on YouTube**

Dissertation in fulfilment of the requirements for
the doctoral degree
at the Faculty of Arts and Humanities
at Kiel University
in Kiel

submitted by
Stephanie Berger

Kiel
25 January 2024

First examiner: Prof. Dr. Margaret Zellers
Second examiner: Prof. Dr. Oliver Niebuhr

Date of the oral examination: 3 May 2024

Approved for printing by the second Vice Dean,
Prof. Dr. Elmar Eggert: 22 August 2024

Data Availability Statement

The data that support the findings of this study are openly available in *opendata@uni-kiel* – Kiel University Research Data Repository with the DOI 10.57892/100-76. Access to TextGrid files or participant metadata for research purposes can be requested from the author.

Open Access

The work is published under the Creative Commons license Attribution 4.0 International (CC-BY 4.0). The deed can be found here: <https://creativecommons.org/licenses/by/4.0/deed.en>.

The electronic version of this doctoral thesis is available on the open access publications server MACAU of the University Library Kiel (<https://macau.uni-kiel.de>) frei verfügbar:

DOI: 10.38071/2024-00858-9

URN: urn:nbn:de:gbv:8:3-2024-00858-9

© Stephanie Berger, 2024

(ORCID: <https://orcid.org/0009-0002-3095-6188>)

Abstract

This dissertation investigates the influence of different acoustic-prosodic features on the perception of charismatic speech of YouTubers. YouTubers are people who produce videos on the platform YouTube and star in them. All YouTubers in the sample of this project have made this into their career. The sample in this dissertation includes ten YouTubers (five male, five female; five from England, five from North America).

The concept charisma is understood in this project as a collection of different attributes. The charisma-adjacent attributes that are relevant for this dissertation are *authentic*, *enthusiastic*, *likable*, and *persuasive*. Charisma is additionally seen as a gradual ability that every person has to varying extent, so that no person is considered uncharismatic, but at most as less charismatic than others (Niebuhr et al., 2016a). Additionally, a charismatic impression of a person is created by the combination of message content and delivery (see Holladay and Coombs, 1993; Awamleh and Gardner, 1999), though the delivery often has a larger influence on charisma ratings (Caspi et al., 2019). Finally, it is important to note that charisma as defined for this project does not involve formal authority of the speaker, but that a charismatic effect is created by similar ideals, world views, and values of speakers and addressees (see Antonakis et al., 2016; Tur et al., 2022).

Charismatic speech is researched in phonetics, and linguistics, to better understand the aspect of the voice in charisma. Most research focuses on speakers from politics and business. Similarly, there is little research on female speakers, and (in regards to English as the investigated language) British English. Research has already identified many acoustic-prosodic feature characteristics that can lead to a charismatic impression of speakers. These feature characteristics are—among others—a higher pitch level for male and a lower pitch level for female speakers, a wider pitch range, more and later high pitch accents, a regular rhythm and more frequently occurring strongly prominent syllables (emphatic accents) as well as a higher tempo (see, e.g., Rosenberg and Hirschberg, 2005, 2009; Biadsky et al., 2007; Signorello et al., 2012a, 2012b; Novák-Tót, 2016; Berger et al., 2017; Niebuhr et al., 2018a; Niebuhr et al., 2020a). Other aspects like content, duration of presented material, speaker gender, origin, familiarity of listeners with the speakers, or audio compression can additionally influence the perception of charismatic speech (see, e.g., Rosenberg and Hirschberg, 2005, 2009; Biadsky et al., 2007; Signorello et al.,

2012a, 2012b; D’Errico et al., 2013; Jokisch et al., 2018; Gutnyk et al., 2019; Niebuhr et al., 2019; Siegert and Niebuhr, 2021a, 2021b).

This dissertation investigates YouTubers as a new speaker group in the area of charismatic speech research. YouTubers can belong to several categories simultaneously, such as entertainment and business, since they entertain an audience with their videos, but are at the same time the face of a brand (their channels) (see Kyncl and Peyvan, 2017). That leads to a specific speaking style that is called “bouncy” in the media (Beck, 2015). The main features of this style (referred to as “YouTube voice”) can be traced back to different strategies of emphasis, for example many strongly prominent syllables (especially in terms of long vowels and consonants), overly clear articulations (hyperarticulations), or a variation of utterances with fast and slow tempo (see, e.g., Beck, 2015).

For this dissertation, roughly 50 minutes of speech material were annotated, which will additionally find use in future studies. This project mainly focuses on perception experiments to investigate how different acoustic features and their characteristics influence the perception of charisma directly, and the four charisma-adjacent attributes mentioned above. Audio sections (stimuli) were digitally modified and presented to experiment participants.

For the first experiment, individual phrases (one per speaker, which was also left unmodified as a stimulus) were digitally modified using the program Praat (Boersma and Weenink, 2018) by changing the pitch level, pitch range, final contour direction, and speech rate. In the second experiment, longer utterances (per speaker: four phrases with three pauses in between) were modified in terms of pause duration and presence of audible breathing noise. The stimuli of both experiments were rated on five Likert scales (*charismatic, authentic, enthusiastic, likable, persuasive*). Along with the *charismatic* ratings, the participants also indicated how familiar they were with the speakers. The third experiment of this dissertation is based on the ratings of the unmodified stimuli of the previous two studies. These ratings were then correlated with acoustic measurements of the stimuli.

The main influences on perceived charismatic speech in the present sample of YouTubers seem to be more frequently occurring prominent syllables, shorter pauses with audible breathing noises, and fast speech rate in longer phrases. These are similar features to other already investigated speech styles such as political speeches or keynote speeches in the business context (see Novák-Tót, 2016; Novák-Tót et al., 2017; Niebuhr et al., 2020a). However, the feature characteristics seem to be more pronounced in the YouTube sample: pauses are even shorter, emphatic accents are even more frequent, and phrases are longer with simultaneously even higher speech rate. Another finding of this dissertation is that the more familiar a speaker was rated, the higher the charisma rating. This was mainly the case, though, when participants indicated that the speaker seemed familiar to them.

When they indicated that they knew the speaker, the charisma rating was mostly as low as for speakers that were identified as unknown by the listeners.

Ultimately, this dissertation offers first indications on charismatic speech in a different speaking style. That also opens up new directions for further research. The more speech styles and situations are investigated in research, the more conclusive and overarching will the understanding of the concept “charisma”, and what it might mean in different contexts, become.

Kurzzusammenfassung

Diese Dissertation beschäftigt sich mit dem Einfluss verschiedener akustisch-prosodischer Merkmale auf die Wahrnehmung von charismatischem Sprechen bei YouTubern. YouTuber sind Menschen, die auf der Plattform YouTube Videos produzieren und in ihnen auftreten. Alle YouTuber:innen in der Stichprobe dieses Projekts haben aus dieser Beschäftigung eine Karriere gemacht haben. Die Stichprobe in dieser Dissertation beinhaltet zehn YouTuber:innen (fünf männlich, fünf weiblich; fünf aus England, fünf aus Nordamerika).

Das Konzept Charisma wird in dieser Arbeit als eine Sammlung verschiedener Attribute angesehen. Die in der Arbeit relevanten, Charisma-nahen Attribute sind authentisch, enthusiastisch, nett und überzeugend. Charisma wird außerdem als eine graduelle Fähigkeit verstanden, die jeder Mensch in verschiedener Stärke besitzt, sodass kein Mensch als uncharismatisch zu bezeichnen ist, sondern höchstens als weniger charismatisch (siehe Niebuhr et al., 2016a). Zusätzlich entsteht ein charismatischer Eindruck einer Person immer aus einer Kombination aus Inhalt und Vortragsweise (siehe Holladay und Coombs, 1993; Awamleh und Gardner, 1999), wobei die Vortragsweise meist einen größeren Einfluss auf Bewertungen hat (Caspi et al., 2019). Zuletzt ist für Charisma bezogen auf YouTuber besonders wichtig, dass Charisma in der Definition dieser Arbeit keine formale Autorität eines Sprechers oder einer Sprecherin erfordert, sondern dass ein charismatischer Effekt durch ähnliche Ideale, Weltansichten und Werte von Sprecher:innen und Angesprochenen entsteht (siehe Antonakis et al., 2016; Tur et al., 2022).

Charismatisches Sprechen wird in der Phonetik, und Linguistik, untersucht, um den Aspekt der Stimme bei Charisma besser zu verstehen. Die meiste Forschung widmet sich dabei Sprecher:innen aus Politik und Wirtschaft. Ebenfalls ist wenig Forschung über Sprecherinnen vorhanden, sowie (bezogen auf Englisch als untersuchte Sprache) Britisches Englisch. Bislang hat die Forschung eine Vielzahl an akustisch-prosodischen Merkmalen identifiziert, die zu einem charismatischen Eindruck bei Sprecher:innen führen können. Dazu zählen—unter anderem—eine höhere Tonlage (*pitch level*) für männliche Sprecher und eine niedrigere für weibliche Sprecherinnen, ein größerer Tonhöhenumfang, häufigere und spätere hohe Tonakzente, ein regelmäßiger Rhythmus und häufigere stark hervorstechende Silben (emphatische Akzente) sowie ein schnelleres Tempo (siehe, u.a., Rosenberg und Hirschberg, 2005, 2009; Biadys et al., 2007; Signorello et al., 2012a,

2012b; Novák-Tót, 2016; Berger et al., 2017; Niebuhr et al., 2018a; Niebuhr et al., 2020a). Zusätzlich können Aspekte wie Inhalt, Länge des präsentierten Materials, Geschlecht der Sprecher:innen, Herkunft, Kenntnis der Sprecher:innen, oder Audiokompression die Wahrnehmung von charismatischem Sprechen beeinflussen (siehe, u.a., Rosenberg und Hirschberg, 2005, 2009; Biadys et al., 2007; Signorello et al., 2012a, 2012b; D’Errico et al., 2013; Jokisch et al., 2018; Gutnyk et al., 2019; Niebuhr et al., 2019; Siegert und Niebuhr, 2021a, 2021b).

Diese Arbeit beschäftigt sich mit YouTubern als neue Sprechergruppe in der Forschung über charismatisches Sprechen. YouTuber sind gleichzeitig in Entertainment und Wirtschaft zu verordnen, da sie mit ihren Videos ein Publikum unterhalten, gleichzeitig aber auch das Gesicht einer Marke (ihres Kanals) sind (siehe Kyncl und Peyvan, 2017). Das führt zu einem besonderen Sprechstil, der in den Medien unter anderem als “hüpfend” bezeichnet wird und deren Hauptmerkmale auf verschiedene Strategien der Emphase (oder Hervorhebung) zurückgeführt werden, wie zum Beispiel viele starke Betonungen in Bezug auf lange Vokale oder Konsonanten, überdeutliche Lautproduktionen (Hyperartikulationen) oder eine Abwechslung von Äußerungen mit schnellem und langsamen Tempo (siehe Beck, 2015).

Es wurden 50 Minuten Sprachmaterial aufgearbeitet und annotiert, welches in späteren Studien weitere Verwendung findet. Vorrangig untersucht dieses Projekt allerdings anhand von Perzeptionsexperimenten, inwieweit verschiedene akustische Merkmale und deren Ausprägungen einen Einfluss auf die Wahrnehmung von Charisma, sowie auf die vier oben genannten anlehrenden Attribute haben. Dazu wurden Audioausschnitte (Stimuli) digital modifiziert und Experimentteilnehmenden präsentiert.

Für das erste Experiment wurden einzelne Phrasen (eine pro Sprecher:in) mit dem Programm Praat (Boersma und Weenink, 2018) digital modifiziert indem (zusätzlich zur unmodifizierten Phrase pro Sprecher:in) diese Äußerungen in ihrem Tonhöhenlevel, Tonhöhenumfang, Richtung der finalen Kontur und der Sprechgeschwindigkeit verändert wurden. Im zweiten Experiment wurden längere Äußerungen (pro Sprecher:in vier Phrasen mit drei Pausen dazwischen) in Pausendauer und Auftreten von hörbaren Atemgeräuschen modifiziert. Bewertet wurden die Stimuli beider Experimente auf fünf Likert-Skalen (*charismatisch, authentisch, enthusiastisch, nett, überzeugend*; Englisch: *charismatic, authentic, enthusiastic, likable, persuasive*). In Verbindung mit der Charisma-Bewertung wurde gleichzeitig abgefragt, wie bekannt die Hörer:innen mit den Sprecher:innen waren. Das dritte Experiment dieser Arbeit basiert auf den Bewertungen der unmodifizierten Stimuli der vorigen beiden Studien. Diese Bewertungen wurden dann mit akustischen Messungen der Stimuli korreliert.

Die Haupteinflüsse auf charismatisches Sprechen in dieser Stichprobe von

YouTubern scheinen häufigere betonte Silben, kürzere Pausen mit hörbaren Atemgeräuschen und schnellerer Sprechgeschwindigkeit in längeren Phrasen zu sein. Dies sind ähnliche Merkmale zu anderen bereits untersuchten Sprechstilen wie politischen Reden oder Reden im wirtschaftlichen Bereich (z.B. Novák-Tót, 2016; Novák-Tót et al., 2017; Niebuhr et al., 2020a). Allerdings scheinen die Merkmale in der YouTube-Stichprobe stärker ausgeprägt zu sein: Pausen sind noch kürzer, emphatische Akzente sind noch häufiger und Phrasen sind länger mit gleichzeitig höherer Sprechgeschwindigkeit. Eine weitere Erkenntnis dieser Arbeit ist, dass je bekannter ein Sprecher oder eine Sprecherin eingeschätzt wurde, desto höher fiel die Charisma-Bewertung aus. Dies ist allerdings vor allem der Fall wenn Experimentteilnehmende angegeben haben, dass der Sprecher oder die Sprecherin ihnen bekannt vorkam. Wenn sie angaben, den Sprecher oder die Sprecherin zu kennen, war die Charisma-Bewertung meist genauso niedrig wie bei nicht bekannten Sprecher:innen.

Diese Arbeit bietet somit erste Hinweise auf charismatisches Sprechen in einem weiteren Sprechstil. Damit werden ebenfalls neue Richtungen für die weitere Forschung aufgezeigt. Je mehr Sprechstile und Situationen in der Forschung untersucht werden, desto deutlicher wird das Verständnis des Konzepts "Charisma" und was es in verschiedenen Kontexten bedeuten kann.

Acknowledgments

This dissertation has accompanied me for the past 2,934,600 minutes (or roughly 5.5 years), and along the way has given me a lot of joy and fun and new experiences, but also resilience to power through some difficult times. Many people have supported me on this journey.

Thank you first and foremost to my supervisors, Meg Zellers and Oliver Niebuhr, for their tremendous help, support, and mentorship. Both were always there to help with my questions, if in the office, on Zoom, or on the phone. I am extremely grateful for all the time and work you invested in me and my project. Traveling together to conferences and meeting colleagues through their connections are equally important to me and offered me amazing experiences abroad—if that is Boston before my PhD work, or Graz, Sønderborg or Prague. Thank you to Meg in particular for the discussions and conversations to put the whole PhD process in perspective, especially when I was having doubts, and the constant feedback to keep growing. It's also fun to know that we got to learn new things together and from each other. You are a great role model of what being a woman in research means, and I am honored to be your first PhD supervision. Thank you to Olli in particular for always motivating me to continue doing research, since early on in my Bachelor studies in Kiel. I started on this journey after you presented phonetics to potential new students, and I probably would not be at this point in my life or career if that, and your continued supervisions, joined work, and conference organizations, had not happened. Both of you made the work on the dissertation and the other joined projects fun by working together at eye level, which made some difficult situations (pandemic..., among others) more bearable.

The first three years of this project were supported financially by the Federal State Grant Schleswig-Holstein, awarded by the graduate center at Kiel University.

Thank you to the phonetics team at Kiel University for their discussions and support: Marlene Böttcher, Martina Rossi, Kathrin Feindt, Benno Peters, and Tina John. The same goes for the others in the department, in particular John Peterson, Eugenie Stapert, and Lee Pratchett, for the discussions and the many new research directions we collected in colloquium talks. I will not get bored... The linguistic data café also helped immensely, in particular Beke Hansen and Robert Dausg, when I was knee-deep in my statistical analyses and needed additional direction.

One of the most special aspects of my PhD journey is the connection we created

among the PhD students (by now present, former, and guests) at the department. The support through the years in our discussions in our PhD meetings and at any point in time were incredibly meaningful to me. So thank you to Ariba Khan, Csilla Kasz, Erika Just, Judith Voß, Tobias Weber, and Anneliese Kelterer (plus Marlene and Martina, you two get two mentions, you deserve it).

There are also some Kiel-external people to thank. Learning and using the DIMA annotation system was an important step in the creation of the thesis. Thank you to Christine Röhr, Jane Mertens and Janina Kalbertodt from Cologne University for creating a DIMA consensus annotation for me in the beginning of my annotation work so I could have a reference to learn from for the rest of the annotation. Furthermore, thank you to Stefan Baumann (and Christine) for the long, helpful, and thought-provoking discussions about DIMA in general, and Frank Kügler for insights into interrater agreement. I also received much help when setting up and carrying out my experiments remotely, specifically from the University of York between 2020 and 2021. Thank you in particular to Dominic Watt for discussions and advertising the experiment to students, and similarly thank you to Richard Ogden, Eleanor Chodroff, Shane Slogget, and Eytan Zweig for organizational help, and Sarah Lapacz for the discussions and help in finding more participants. Thank you also to Salina Cuddy for the insights into the use and set-up of Prolific. Furthermore, Jana Neitsch also always helped motivate me to continue in our numerous (research and non-research related conversations), and I had so much fun at conferences and (sometimes stressful) conference organizations with you.

Another thank you goes to my friend group that exists outside my academic bubble, starting with Jenny Voß (we shall be conquering more galaxies soon) for always being there and sending funny (and sometimes annoying) videos and jokes when I need them. The same goes to Nele Kiupel, thanks for always being there when needed, and for reading this whole thing—your comments and the discussions we had were incredibly motivating. Also thanks to the others in our group—Megan, Josy, Mimi, Regina, and Sonja—for the years of friendship and me knowing I have people in my corner.

Last, but definitely not least, I would like to thank my parents—Birgit and Thomas Berger—for allowing me to be who I am, do what I do, and being my biggest supporters through everything. Despite having so much fun with this project and work, the past few years have been incredibly difficult at times, and you were always there to cheer me on and cheer me up. Thank you!

Tho' the story never ends.

Jonathan Larson (RENT)

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| I | Background | 5 |
| 2 | Charisma and its many definitions | 7 |
| 2.1 | Traditional understandings of charisma | 7 |
| 2.1.1 | From the original understanding to Max Weber | 7 |
| 2.1.2 | Popularization of charisma and the media influence | 8 |
| 2.1.3 | Contemporary understandings of charisma | 8 |
| 2.2 | Charismatic leadership | 9 |
| 2.3 | A new approach: charisma as a signal | 11 |
| 2.3.1 | Problems with previous charisma definitions | 11 |
| 2.3.2 | Can charisma be taught? | 12 |
| 2.3.3 | Charisma as leadership signaling | 12 |
| 2.3.4 | Charisma signaling and informal leadership | 14 |
| 2.4 | Charisma and delivery | 15 |
| 2.5 | Summary: Charisma in this project | 16 |
| 3 | Charismatic speech: Phonetic and non-phonetic influences | 19 |
| 3.1 | Moving from rhetorics to phonetics | 19 |
| 3.2 | Charismatic speech: Acoustics and prosody | 20 |
| 3.2.1 | Pitch | 20 |
| 3.2.2 | Intonation | 24 |
| 3.2.3 | Rhythm and phrasing | 26 |
| 3.2.4 | Tempo | 29 |
| 3.2.5 | Phonation quality and loudness | 31 |
| 3.2.6 | Articulation space | 33 |
| 3.2.7 | Additional findings and applications | 35 |
| 3.3 | Non-phonetic influences on charismatic speech | 36 |
| 3.3.1 | Content | 36 |
| 3.3.2 | Duration of presented material | 37 |
| 3.3.3 | Speaker gender | 38 |
| 3.3.4 | Culture | 39 |
| 3.3.5 | Familiarity | 40 |
| 3.3.6 | Audio compression | 41 |
| 3.4 | Summary and observations | 41 |
| 4 | YouTube: Lights, camera, business | 45 |
| 4.1 | A history of YouTube | 45 |
| 4.1.1 | Beginnings and growth of a platform | 45 |
| 4.1.2 | Monetization and business on YouTube | 46 |
| 4.1.3 | Video style: vlog | 47 |

| | | |
|------------|--|-----------|
| 4.1.4 | Becoming successful on YouTube | 48 |
| 4.1.5 | YouTube statistics and audience demographics | 49 |
| 4.2 | Youtubers as leaders: Charisma online | 50 |
| 4.2.1 | YouTube and authority | 51 |
| 4.2.2 | YouTube, authenticity, and intimacy | 51 |
| 4.2.3 | YouTube and enthusiasm | 52 |
| 4.2.4 | YouTube and charisma | 53 |
| 4.3 | YouTube voice | 54 |
| 4.3.1 | The features of “YouTube voice” | 55 |
| 4.3.2 | Comparisons with other speaker groups | 56 |
| 4.3.3 | Influence of video type on YouTube voice | 57 |
| 4.3.4 | Vocal effort and charisma (on YouTube) | 58 |
| 4.4 | Summary | 60 |
| 5 | Research questions and hypotheses | 61 |
| II | General Methodology | 65 |
| 6 | Data selection | 67 |
| 6.1 | Speaker demographics | 67 |
| 6.2 | Speaker selection | 68 |
| 6.3 | Video selection | 70 |
| 6.3.1 | Video overview | 70 |
| 6.3.2 | Inclusion criteria | 71 |
| 6.3.3 | Exclusion criteria | 72 |
| 7 | Data treatment | 75 |
| 7.1 | Data preparation | 75 |
| 7.2 | Annotation for acoustic measurements | 75 |
| 7.3 | DIMA annotation for English data | 78 |
| 7.4 | Measurements and scripts | 81 |
| III | Studies | 83 |
| 8 | Perception I: Prosodic manipulations | 85 |
| 8.1 | Introduction | 85 |
| 8.2 | Hypotheses | 85 |
| 8.3 | Methods | 88 |
| 8.3.1 | Stimulus selection and creation | 88 |
| 8.3.2 | Stimulus measurements | 92 |
| 8.3.3 | Experiment procedure | 93 |
| 8.3.4 | Experiment design | 94 |
| 8.3.5 | Participant recruitment and demographics | 97 |
| 8.3.6 | Statistical analyses | 100 |
| 8.4 | Results I: Charisma-adjacent attribute ratings | 103 |
| 8.4.1 | Pitch level | 103 |
| 8.4.2 | Pitch range | 106 |
| 8.4.3 | Final contour direction | 109 |

| | | |
|-----------|---|------------|
| 8.4.4 | Speech rate | 113 |
| 8.5 | Results II: Direct charisma ratings and familiarity | 116 |
| 8.5.1 | Pitch level | 117 |
| 8.5.2 | Pitch range | 120 |
| 8.5.3 | Final contour direction | 122 |
| 8.5.4 | Speech rate | 125 |
| 8.6 | General discussion | 127 |
| 8.6.1 | Summary of the experiment results | 127 |
| 8.6.2 | Manipulation effects on direct charisma ratings | 129 |
| 8.6.3 | Manipulation effects on attribute ratings | 130 |
| 8.6.4 | Speaker gender effects | 133 |
| 8.6.5 | Speaker origin effects | 136 |
| 8.6.6 | Familiarity effects | 138 |
| 9 | Perception 2: Pauses, breathing & cuts | 141 |
| 9.1 | Introduction | 141 |
| 9.2 | Hypotheses | 143 |
| 9.3 | Method | 145 |
| 9.3.1 | Stimulus creation: Pause manipulations | 145 |
| 9.3.2 | Stimulus measurements | 149 |
| 9.3.3 | Experiment design and participants | 149 |
| 9.3.4 | Statistical analyses | 150 |
| 9.4 | Results I: Charisma-adjacent attribute ratings | 151 |
| 9.4.1 | Pause duration | 151 |
| 9.4.2 | Presence of breathing noises | 155 |
| 9.5 | Results II: Direct charisma ratings and familiarity | 162 |
| 9.5.1 | Pause duration | 162 |
| 9.5.2 | Presence of breathing noises | 165 |
| 9.6 | Discussion | 167 |
| 9.6.1 | Summary of the experiment results | 167 |
| 9.6.2 | Manipulation effects on direct charisma ratings | 171 |
| 9.6.3 | Pause duration effects on attribute ratings | 172 |
| 9.6.4 | Breathing noise effects on attribute ratings | 175 |
| 9.6.5 | Speaker gender effects | 176 |
| 9.6.6 | Speaker origin effects | 178 |
| 9.6.7 | Familiarity effects | 180 |
| 10 | Perception 3: Acoustics and perception | 183 |
| 10.1 | Introduction | 183 |
| 10.2 | Hypotheses | 184 |
| 10.3 | Methods | 188 |
| 10.3.1 | Measurements | 188 |
| 10.3.2 | Experiment design | 191 |
| 10.3.3 | Statistical analyses | 191 |
| 10.4 | Results | 193 |
| 10.4.1 | Pitch | 193 |
| 10.4.2 | Intonation | 199 |
| 10.4.3 | Duration and tempo | 205 |
| 10.5 | Discussion | 210 |
| 10.5.1 | Summary of the results | 211 |

| | | |
|-----------|--|---------------|
| 10.5.2 | Interpretation of the pitch-related findings | 212 |
| 10.5.3 | Interpretation of the intonation-related findings | 216 |
| 10.5.4 | Interpretation of the tempo-related findings | 221 |
| 10.5.5 | Observations regarding gender, origin, and speaking style . . | 224 |
| IV | Discussion | 227 |
| 11 | Discussion | 229 |
| 11.1 | Findings and implications | 229 |
| 11.1.1 | Acoustic feature configuration for charisma perception | 229 |
| 11.1.2 | Ratings and acoustic features | 232 |
| 11.1.3 | Familiarity and charisma | 238 |
| 11.2 | Development of remote perception experiment method | 239 |
| 11.3 | Limitations | 241 |
| 11.3.1 | Data treatment and measurements | 241 |
| 11.3.2 | Experiment methodology | 244 |
| 11.3.3 | Statistical methodology | 248 |
| 11.4 | Future research | 248 |
| 12 | Conclusions | 253 |
| V | Appendix | 271 |
| A | YouTube statistics | i |
| B | Measurements of short stimuli | iii |
| C | Experiment file (example) | v |
| D | Experiment instructions | xix |
| E | Linear mixed model codes (Perception 1) | xxiii |
| F | Outputs: Linear mixed models (Perception 1) | xxvii |
| G | Output: Estimated marginal means (Perception 1) | xxxi |
| H | Measurements of long stimuli | xxxiii |
| I | Linear mixed model codes (Perception 2) | xxxv |
| J | Outputs: Linear mixed models (Perception 2) | xxxvii |
| K | Additional figures (Perception 3) | xli |

List of Tables

| | | |
|------|---|-----|
| 3.1 | Overview: pitch features of charismatic speech. | 25 |
| 3.2 | Overview: intonation features of charismatic speech. | 27 |
| 3.3 | Overview: rhythm and phrasing features of charismatic speech. . . | 30 |
| 3.4 | Overview: tempo features of charismatic speech. | 31 |
| 3.5 | Overview: phonation quality and loudness features of charismatic speech. | 33 |
| 3.6 | Overview: articulation space features of charismatic speech. | 35 |
| 6.1 | Speakers with age, gender, and origin | 68 |
| 6.2 | Channel overview | 70 |
| 6.3 | Channel and video overview | 71 |
| 7.1 | Annotation tiers for acoustic measurements. | 77 |
| 7.2 | DIMA inventory | 80 |
| 7.3 | DIMA interrater agreement (Cohen’s kappa) | 81 |
| 7.4 | Measurements for analyses and scripts. | 82 |
| 8.1 | Predictions of acoustic feature results of prosodic manipulations . . | 87 |
| 8.2 | Overview of prosody-related stimulus manipulations | 92 |
| 8.3 | Experimental lists for prosodic manipulations | 97 |
| 8.4 | Participant demographics | 101 |
| 8.5 | Model output for pitch level and <i>authentic</i> ratings | 104 |
| 8.6 | Model output for pitch level and <i>enthusiastic</i> ratings | 105 |
| 8.7 | Model output for pitch range and <i>enthusiastic</i> ratings | 108 |
| 8.8 | Model output for final contour and <i>authentic</i> ratings | 110 |
| 8.9 | Model output for final contour and <i>persuasive</i> ratings | 112 |
| 8.10 | Model output for speech rate and <i>authentic</i> ratings | 114 |
| 8.11 | Model output for speech rate and <i>enthusiastic</i> ratings | 115 |
| 8.12 | Model output for pitch level and <i>charismatic</i> ratings | 118 |
| 8.13 | Model output for pitch range and <i>charismatic</i> ratings | 121 |
| 8.14 | Model output for final contour and <i>charismatic</i> ratings | 124 |
| 8.15 | Emmeans output for final contour and <i>charismatic</i> ratings | 125 |
| 8.16 | Overview of the effects of the prosodic manipulations | 128 |
| 9.1 | Predictions of pause duration effect on attribute ratings | 145 |
| 9.2 | Overview of pause-related stimulus manipulations | 149 |
| 9.3 | Experimental lists for pause manipulations | 150 |
| 9.4 | Model output for pause duration and <i>likable</i> ratings | 155 |
| 9.5 | Model output for breathing noise and <i>authentic</i> ratings | 157 |
| 9.6 | Model output for breathing noise and <i>persuasive</i> ratings | 160 |
| 9.7 | Model output for breathing noise and <i>charismatic</i> ratings | 163 |
| 9.8 | Emmeans output for pause duration and <i>charismatic</i> ratings | 163 |
| 9.9 | Model output for breathing noise and <i>charismatic</i> ratings | 166 |
| 9.10 | Overview of the effects of pause-related manipulations | 168 |

| | | |
|------|--|---------|
| 10.1 | Acoustic measurements of short stimuli | 189 |
| 10.2 | Acoustic measurements of long stimuli | 190 |
| 10.3 | Correlations of pitch-related features | 195 |
| 10.4 | Correlations of pitch peak timing and prominence ratio | 201 |
| 10.5 | Correlations of prominence level frequency | 205 |
| 10.6 | Correlations of phrase duration and speech rate | 207 |
| 10.7 | Correlations of speech rate variability and stimulus duration | 207 |
| 10.8 | Summary of correlation results of short stimuli | 212 |
| 10.9 | Summary of correlation results of long stimuli | 213 |
| 11.1 | Remote experiment recommendations | 242 |
| A.1 | YouTube statistics | i |
| B.1 | Measurements of prosody-related stimuli (short) | iii |
| E.1 | Model code for pitch level | xxiii |
| E.2 | Model code for pitch range | xxiv |
| E.3 | Model code for final contour direction | xxiv |
| E.4 | Model code for speech rate | xxv |
| F.1 | Model output for pitch level and <i>likable</i> ratings | xxvii |
| F.2 | Model output for pitch level and <i>persuasive</i> ratings | xxvii |
| F.3 | Model output for pitch range and <i>authentic</i> ratings | xxviii |
| F.4 | Model output for pitch range and <i>likable</i> ratings | xxviii |
| F.5 | Model output for pitch range and <i>persuasive</i> ratings | xxviii |
| F.6 | Model output for final contour and <i>enthusiastic</i> ratings | xxix |
| F.7 | Model output for final contour and <i>likable</i> ratings | xxix |
| F.8 | Model output for speech rate and <i>likable</i> ratings | xxix |
| F.9 | Model output for speech rate and <i>persuasive</i> ratings | xxx |
| F.10 | Model output for speech rate and <i>charismatic</i> ratings | xxx |
| G.1 | Emmeans output for pitch range and familiarity | xxxi |
| H.1 | Measurements of pause-related stimuli (long) | xxxiii |
| I.1 | Model code for pause duration | xxxv |
| I.2 | Model code for breathing noises | xxxvi |
| J.1 | Model output for pause duration and <i>authentic</i> ratings | xxxvii |
| J.2 | Model output for breathing and <i>enthusiastic</i> ratings | xxxviii |
| J.3 | Model output for breathing and <i>charismatic</i> ratings | xxxix |
| J.4 | Model output for breathing and <i>charismatic</i> ratings | xl |

List of Figures

| | | |
|------|--|-----|
| 7.1 | Example annotation tiers | 76 |
| 8.1 | Instructions and rating screens for perception experiments | 95 |
| 8.2 | Experiment participants per recruitment method | 100 |
| 8.3 | Results for pitch level and <i>authentic</i> ratings | 104 |
| 8.4 | Results for pitch level and <i>enthusiastic</i> ratings | 105 |
| 8.5 | Results for pitch level and <i>likable</i> ratings | 106 |
| 8.6 | Results for pitch level and <i>persuasive</i> ratings | 106 |
| 8.7 | Results for pitch range and <i>authentic</i> ratings | 107 |
| 8.8 | Results for pitch range and <i>enthusiastic</i> ratings | 108 |
| 8.9 | Results for pitch range and <i>likable</i> ratings | 109 |
| 8.10 | Results for pitch range and <i>persuasive</i> ratings | 109 |
| 8.11 | Results for final contour and <i>authentic</i> ratings | 110 |
| 8.12 | Results for final contour and <i>enthusiastic</i> ratings | 111 |
| 8.13 | Results for final contour and <i>likable</i> ratings | 111 |
| 8.14 | Results for final contour and <i>persuasive</i> ratings | 112 |
| 8.15 | Results for speech rate and <i>authentic</i> ratings | 114 |
| 8.16 | Results for speech rate and <i>enthusiastic</i> ratings | 115 |
| 8.17 | Results for speech rate and <i>likable</i> ratings | 116 |
| 8.18 | Results for speech rate and <i>persuasive</i> ratings | 117 |
| 8.19 | Results for pitch level and <i>charismatic</i> ratings | 119 |
| 8.20 | Results for pitch level, <i>charismatic</i> ratings, and familiarity | 119 |
| 8.21 | Results for pitch range and <i>charismatic</i> ratings | 122 |
| 8.22 | Results for pitch range, <i>charismatic</i> ratings, and familiarity | 122 |
| 8.23 | Results for final contour and <i>charismatic</i> ratings | 125 |
| 8.24 | Results for final contour, <i>charismatic</i> ratings, and familiarity | 125 |
| 8.25 | Results for speech rate and <i>charismatic</i> ratings | 126 |
| 8.26 | Results for speech rate, <i>charismatic</i> ratings, and familiarity | 126 |
| 9.1 | Pause durations in the corpus data | 146 |
| 9.2 | Results for pause duration and <i>authentic</i> ratings | 152 |
| 9.3 | Results for pause duration and <i>enthusiastic</i> ratings | 154 |
| 9.4 | Results for pause duration and <i>likable</i> ratings | 156 |
| 9.5 | Results for pause duration and <i>persuasive</i> ratings | 156 |
| 9.6 | Results for breathing noise and <i>authentic</i> ratings | 158 |
| 9.7 | Results for breathing noise and <i>enthusiastic</i> ratings | 159 |
| 9.8 | Results for the breathing noise and the <i>likable</i> ratings | 159 |
| 9.9 | Results for breathing noise and <i>persuasive</i> ratings | 161 |
| 9.10 | Results for pause duration and <i>charismatic</i> ratings | 164 |
| 9.11 | Results for pause duration, <i>charismatic</i> ratings, and familiarity | 165 |
| 9.12 | Results for breathing noise and <i>charismatic</i> ratings | 167 |
| 9.13 | Results for breathing noise, <i>charismatic</i> ratings, and familiarity | 167 |

| | | |
|-------|---|------|
| 10.1 | Stimuli and corpus measurements for pitch-related features | 194 |
| 10.2 | <i>Charismatic</i> ratings and mean F0 (short stimuli) | 196 |
| 10.3 | <i>Authentic</i> ratings and pitch variability (short stimuli) | 197 |
| 10.4 | <i>Enthusiastic</i> ratings and median pitch (long stimuli) | 198 |
| 10.5 | <i>Likable</i> ratings and pitch variability (long stimuli) | 199 |
| 10.6 | <i>Charismatic</i> ratings and mean F0 (long stimuli) per gender | 199 |
| 10.7 | Peak timing and prominence ratio (stimuli and corpus) | 200 |
| 10.8 | Frequency of prominent syllables (stimuli and corpus) | 201 |
| 10.9 | <i>Authentic</i> ratings and peak timing (short stimuli) | 202 |
| 10.10 | <i>Persuasive</i> ratings and peak timing (short stimuli) by gender. | 202 |
| 10.11 | <i>Enthusiastic</i> ratings and prominence ratio (short stimuli) | 203 |
| 10.12 | <i>Likable</i> ratings and peak timing (long stimuli) | 204 |
| 10.13 | <i>Charismatic</i> ratings and prominence ratio (long stimuli) | 204 |
| 10.14 | <i>Likable</i> ratings and emphatic accent frequency (long stimuli) | 205 |
| 10.15 | Phrase duration and speech rate (stimuli and corpus) | 206 |
| 10.16 | Standard deviation of speech rate (stimuli and corpus) | 206 |
| 10.17 | <i>Enthusiastic</i> ratings and phrase duration (short stimuli) | 208 |
| 10.18 | <i>Persuasive</i> ratings and speech rate (long stimuli) | 209 |
| 10.19 | <i>Likable</i> ratings and stimulus duration (long stimuli) | 210 |
| 10.20 | <i>Authentic</i> ratings and speech rate (long stimuli) | 211 |
| K.1 | <i>Enthusiastic</i> ratings and phrase duration in stimuli (long) | xli |
| K.2 | <i>Likable</i> ratings and phrase duration in stimuli (long) | xlii |

Abbreviations

| Abbreviation | Explanation |
|---------------|--|
| AU | Authentic |
| BID | Bio-informational Dimensions |
| CA | Canada |
| CEO | Chief Executive Officer |
| CH | Charismatic |
| cpm | Count per minute |
| CUT | Stimuli with cuts instead of pauses |
| dB | Decibel |
| DIMA | Deutsche Intonation – Modellierung und Annotation (engl. <i>German intonation – modeling and annotation</i>) |
| EMMs | Estimated Marginal Means |
| EN | Enthusiastic |
| ENG | England |
| EPr | Emphatic prominences |
| Exc | Excursion size (difference MaxF0-MinF0) in a phrase |
| ExperimentMFC | Experiment Multiple Forced Choice |
| F0 | Fundamental frequency |
| F1 | First formant |
| F2 | Second formant |
| Fam. | Familiarity |
| H | High tone |
| HF0 | Stimuli with increased pitch level |
| HF0R | Stimuli with widened pitch range |
| HSR | Stimuli with high speech rate |
| Hz | Hertz |
| IPU | Inter-pausal unit |
| JFK | John F. Kennedy |
| kHz | Kilohertz |
| L | Low tone |
| L1 | First language |
| LF0 | Stimuli with decreased pitch level |
| LF0R | Stimuli with narrowed pitch range |
| LI | Likable |
| LMM | Linear mixed-effects model |
| LONG | Stimuli with long pauses |
| LONG_BR | Stimuli with long pauses with breathing noise |
| LONG_NBR | Stimuli with long pauses without breathing noise |
| LSR | Stimuli with low speech rate |
| LTAS | Long-Term Average Spectrum |
| MaxF0 | Maximum F0 in a phrase |

| Abbreviation | Explanation |
|---------------------|---|
| MeanF0 | Mean F0 in a phrase |
| MED | Stimuli with medium length pauses |
| MED_BR | Stimuli with medium pauses with breathing noise |
| MED_NBR | Stimuli with medium pauses without breathing noise |
| MedP | Median pitch in a phrase |
| MinF0 | Minimum F0 in a phrase |
| MIX_BR | Stimuli with one pause without and two pauses with breathing noise |
| MIX_L | Stimuli with one short, one medium, and one long pause |
| ms | Milliseconds |
| MSR | Stimuli with medium speech rate |
| MZ | Mark Zuckerberg |
| NAM | North America |
| ORIG | Unchanged stimulus |
| PE | Persuasive |
| PhrD | Phrase duration |
| PkT | Pitch peak timing |
| PrR | Prominence ratio |
| RMN | Richard M. Nixon |
| s | Seconds |
| SAMPA | Speech Assessment Methods Phonetic Alphabet |
| SHORT | Stimuli with short pauses without breathing noise |
| SJ | Steve Jobs |
| SPr | Strong prominences |
| SR | Speech rate |
| SRV | Speech rate variation |
| SSBE | Standard Southern British English |
| st | Semitones |
| StimD | Stimulus duration |
| syll/s | Syllables per second |
| TD-PSOLA | Time-Domain Pitch-Synchronous Overlap-and-Add resynthesis algorithm |
| ToBI | Tones and Break Indices |
| Var | Pitch variability |
| WPr | Weak prominences |

Chapter 1

Introduction

When people are talking, they always—consciously and subconsciously—present themselves and their personalities through the appearance, attire, and body language, as well as through the sound of their voice. How we are perceived by others is ultimately affected by *how* we speak, not only by *what* we are saying.

The following research in this dissertation project investigates specific acoustic and prosodic features of a sample of English-speaking voices to find out which of these contribute more strongly to the way the speaker is perceived, and if specific acoustic feature characteristics emerge that are rated as positively and engaging as possible in the context of YouTube.

Attributes like *friendly, convincing, authentic, attractive, enthusiastic, motivating, or inspiring* are frequently combined under the umbrella concept *charisma* (see, for example, Klein and House, 1995; Rosenberg and Hirschberg, 2005; Searle and Hanrahan, 2011; Signorello et al., 2012a; Signorello et al., 2012b). This project will both take a look at judgments of charisma directly, as well as some of the related attributes.

Charisma as a concept has been used for non-religious purposes since the early 20th century, and in the religious context for much longer. Max Weber was the first to use the term charisma outside the religious context. He defined charisma as “a certain quality of an individual personality by virtue of which [the speaker] is considered extra-ordinary and treated as endowed with supernatural, superhuman, or at least specifically exceptional powers or qualities” (Weber, 1968, p. 241; cited in Potts, 2009, p. 117). However, today charisma is defined more in terms of the relationship between a speaker or leader, their audience or followers, and the context in which leader and audience interact, as well as the necessity for shared values and ideals between leader and audience members (Klein and House, 1995). Charisma—like the voice—is nowadays believed to be an attribute that all people possess, just to differing degrees (Niebuhr et al., 2016b).

Over the past years, more and more phonetic studies on charismatic voices have appeared. Most of these studies either investigate politicians, or prominent leaders of companies. Both speaker groups are extremely important for gathering data on charismatic voices. Politicians and business leaders alike rely on their leadership

skills. They need to convince their audiences of their ideas and values. They need to convince them to buy or invest in a product, or give them their vote and confidence for political leadership. In this process, the voice has a major impact on their persuasive power.

This project, however, takes a different approach by researching a different speaker group: internet personalities, more specifically: YouTubers (also referred to as content creators). A YouTuber is a person who produces videos on the platform YouTube (Merriam-Webster, n.d.) and stars in them. The YouTuber and their name are a brand for entertainment and business purposes. The YouTubers in this study are all from North America and England, and all of them started their careers on YouTube, but also branched out to other ventures (production companies, clothing, stage shows, writing, etc.). Since these YouTubers built businesses out of their video making, they are considered entrepreneurs who grew their brands by word of mouth first, and then investing back into their product to increase the video quality and reach more and more people (see Kyncl and Peyvan, 2017).

Charisma and, in particular, a charismatic voice are essential also for the success of a YouTube personality, since many other visual channels like appearance and body language are limited in how much can be conveyed over video, depending on the individual video. For example, hand gestures are not always visible in the video frame, and, while the face and its gestures are the center of visual attention, the voice is the one element that is always present. It is one of the first tools a YouTuber can use to grab the attention of a potential audience member, and—even more importantly—to keep the audience engaged and continuing to watch the video instead of swiftly moving on to the next enticing suggestion.

This research project offers the opportunity to further advance several aspects of phonetic research in general and phonetic charisma research in particular. Several topics included and studied in this project have as of yet not been studied before, or only infrequently been included in phonetic (charisma) research.

First, this study adds to the already existing research on phonetic characteristics of charismatic voices which can be applied to and used in presentational and rhetoric training. Conventional training guides mostly focus on appearance, body language, gestures, and content of the message. The voice is often brushed over, and if it is discussed, the explanations and recommendations are either vague or contradictory to other guides. Phonetic research like this study can be used to build a definition of a “charismatic voice” that can be measured and ultimately implemented in digital training materials (e.g., see Niebuhr et al., 2019). This way, the results of the project can potentially also play a role in connecting research and academia to public interest.

Second, YouTube is becoming a more and more popular medium of entertainment. Understanding the voices on the platform and perhaps finding commonali-

ties between different speakers could hint at a specific genre of speech that might be termed “YouTube voice” (see Beck, 2015). This project therefore offers some first indications of what might be expected from speech material from the platform in order to account for specialties in future research—at least when it comes to English-speaking YouTubers and the particular style of video that is analyzed here.

Additionally, YouTubers are a speaker group that combine entertainment with business. They are rarely professionally trained speakers. In contrast to other public speakers like politicians or business speakers, YouTubers rarely if ever talk to an audience in the same room as them: the audiences they address are virtual. Investigating their speech patterns and voice features may also offer new insights into characteristics for charismatic voices in virtual presentations. Especially in today’s day and age, after the Covid-19 pandemic hit in 2020, virtual presentations via Zoom and other video-conferencing platforms are more and more normal. As on YouTube and depending on the format of the virtual presentation, presenters may not be able to see the majority of their audience, if they even see anyone. They are therefore also presenting to a more or less imaginary audience, which can greatly affect the amount of vocal effort a speaker puts into their presentation unless they actively try. Researching charismatic voices on YouTube can therefore also point phonetic research on virtual presentation in the direction of features of interest.

The majority of the project was carried out during or shortly after the Covid-19 pandemic. Therefore, the perception experiments that form the core of the study could not be carried out in person in a laboratory, but were moved online. At the time of creating the experiments, there was not much research available that used virtual, but supervised perception experiments. A major part of this study is the development of experimental methods and set-ups that would cater to all the needs of this study: data security, offline storage, and easy access for participants. Therefore, recommendations and a sample set-up, together with difficulties, are presented. These recommendations can also be used or adapted further for cost-effective and more sustainable perception research in other research projects.

This study additionally uses an annotation system for the annotation of intonation (tones, boundaries, prominences) that was created as a consensus system for German (*Deutsche Intonation – Modellierung und Annotation*, or short DIMA; engl. *German intonation – modeling and annotation*, see Kügler and Baumann, 2019a; Kügler et al., 2019). This system is applied to English data which at the time of writing had been done only once before (Niebuhr et al., 2018b). In general, the amount of data annotated for this investigation far surpasses the annotated data for German (S. Baumann, p.c.). Therefore, this study also helps to validate DIMA and open its doors to international application, which is one of the future goals for the system (Kügler et al., 2019).

Perhaps, however, the most important and new contribution of this project is the

annotation of around 50 minutes of speech data. The majority of phonetic charisma studies are based on holistic analyses without detailed annotations, or at least no information on annotation processes. Aside from the intonation annotation with DIMA, the speech material is also annotated in terms of phrases, words, syllables, and segments, as well as discontinuities (like repairs and hesitation particles), emphatic accents, breathing, and (YouTube-specific) post-production phenomena (i.e., cuts). While not all annotated data could be used in this project, this offers opportunities for future investigations.

Chapter 2 reviews literature and theories on charisma, leadership, and the intersection of charisma and delivery (from a non-phonetic standpoint). It also introduces the charisma definition that is used in this project. Then, Chapter 3 reviews the available previous research on charismatic speech, which includes findings on charismatic voices and their acoustic-prosodic features, as well as other influences (including social factors). Chapter 4 introduces YouTube as a platform in more detail, charisma on YouTube, and the phenomenon of “YouTube voice”. Chapter 5 presents the research questions and hypotheses for the empirical studies.

Chapter 6 describes the speakers and the videos selected. In Chapter 7, the data treatment methods are explained in detail, and the annotation process and the different measurements taken for the study are documented.

Chapters 8 through 10 are the experimental chapters of this dissertation. In the experiment reported in Chapter 8, short stimuli with acoustic-prosodic manipulations (pitch level, pitch range, speech rate, final contour direction) were presented to participants and rated in terms of charisma-adjacent attributes, charisma directly, and familiarity with the speakers. The same is done in Chapter 9 for the second experiment with longer stimuli and pause duration as well as breathing noise manipulations. Acoustics and perception are then combined in Chapter 10 as the unmodified stimuli are acoustically analyzed and correlated with the perception ratings of the charisma-adjacent attributes and charisma directly.

Finally, Chapter 11 discusses all results that were presented before. Since the majority of the project work took part during a global pandemic, remote experiment methods had to be developed, which are also discussed and evaluated. Limitations are highlighted. Future research ideas are also introduced. The thesis is finished with some concluding remarks and a final summary in Chapter 12.

Part I

Background

Chapter 2

Charisma and its many definitions

2.1 Traditional understandings of charisma

2.1.1 From the original understanding to Max Weber

The concept of charisma originated in ancient times with its first written use by the apostle Paul (Potts, 2009). At that time and for almost two millennia afterwards, the concept was purely religious with no reference to leadership and authority or dominance (Potts, 2009). The word 'charisma' came from ancient Greek *charis*. Paul was the first to write down the word 'charisma' in his epistles between 50 and 62 AD with the meaning " 'the gift of God's grace'; it is usually translated as 'spiritual gift' " (Potts, 2009, p. 5). Other meanings include "grace, talent bestowed by God" (Potts, 2009, p. 109). However, the meaning of the word as simply a 'spiritual gift' has changed substantially over time. Until after the 1920s, the term was rarely used, and if so then in the context of theology.

In the 1920s, Max Weber's work *Economy and Society* was published, and it included a reinvention of the concept 'charisma' and new application of the word away from being "divine grace available for the benefit of the community to innate power residing in extraordinary leaders" (Potts, 2009, p. 122) and authority figures. Max Weber defines charisma as follows:

The term 'charisma' will be applied to a certain quality of an individual personality by virtue of which he is considered extraordinary and treated as endowed with supernatural, superhuman, or at least specifically exceptional powers or qualities. (Weber, 1968, p. 241; cited in Potts, 2009, p. 117)

Weber's definition still maintains the notion of a 'spiritual gift', but applies the concept to exceptional leaders, who in conjunction with trust by their followers gain charismatic authority. According to Potts, Weber assumes two types of charisma: "primary charisma" which is assumed to be innate to a person and therefore cannot be learned or practiced; and a "secondary type of charisma" which can be developed and can arise from outside sources, even though there should be a hint of power in a person (Potts, 2009, p. 121). This second type of charisma seems

to be what most subsequent research on charisma focuses on. It can be seen as “some extraordinariness granted a leader by his or her followers” (Holladay and Coombs, 1993, p. 406), though attribution of charisma by followers can be increased by developing the small innate part (Potts calls this “the germ of such power”, 2009, p. 121) with practice.

Additionally, Weber understood charismatic leaders as creating loyal and devoted followers who join the leader in a mission that can have different sources, such as enthusiasm (most likely for a cause), but also desperation or—as the opposite possibility—hope (see Antonakis et al., 2016) and that “the rise of such individuals was often associated with rapid and radical changes” (Grabo et al., 2017, p. 474). Weber’s original charisma definition also suggests that charisma is created mostly in crisis situations (see Antonakis et al., 2016).

Despite its reinvention at the beginning of the twentieth century, the term ‘charisma’ remained rarely used by the public. This changed roughly in the 1960s, but according to Potts (2009) this can be connected directly to the increasing reach of Weber’s theory of charisma in society.

2.1.2 Popularization of charisma and the media influence

The concept charisma was popularized over time. It was not part of general language usage in the United States when Hollywood stars started coming up in the 1920s and 1930s. However, the term had moved into media outlets worldwide by the mid 1960s via reports about politics, mainly referring to John F. Kennedy post his election to President of the United States of America (Potts, 2009).

Within media comments, there was a distinction made between people considered to be celebrities, and charismatic people. Celebrity, then, was seen as the construct of mass media, while charisma was not constructed (Potts, 2009). Consequently, only few public personalities “who seem to transcend normal celebrity” (Potts, 2009, p. 180) received the attribute *charismatic* in media and society when they were felt to have a genuineness and presence around them; popular examples from history and contemporary times would be John F. Kennedy, Martin Luther King Jr., Princess Diana, or Barack Obama (see Potts, 2009).

2.1.3 Contemporary understandings of charisma

According to Potts, the modern meaning of the concept charisma still includes its innateness “that sets certain individuals apart and draws others to them” (Potts, 2009, p. 2). It is therefore no longer seen as a spiritual gift that spreads over a community, but as something exceptional that only few individuals possess. However, unlike in Weber’s definition, the term is not only applied to leaders and authority figures from religion and politics who would traditionally have the power to

help followers through a crisis. It is now also used for entertainers, celebrities, athletes, or conservationists, to name a few—people who lead others in many aspects of modern society (Grabo et al., 2017) by offering opinions and raising awareness about topics, and thereby inspiring followers to do the same in areas that are often not heard or addressed in the political landscape in a way that leads to changes.

When looking at modern day lay definitions of the concept, Grabo et al. (2017) review that “charisma has become more down-to-earth: typically understood as a personality trait related to charm, magnetism or likeability” (p. 473). There is a large range of characteristics attributed to the concept of charisma, ranging from more positive attributes like charm, enthusiasm, motivation, inspiration, likability, persuasion (see, for example, Shamir et al., 1993; Rosenberg and Hirschberg, 2009; Searle and Hanrahan, 2011; Grabo et al., 2017) to more negative attributes like “frightening intensity” (Potts, 2009, p. 185).

In general, many people will be able to name a charismatic person, but they will struggle with defining what it actually is that makes them charismatic (Potts, 2009). And this problem does not only arise in the popular understanding of the concept, but also in the area of leadership research, since the definition of the concept is not clear. According to Grabo et al. (2017), there is no consensus in research yet on how to approach and study charismatic leadership. And even less research has been conducted on an empirical and measurable definition of the concept—that started roughly in 2005, at least for possible influences of the voice on charisma, with the research of Rosenberg and Hirschberg (2005, see Chapter 3 for an overview of empirical voice-related charisma findings).

2.2 Charismatic leadership

Research on charismatic leadership has been carried out in organizational psychology for years. Much of the available research centers on the leader and their characteristics and personality traits, and some focus on the followers and their relationship with leaders. For an extensive overview of (leader-centric) charismatic leadership research, see Antonakis et al. (2016) and Antonakis (2017). All literature in this section refers to leaders in the business context, for example chief executive officers (CEOs), work group leaders, and so on. The strong connection and interaction between YouTubers and their followers (see Chapter 4) is also the reason for the focus of the literature review on leadership research that includes or even focuses on the relationship between followers and leaders.

Even early on in charismatic leadership research, some researchers suggested that charisma could be found in the relationship between a leader and their group of followers, rather than being a specific characteristic or personality trait of the leader (Antonakis et al., 2016; see also Davies, 1954). More specifically, Anton-

akis et al. (2016) refer to the 1977 theory from another researcher, Robert J. House, and they write that “[for] House, charisma was not really a gift or some magical ability, but rather a complex interaction between the leader, the prevailing context, and follower needs” (p. 300). Howell and Frost (1989) also maintain that “the extraordinary, intensely personal relationship between charismatic leaders and their followers” is at the heart of charisma, and based mostly on “emotional rather than rational grounds”, making followers show, amongst others, “loyalty, commitment, and devotion to the leader” (p. 244). It is this interaction between leader and followers that is relevant for the YouTube context where a content creator—in a way a leader of a group or community of viewers—is speaking directly to the viewers and interacting via comments and on Twitter, Instagram, and other social media.

Klein and House (1995) suggest that “[charisma] resides not in a leader, but in the relationship between a leader who has charismatic qualities and a follower who is open to charisma, within a charisma-conducive environment” (p. 183). They use a fire-starting metaphor in their research: the leader with charismatic qualities is the spark, the follower open to charisma is the flammable material, the “charisma-conducive environment” is the oxygen that feeds the fire, which is the charisma that is created in the relationship between the leader and the follower (Klein and House, 1995, 183ff.).

In this framework, the leader shows charismatic behaviors that previous leader-centric research identified. These behaviors include, for example, confidence in the followers and explicit communication thereof as well as directly referring to a collective and a shared identity between followers and leaders (Klein and House, 1995; see also Bass, 1985; Conger and Kanungo, 1987; Shamir et al., 1993). Klein and House also suggest that followers who are part of a charismatic relationship are not weak and easily influenced (a different possibility suggested by Conger and Kanungo, 1988). Rather, they argue (based on findings of Shamir et al., 1993) that the followers already share the vision the leader puts forward, thereby sharing ideals and values. They write that “leaders in charismatic relationships are powerful at least in part because their followers agree with them, in large measure, from the outset” (Klein and House, 1995, p. 185). This framework also does not rely on a crisis for charisma to emerge (as Weber suggested, see Section 2.1.1), but rather different environments or situations that can cause uncertainty. The environment and its conditions refer to a higher level of environment: “the group, organization, industry, community, society, country, or time that provides a potential context for the development of charismatic relationships” (Klein and House, 1995, p. 186).

Of course, there is also criticism of Klein and House’s framework and fire-starting metaphor. Howell and Shamir (2005) suggest that it is problematic to classify the followers as flammable material that needs a spark (i.e., leader) to ignite.

In their opinion, this forces followers into “a limited and passive role” (Howell and Shamir, 2005, p. 99) which may be too restrictive.

One of the outcomes of charismatic relationships between followers and leaders can be that followers are motivated by the leaders. In an organizational context, this could mean that they are motivated to complete tasks more easily, or that they are “willing to make personal sacrifices to achieve the vision the leader has defined” (Klein and House, 1995, p. 187; see also, e.g., Bass, 1988). Looking at the YouTube context of the present investigation, motivation would mean to subscribe, view other videos, or interact with other viewers or the content creator in the comments, while viewers may sacrifice time and—in some cases—money to support the YouTubers. Aside from triggering motivation as well as inspiration (see Searle and Hanrahan, 2011) in followers, Bono and Ilies (2006) suggest that a charismatic leader tends to express positivity which is expected to affect the emotions of the followers (see also Ilies et al., 2012). In an organizational context, they found that this affected motivation and task performance positively, also when there were only short interactions between leaders and followers.

However, charismatic leadership research also calls for caution, and tends to include warnings of the negative aspects of charisma (see, e.g., Caspi et al., 2019). When listening to a leader, especially a charismatic one, an audience should listen with caution, as the person may reveal a darker undertone, meaning or intention after a while (Caspi et al., 2019).

2.3 A new approach: charisma as a signal

2.3.1 Problems with previous charisma definitions

Antonakis et al. (2016) maintain that charisma is a “fuzzy concept”, and that “the concept is ill-defined by using exemplars or by defining it by its outcomes” (p. 292), and earlier (Antonakis et al., 2011), the research group set out to find a new definition of the concept. The current project deals with a more informal type of charismatic leader (not in the sense of CEOs or workplace leaders; see also Tur et al., 2022), but the findings and theories can be applied to all types of leaders.

According to Antonakis et al. (2016), previous definitions of charisma have focused on characteristics of leaders and the strong effects that these characteristics have on followers. They call for an empirical investigation of the outcomes (after an effect of charisma) in order to find out how the construct can be used, “but the construct cannot be defined in terms of the outcomes it should produce” (Antonakis et al., 2016, p. 302).

From previous literature, Antonakis et al. (2016) identified several elements that they bring into their own definition of charisma (see Section 2.3.3). They suggest

that charisma is defined “as a type of leadership whose nature is based on values (i.e., morals), beliefs and symbolism, as well as on emotion, which is expressive in its transmission of information” (Antonakis et al., 2016, p. 303).

2.3.2 Can charisma be taught?

Another important aspect to note is whether or not charisma is seen as an innate quality or something that can be taught and learned. Potts seems to be against the notion of charisma being something that can be taught (Potts, 2009), as he strongly criticizes self-help guides that promise increased charisma by improving a person’s communication skills.

Psychologist Frank Bernieri maintains that charisma is not a skill that can be learned, but training of communication skills and strategies can lead to an approximation (Flora, 2005). Additionally, as part of his definition of charisma, Riggio (1998) calls charismatic persons “brilliant and effective communicators” (p. 387) who convey their feelings to followers (see also Holladay and Coombs, 1993). These observations suggest a focus on communication for charisma, and that (at least) the charismatic impact of a person can be learned.

As a consequence of charisma as a skill that can be learned, charisma may be regarded as a gradual concept and no longer as something categorical and all-or-nothing (Niebuhr et al., 2016b). It follows that the charisma of a person can develop as time progresses, but also that no speaker can be seen as completely un-charismatic. Rather, it means “that some speakers are better than other speakers in producing charismatic speech – and in adapting their speech to certain contextual requirements” (Niebuhr et al., 2016b, p. 368).

While charisma is a teachable skill, Pauser and Wagner (2018) found that—at least in the case of sales personnel—too much training can also be detrimental for charisma perception. This suggests that changing too much of a person’s presence can lead to a negative influence, perhaps in connection with losing part of a person’s authenticity or genuineness through training.

Perhaps there are even two layers of charisma as such: one that is innate (though this is difficult to prove), and one that is a skill, which achieves a charismatic effect with communication. The first one could not be taught to a person. The second one could be—and thereby approximating innate charisma. People with innate charisma (if this exists) would also be able to hone their communication skills to strengthen any natural effect.

2.3.3 Charisma as leadership signaling

Antonakis et al. (2016) introduce and explain a new definition of charisma that addresses many of their criticisms, but at the same time includes elements that

previous definitions had in common. This new definition is based on two assumptions: a) that leadership, when influencing of followers is involved, is not based on authority, but following the leader voluntarily; and b) that there is signaling involved—both verbally and non-verbally (Antonakis et al., 2016, also referring to Awamleh and Gardner, 1999; Frese et al., 2003; Towler, 2003). The following definition of charisma is posed based on these elements and assumptions:

Charisma is values-based, symbolic, and emotion-laden leader signaling. (Antonakis et al., 2016, p. 304)

According to Antonakis et al. (2016), a person can be charismatic without having formal influence over his/her followers. Additionally, the definition does not include a judgment of the morality of the speaker or the signal (as some previous definitions do; e.g., Shamir et al., 1993; for a discussion see Antonakis, 2012) because the inclusion of “values-based” already implies that a leader always communicates their values and morals, and these have to align with those of the followers for the leader to be accepted and for a charismatic effect to occur (Antonakis et al., 2016). Furthermore, “[leaders] cannot say one thing and do another, or signal unrealizable actions, because in the long run they risk losing their credibility and hence the charismatic effect” (Antonakis et al., 2016, p. 305). Therefore, in order to have effective signaling, followers need to gain the impression of an honest leader who communicates information that is truthful (Bastardo, 2020).

Grabo et al. (2017) further suggest that information signaling by leaders already happens when the leader puts increased effort and energy into attracting audience attention. Charismatic leadership, by means of active verbal and nonverbal signals (i.e., content and primarily voice/gestures etc.), is in their view “able [...] to attract the attention of followers, engage and synchronize their emotions, offer a vision, reinforce cultural values and norms, and to provide a shared sense of identity behind which a group of followers can rally” (Grabo et al., 2017, p. 480). Reh et al. (2017) additionally suggest that the perception part of the signaling process is not completely active, but fast and often happening subconsciously based on perceptual cues coming from both the people and the environment.

In order to produce a signal that—as Antonakis et al.’s (2016) definition includes—is based on specific values and symbols, the leader needs intelligence and specific intentions, both of which are indicated by the charisma signal and given to the followers as knowledge about the leader (Bastardo, 2020). Antonakis and colleagues identified three categories of Charismatic Leadership Tactics: framing, substance, and delivery. *Framing* is concerned with the construction of the signal; this includes rhetoric devices such as metaphors, stories, rhetorical questions and others, and it refers to the actual writing and storytelling of the signal (Antonakis, 2017). Framing is an indicator for a leader’s intelligence as it shows the ability

for abstraction by using symbolism (Bastardo, 2020). *Substance* is concerned with the content of the signal which serves as a justification for a leader's agenda (Antonakis, 2017) and is the means to convey the vision of a leader to their followers (Bastardo, 2020). *Delivery* is then about the presentation of the signal, which has to align with the verbal signals from framing and substance (Bastardo, 2020). For Antonakis (2017), the delivery demonstrates that the speaker is passionate and convinced of their vision, and the leader shows this by using non-verbal means like the voice or facial and body gestures to signal confidence and their emotions.

2.3.4 Charisma signaling and informal leadership

Most research on charismatic leadership has in common that the focus is on leaders who have formal authority. Tur et al. (2022) are interested in leaders without formal authority, specifically in digital environments like social media. They investigate TED talks¹ and Twitter posts since they see interacting with these media (e.g. by sharing or promoting them, or watching a talk) as a voluntary and intentional action of the followers. They focus on this informal leadership of social media influencers, many of which are anecdotally referred to as charismatic (Tur et al., 2022) which could occur in comments or reactions, for example.

Tur et al. (2022) define formal leaders as leaders who have “representative authority either by election or by formal appointment in an organization” (p. 3) and the leader has the power to reward or reprimand the followers. Subsequently, informal leaders are defined “as those individuals who signal their beliefs and preferences to others but have no formal influence over them” (Tur et al., 2022, p. 3, italics removed). Being a follower can also involve a cost, like supporting ideas or promoting them, but generally there are no reprimands or consequences for either not following or specifically following certain people. In general, Tur et al. (2022) found in their study that the use of charismatic signaling (only in terms of framing and substance tactics, so delivery was excluded) tends to go together with higher numbers of views and higher ratings of the TED talks, and more sharing of Twitter posts.

YouTubers as the speaker group of this project also do not have formal authority—perhaps even less authority than the examples mentioned by Tur and colleagues (Dr. Anthony Fauci, a scientist, and Greta Thunberg, an activist; see Tur et al., 2022). Content creators on YouTube are primarily entertainers. It is therefore important to keep the notion of informal leaders in mind.

¹TED is an organization that publishes videos of talks and educational content and holds events “to discover and spread ideas that spark imagination, embrace possibility and catalyze impact”, see <https://www.ted.com/about/our-organization>

2.4 Charisma and delivery

Holladay and Coombs (1993; 1994) specifically studied the effects of *delivery* on charisma perception. Their investigation suggests that a charismatic effect of a message is created both by its language (meaning what is written, the structure, as well as the vision/content) and the delivery (meaning voice and prosody, posture, gestures etc.). Similar assertions have been made by, for example, Awamleh and Gardner (1999), Johnson and Dipboye (2008), and Caspi et al. (2019). Holladay and Coombs (1993) suggest that delivery has an effect of leader perception by followers due to their assumption that charisma is a concept that is attributed to a leader by followers. They review previous research that found that a) speakers with stronger deliveries—e.g., using body language like eye contact, gestures (e.g., hand or head movements), and facial expressions, but also speaking fluently and variably—tend to be perceived as more credible than speakers with weak delivery but the same content (see, e.g., Burgoon et al., 1990); and b) that “[charismatic] people are emotionally expressive and energetic” (Holladay and Coombs, 1993, p. 409f., referring to Riggio, 1987).

In their study, Holladay and Coombs (1993) had a trained speaker produce two versions of the same content message: once with a strong delivery (including eye contact, natural gestures, variable pitch and lack of hesitations) and once with a weak delivery (the opposite features). The study found that the stronger delivery condition also resulted in more charismatic perceptions of the speaker (Holladay and Coombs, 1993). In a later study, they also varied the content of the message so that there were four combinations: visionary content and strong delivery, visionary content and weak delivery, non-visionary content and strong delivery, and non-visionary content and weak delivery (Holladay and Coombs, 1994). They found that delivery and content are the main components that come together for the perception of charisma, but that delivery is much more impactful for charisma perception than message content (Holladay and Coombs, 1994). In their study, charisma ratings of the strong delivery conditions (both with visionary and non-visionary content) were higher than the ratings of the weak delivery conditions (again with both content types).

Continuing on from this research on charisma and delivery, Caspi et al. (2019) write that perceived charisma is created by the combination of the content of a message and the more dominant delivery. Thereby they accept and include the results of Holladay and Coombs (1993; 1994). They expand this assumption by saying that when the delivery is good, the speaker can be perceived as charismatic even with a bad, empty, immoral vision, but an important message may be lost to a weak delivery (Caspi et al., 2019). As the basis of their study, Caspi et al. (2019) make the claim that there is a difference in processing speed between message and delivery in that

the delivery is processed much faster than the message of a signal. Something similar concerning embodied signals was put forward by Reh et al. (2017) who argue that embodied signals (like the voice or body language, and therefore related to delivery signals) are more easily processed, often automatically, than content-related signals. Aside from eye contact, gestures, posture, and facial expressions, Caspi et al. (2019) mention “a captivating tone of voice [and] exhibiting verbal fluency” as features of good delivery for rhetoric purposes (p. 4), which likely means variable (i.e., not monotone) pitch and few hesitations and disfluencies.

Caspi et al. (2019) hypothesize that a) delivery is more important for the perception of charisma than the content, b) delivery has a faster impact on charisma perception than the message content, especially when perception is measured in a number of consecutive moments; and c) when there is a misalignment between content and delivery, the first impression is re-evaluated once the content gets processed. The results show that stronger delivery was rated more charismatic than the weak delivery; the same was the case for visionary content which received higher charisma ratings than non-visionary content (Caspi et al., 2019).

Additionally, Caspi et al. (2019) found that when there is alignment between the delivery and the content, the immediate impression that raters have of a speaker—whether this is a charismatic or non-charismatic impression—is usually confirmed as more speech material is presented. That means, for example, that a strong delivery with a visionary content tends to be perceived as more and more charismatic as time goes on (see Caspi et al., 2019, p. 15, Figure 2). However, their results also show that if the delivery and the content are not aligned, raters tend to change their evaluation of a speaker as charismatic or non-charismatic compared to their first impression as more material is presented. For example, after about six minutes of speech, the ratings of strong delivery/visionary content and strong delivery/non-visionary content cross and, while the charisma rating of the former condition keeps rising over time, the rating of the latter condition starts decreasing (see Caspi et al., 2019, p. 15, Figure 2). To Caspi et al. (2019), that suggests that “delivery is processed rapidly and generates the first impression, whereas content is processed slowly and provides either compatible or incompatible perception” (Caspi et al., 2019, p. 20).

2.5 Summary: Charisma in this project

Charisma is—in this project—understood as a collection of other attributes that together create the concept “charisma”. Some of these attributes that have been mentioned in conjunction with charisma in previous research include genuineness (Potts, 2009) and honesty (Bastardo, 2020), likability, charm, enthusiasm, and per-

suasion (e.g., Shamir et al., 1993; Rosenberg and Hirschberg, 2009; Grabo et al., 2017). This list of attributes is by no means exhaustive, though.

Speakers signal several of these and other attributes to their followers by means of content (though this is not explored in this project) and delivery (gestures, body language, and—focus of the project—the voice) of their message. The signaling is one aspect, but a charismatic impression can only emerge from the relationship with the followers. This relationship is heavily tied to the situation (Klein and House, 1995)—or better yet medium, as the relationship between leaders on YouTube and their followers tends to differ from the relationship between a political or organizational leader and their followers. In particular, YouTubers can be seen as informal leaders without authority over the followers (see Tur et al., 2022) since followers actively choose to follow specific speakers from an endless pool of potential choices. Most importantly, there are usually no consequences for following a person—unlike in politics, for example, where the choice of leader can change the way society develops. Nonetheless, YouTubers lead by expressing opinions, signaling their values and emotions. The followers' world views and interests have to be in alignment with those of the YouTubers (see Antonakis et al., 2016; Tur et al., 2022). If the world views and values do not align, a charismatic impression would not occur.

The charismatic abilities of people are gradual: no person is *un-charismatic*. Some people simply are more charismatic than others (see Niebuhr et al., 2016b). However, charisma can be learned and practiced (Antonakis et al., 2016; Niebuhr et al., 2016b), both in terms of message content and delivery (see also Holladay and Coombs, 1993; Awamleh and Gardner, 1999; Caspi et al., 2019). Both these influences on charisma can be honed, elevated and refined through training, but the most immediate option for improvement is the delivery part, especially for short interactions (see Caspi et al., 2019). A person can learn to move or dress differently. To develop the facial expressions might be somewhat difficult. Changing the way a person speaks is manageable once advice from rhetorical training like a 'lively' voice or 'speaking fluently' have measurable acoustic correlates and educated trainers. This project helps to create an empirical basis for understanding how charisma can be conveyed via the voice.

Chapter 3

Charismatic speech: Phonetic and non-phonetic influences

Why should we even think about studying charisma from a phonetic standpoint? Many rhetorics guides and presentational training formats offer some advice on how a “charismatic” or “pleasant” voice should sound for a talk. The advice regarding the voice is mostly vague, abstract and differs from guide to guide. The voice can be trained though, and listeners will always react to the voice as well as the appearance and body language. Empirical information on particular voice characteristics and how they are perceived is needed for an encompassing understanding of charisma.

3.1 Moving from rhetorics to phonetics

In recent years (in particular starting with Rosenberg and Hirschberg, 2005), phonetic investigations of several aspects of charismatic voices, charismatic speaking styles and specific speakers who are perceived as charismatic (or not) have gained ground rapidly. These studies mainly focus on politics (e.g., Rosenberg and Hirschberg, 2005, 2009; Biadys et al., 2007, 2008; Signorello et al., 2012a, 2012b; Signorello and Demolin, 2013; D’Errico et al., 2013; Hiroyuki and Rathcke, 2016; Bosker, 2017; Jokisch et al., 2018; Berger et al., 2020) and on speakers from the business world (Niebuhr et al., 2016a, 2016b, 2017, 2018a, 2018b, 2020a; Novák-Tót et al., 2017; Mixdorff et al., 2018). Berger et al. (2017) and Berger (2017) looked at charismatic speech from a more controlled angle—with lab-recorded read phrases that were then digitally altered for two university lecturers.

Holladay and Coombs (1993) showed that presenters with a stronger delivery were perceived as more charismatic than other speakers, even though the content was identical. Even more so, strong delivery and non-visionary content is still perceived as more charismatic than when visionary content is presented with weak delivery (Holladay and Coombs, 1994; see also Chapter 2). That is why it is important to understand the phonetic features that are relevant for charisma perception.

While a lot of research goes into the concept of charisma and charismatic

speakers in general, knowledge about the *speech* aspect in charismatic speech has not been collected until fairly recently, and what we know is still expanding. That is, phonetic charisma research can offer additional information to the understanding of charisma that are—in contrast to the abstract nature of the phenomenon—measurable in the speech signal and the reactions of listeners to the speech signal. Presentational training guides mostly use vague terms to describe speaker’s voices such as rich, full, or fluent (see Niebuhr et al., 2017, 2020a). Vague terms like these are difficult to convey as an instructor and to implement as a learner (who may not be as familiar with their voice and its impact). Phonetic charisma research can complement these vague terms by offering measurable features of the speech signal. For example, Signorello et al. (2012a) have the hypothesis “that the richer the prosody (higher f_0 values as max, min and range, focus on key words, tonal jumps, rising and falling contour, etc.) the more charismatic the speaker will be perceived” (p. 437)—and therefore also more effortful productions. And knowledge of phonetic features involved in charisma perception can even allow the development of software and systems that are able to monitor and assess a learner’s progress (Niebuhr et al., 2017).

3.2 Charismatic speech: Acoustics and prosody

3.2.1 Pitch

Several phonetic studies have—over the years—included the impact of **mean pitch** or **overall pitch level** on the charisma of voices. Pitch is connected to the so-called fundamental frequency (F0, sometimes also f_0). F0 is the frequency with which the vocal folds come together and move apart, and the lowest frequency in the complex speech signal. The speed of this movement of the vocal folds is measured in Hertz (Hz): how many times per second do the vocal folds connect and move apart? Perceptually, this is the main correlate for pitch (Niebuhr et al., 2020b). The faster the vocal folds move, the higher the perceived voice.

For American English-speaking politicians, results of a study by Rosenberg and Hirschberg (2009; see also 2005) suggest that a higher mean pitch triggered a more charismatic perception. This was also found for a separate experiment with an all-male subset of the same speakers (Biadys et al., 2007). Signorello et al. (2012a, 2012b) and D’Errico et al. (2013) found similar results for Italian speakers.

Studies into business speakers have shown that for male speakers—Steve Jobs (SJ) and Mark Zuckerberg (MZ) as case studies—the speaker perceived as more charismatic (SJ) also spoke with significantly higher mean F0 levels. These F0 levels were in fact “so high that it comes close to average f_0 levels that have been

determined for female speakers” (Niebuhr et al., 2020a, p. 19; see also Niebuhr et al., 2016a).

Niebuhr et al. (2018a) also found that male speakers were perceived as more charismatic (and it was more likely that listeners would invest in their company) when their F0 level was higher, i.e. further away from their baseline. For female speakers, a lower F0 level compared to the baseline was more advantageous.

Part of the charisma definition in this project is that it involves a form of leadership without formal authority. Low pitch is often associated with authority and having power over others, while higher pitch tends to be interpreted as more open and friendly (Gussenhoven, 2016). This part of the definition kept in mind, the findings suggest two things.

A low-pitched speaker conveys power and authority and, on this basis, tells people what to think and do. A high-pitched speaker conveys charisma and, on this basis, makes people adopt his or her point of view so that the intended thoughts and actions are elicited on a voluntary basis. (Michalsky and Niebuhr, 2019, p. 37)

In a controlled lab study—in part similar to the present project—Berger et al. (2017) manipulated pitch level (among other features) for an utterance of two male American English-speaking university lecturers (see also Berger, 2017). There was no effect of unchanged and decreased pitch level on the perceptual rating scales *charismatic*, *convincing*, *attractive*, and *motivating*. Increased pitch level only had an effect on the attractiveness rating: a higher pitch level was perceived as less attractive than the other manipulation levels, which is in line with attractiveness research (see, e.g., Quené et al., 2016) and the Frequency Code (Ohala, 1984, though this theory is nowadays criticized, see Winter et al., 2021). These findings suggest that other features may be more relevant for charisma perception than overall pitch level alone (see also Michalsky and Niebuhr, 2019). Actually, from the perspective of the Frequency Code, one possible interpretation of the higher pitch level that is perceived as more charismatic is submissiveness (Ohala, 1984; Gussenhoven, 2002; Gussenhoven, 2016). According to Niebuhr et al. (2016a), it likely is the combination of pitch level with other features like other pitch features, loudness, tempo, or rhythm that is important for charismatic speech to emerge and turning a voice that could be interpreted as submissive into one that can be interpreted as charismatic.

Novák-Tót (2016) investigated three female speakers, Oprah Winfrey (among other things, CEO of Oprah Winfrey Network) as well as Meg Whitman (CEO of Hewlett Packard Enterprise) and Ginny Rometty (CEO of IBM). Compared to reference values from the literature, these three speakers had lower mean F0 values than the average female American English speaker, or they were at the lower end of the reference values (Novák-Tót, 2016). This suggests that female CEOs tend to

speak with lower voices than other female speakers and perhaps makes them more similar to male speakers in pitch, as well as perceived authority and competence.

Connected to the mean pitch level are the measures of **minimum** and **maximum F0** in a phrase. This refers to the point in a phrase that has the highest measurement of F0 (usually, but not exclusively, this is a pitch peak and connected to a prominent syllable) or the lowest F0 (often, but not exclusively, this coincides with the end of a phrase). Signorello et al. (2012a, 2012b) and D'Errico et al. (2013) found that higher minimum and maximum F0 were perceived as more charismatic for Italian speakers, and Italian and French listeners. Additionally, Biadys et al. (2007) found that higher charisma ratings of American speakers were achieved with higher minimum F0, while the maximum F0 had no effect for their American English sample.

Since all pitch features are connected, the previously mentioned features also influence the **variability of pitch** and the **pitch range** and vice versa. The variability of pitch is often measured as the standard deviation of pitch overall or specifically maximum pitch within one phrase. The pitch range is mostly measured as the difference between minimum and maximum F0 in a phrase (also sometimes referred to as excursion size).

Rosenberg and Hirschberg (2009) found that a larger standard deviation of F0 within one token (sometimes several utterances) triggered a more charismatic perception. They also investigated the variation of pitch in smaller excerpts of the tokens—on phrase level—to account for the liveliness and expressiveness of the intonation on charisma ratings. They found that the standard deviation of maximum pitch, which Rosenberg and Hirschberg equate with pitch range, approached significance in that phrases with more variation in terms of maximum pitch received higher charisma ratings than phrases that varied less in terms of maximum F0 (Rosenberg and Hirschberg, 2009). Because of the quick changes that happen, this may be one way to convey passion and enthusiasm as attributes connected to charisma (Rosenberg and Hirschberg, 2009).

Niebuhr et al. (2018a) also found that speakers with wider pitch ranges—measured in terms of a) larger standard deviation of pitch, and b) higher values of the F0 percentile range—were perceived as more charismatic. They equated charisma with the likelihood of participants investing in a speaker's company. For the two business speakers SJ and MZ, Niebuhr et al. (2020a) also found that SJ as the speaker perceived to be more charismatic used on average about twice as large F0 ranges as MZ.

Berger et al. (2017) found that stimuli (produced by American English speakers and rated by American English-speaking listeners) with digitally expanded pitch range had a positive effect on the ratings of the voices, especially in terms of the attributes *convincing* and *charismatic*. Meanwhile, a narrowed pitch range was rated negatively on those two scales as well as *attractive* and *motivating*.

Contrary to these findings, Biadys et al. (2007) did not find a significant effect of F0 standard deviation on their American English sample.

Mixdorff et al. (2018) make the connection between all these pitch features and more—again investigating the business speakers SJ and MZ. Their results suggest that the high pitch level that was found for SJ is connected to both a higher minimum F0 and larger pitch ranges overall, but especially in terms of higher maximum F0 on pitch accents. MZ, on the other hand, had a lower minimum F0 to begin with. Additionally, other aspects like a lower pitch baseline with longer, higher and faster F0 slopes of pitch accents and boundary tones can also affect the high pitch level often reported for charismatic voices (Mixdorff et al., 2018; see also Michalsky and Niebuhr, 2019).

Niebuhr and Skarnitzl (2019) found that mean and median F0 correlate significantly and positively with charisma perceptions, baseline F0, however, is not correlated. Regarding pitch variability, they found that the 80-percentile range (which is the range calculated between the F0 values at the edges of the 90th and 10th percentiles of all values) had a positive correlation with charisma perception, while the range between maximum and minimum showed no correlation, which suggests that this measure is more likely to be affected by errors that can happen when F0 is measured automatically (Niebuhr and Skarnitzl, 2019). The standard deviation of pitch from the mean was not significantly correlated with charisma ratings. However, Niebuhr and Skarnitzl (2019) include two additional measures relating to the F0 distribution in a measurement sample: skewness (i.e., how symmetrical is the value distribution around the sample's mean) and kurtosis (i.e., how thin are the tails of the distribution). Both measures were negatively correlated with charisma perceptions. For skewness this suggests that a speaker who was perceived as more charismatic mostly used pitch values that were further away from the bottom of their individual pitch range; the kurtosis result suggests that a speaker used their full range more equally.

That means that features like a larger pitch range (especially in terms of the percentile range), higher minimum F0, lower pitch baseline, higher pitch accents, and faster slopes are overall more reliable predictors of speaker charisma than mean pitch level, as it is influenced by all other features. The findings on pitch-related features are aptly summarized by Michalsky and Niebuhr:

Speakers should raise rather than lower their global pitch level. However, local pitch levels that are reached between expressively stressed, high-pitched words and at the ends of utterances should indeed be lowered to the bottom of a speaker's individual pitch range (Michalsky and Niebuhr, 2019).

That suggests that variation is key, which other research also shows (see, for ex-

ample, Rosenberg and Hirschberg, 2005, 2009; Berger et al., 2017; Niebuhr et al., 2020a). Charismatic speech is therefore less about global pitch level, but local changes. Table 3.1 provides an overview of the pitch-related results from previous literature.

3.2.2 Intonation

While the overall pitch level and variability of pitch are extremely relevant for the perception of charismatic voices, these features can be included in the specific context of intonation. Intonation refers to overarching patterns of speech melody created by pitch movements (Crystal, 2003), in this project included in terms of some tonal movement aspects, though this list is by no means exhaustive.

One important aspect for intonation and its impact on charismatic voices is the **phrase boundary behavior** or **final contour direction**—is the direction of the final pitch contour in a phrase rising (i.e., the pitch goes higher), falling (i.e., the pitch goes to the lower end of a speaker’s range), or a plateau (i.e., the pitch contour stays level), and which direction is perceived more positively? Rosenberg and Hirschberg (2009) found that the more rising phrase boundaries were in an experiment token (which could be more than one phrase), the more negatively the listeners rated the token. According to the authors, listeners can associate a final rise with questions or uncertainty of the speakers, which could be interpreted as the opposite of being persuasive and charismatic.

For **pitch accent types**¹, which basically distinguishes between different high tones (H) and low tones (L), Rosenberg and Hirschberg (2009) found that having a higher proportion of H* pitch accents (i.e., accents where a high tone falls within the stressed syllable) in a token resulted in more positive charisma ratings (see also Niebuhr et al., 2018b). Meanwhile, a higher proportion of pitch accents of other types, namely the early peak L*+H and the low pitch accent L*, were perceived as less charismatic (all pitch-accent related notations referring to Rosenberg and Hirschberg are made in with the *Tones and Break Indices* system, or ToBI; see Beckman et al., 2005). In American English, the H* pitch accent is mostly used to introduce new information into a phrase, while the L*+H is used for uncertainty or incredulity, and the L* pitch accent is used to refer to given or old information (Rosenberg and Hirschberg, 2009).

While referring to the same data, Biadisy et al. (2007) found—in contrast to Rosenberg and Hirschberg—that American English stimuli were rated more positively in terms of charisma the more downstepped !H* (i.e., high pitch peaks that are lower in value and perception than the preceding high peak) and late L+H* accents were

¹Pitch accents are tonal maxima or minima in a phrase that perceptually make a syllable that is associated with this movement stand out (i.e., the syllable is stressed or prominent).

Table 3.1: An overview of the previous literature on charismatic speech and the results on pitch features. These features include pitch level, pitch baseline, minimum F0 (= MinF0), maximum F0 (= MaxF0), the variation of pitch (= Var. Pitch), and the pitch range. The gender of the speakers included in each study is provided in parentheses (m = male, f = female). Furthermore, the genre of speech is also provided (P = politics, B = business, L = laboratory).

| Study | Genre | Pitch level | Baseline | MinF0 | MaxF0 | Var. Pitch | Pitch Range |
|--------------------------------|-------|--|---------------------|---------------|------------------|------------------|---------------|
| Biadys et al., 2007 | P | higher (m, f) | - | higher (m, f) | no effect (m, f) | larger (m, f) | - |
| Rosenberg and Hirschberg, 2009 | P | higher (m, f) | - | - | - | larger (m, f) | - |
| Signorello et al., 2012a | P | higher (m) | - | higher (m) | higher (m) | - | - |
| Signorello et al., 2012b | P | higher (m) | - | higher (m) | higher (m) | - | - |
| D’Errico et al., 2013 | P | higher (m) | - | higher (m) | higher (m) | - | - |
| Niebuhr et al., 2016b | B | higher (m) | - | - | - | - | larger (m) |
| Novák-Tót, 2016 | B | lower (f) | - | - | - | - | - |
| Berger et al., 2017 | L | no effect (m) | - | - | - | - | larger (m) |
| Mixdorff et al., 2018 | B | higher (m), affected by other features | lower (m) | higher (m) | - | - | larger (m) |
| Niebuhr et al., 2018a | B | higher than baseline (m); lower than baseline (f) | - | - | - | larger (m, f) | larger (m, f) |
| Niebuhr and Skarnitzl, 2019 | B | higher (m, f) | no effect (m, f) | - | - | no effect (m, f) | larger (m, f) |
| Niebuhr et al., 2020a | B | higher (m) | - | - | - | - | larger (m) |

produced. According to Biadys and colleagues, the late accent is connected to emphasis and contrast. This could be interpreted by listeners as increased enthusiasm or immediacy. Meanwhile, the downstepped !H* is in English connected to lists which are frequently used as rhetorical devices in speeches, and therefore this type of pitch accent could increase a rhetorical effect and charisma (see Biadys et al., 2007). L* accents were reported as negatively correlated with charisma perception, since it is (in English) often used to present known information which could lead to speakers being perceived as boring (Biadys et al., 2007).

Berger et al. (2020) studied one intonation-based feature of John F. Kennedy's (JFK) and Richard M. Nixon's (RMN) voices from the first televised presidential debate in 1960: **pitch-accent timing** or—if only peaks are considered—**pitch peak timing**, which is closely related to the pitch accent type. This measure refers to the time difference between the vowel onset of an accented syllable and the associated F0 maximum (see Berger et al., 2020). Later pitch accents are perceived as more charismatic, conveying new information and signaling something that is considered unusual or surprising, while also conveying more vocal effort through longer rises that can also be substitutes for pitch height (Gussenhoven, 2002). RMN produced pitch accents on average about 100 ms later (relative to accented vowel onset) compared to JFK, “which is large enough a delay to be indicative of a more expressive, emotional prosody” (Berger et al., 2020, 126f.). From a modern perspective that would suggest that RMN may have been perceived as more charismatic. However, to the author's knowledge, there is no further research available on charisma and pitch accent timing.

All these results “might then suggest that charismatic speakers are expected to present ‘new’ rather than ‘old’ information to listeners and are not expected to convey ‘uncertainty’ or ‘incredulity’ in their messages” (Rosenberg and Hirschberg, 2009, p. 647). The late pitch accents play into this. They require more effort of a speaker to produce, and show emphasis and importance (see Gussenhoven, 2002). Table 3.2 provides an overview of the intonation-related results from previous literature.

3.2.3 Rhythm and phrasing

This section reviews results and findings from the literature on several topics relating to rhythm and phrasing (including pause and phrase duration, emphatic accents, and breathing). Rhythm and phrasing again combine aspects of the previous topics pitch and intonation with additional features. The basis for this section's literature review is a case study of the two American CEOs SJ (commonly perceived as a very charismatic speaker) and MZ (known as a less charismatic speaker);

Table 3.2: An overview of the previous literature on charismatic speech and the results on intonation features. These features include the final contour shape, the type of pitch accents (= PA), and the pitch accent timing. The gender of the speakers included in each study is provided in parentheses (m = male, f = female). Furthermore, the genre of speech is also provided (P = politics, B = business).

| Study | Genre | Final contour | PA type | PA timing |
|--------------------------------|-------|------------------------------|---|-------------------------|
| Biadisy et al., 2007 | P | - | more !H* and late accents; less L* (m, f) | - |
| Rosenberg and Hirschberg, 2009 | P | fewer rising contours (m, f) | more H*; fewer early and low peaks (m, f) | - |
| Niebuhr et al., 2018b | B | - | more H* (m) | - |
| Berger et al., 2020 | P | - | - | later pitch accents (m) |

Niebuhr et al., 2020a). This case study investigated all major aspects of rhythm and phrasing. Additional results from other studies are included where fitting.

Rhythm—and more importantly a regular rhythm—plays an important role in intelligibility and communication alike in that “syntactic processing [...], semantic processing [...], and recognition memory are all facilitated by regular meter” (Bosker, 2017, p. 2228). Niebuhr et al. (2020a) investigated two rhythm features: %V which is the proportion of vowels to consonants in an inter-pausal unit (IPU), and VarcoV which is the standard deviation of vowel durations in an IPU normalized by speech rate (see, e.g., White and Mattys, 2007; Wiget et al., 2010; Arvaniti, 2012; Niebuhr et al., 2016b; 2020a), finding that higher variability in these metrics is associated with more charismatic speech.

Furthermore, Niebuhr et al. (2020a) investigated the effect of **pause duration** on charismatic speech. They found that SJ used significantly shorter pauses than MZ, but only when the speech was directed at customers. When directed at investors, MZ used the on average shorter pauses. That suggests different charisma strategies and most importantly different relevance of audience type (customer or investor) for the two speakers. Similarly, D’Errico et al. (2013) also found for manipulated stimuli of Italian and French speakers that shorter pauses generally received higher ratings in terms charisma and 67 charisma-adjacent attributes than longer pauses when rated by listeners without knowledge of the respective language. While shorter pauses therefore seem to be preferred when it comes to charisma, the variation of pause duration (i.e., standard deviation of pause duration between pauses in a sample) was negatively correlated with perceived charisma for American English stimuli, suggesting that pause duration variation may not be perceived as charismatic in American English (Biadisy et al., 2007).

In addition to the pause duration, the **phrase duration** also has an influence on

charisma perception. Niebuhr et al. (2020a) found that prosodic phrases of SJ were significantly shorter than MZ's, both in customer- and investor-directed speech. This was also suggested by a comparison of JFK and RMN, where the phrases in RMN's speech material were significantly shorter than those of JFK (Berger et al., 2020).

Therefore, previous research generally (with some exceptions) suggests that there is a negative correlation between charisma perception and phrase and pause duration—shorter pauses and phrases are perceived as more charismatic. It would follow that more frequent pauses are also advantageous for charisma perception—since shorter pauses and phrases give the speaker the opportunity to produce more pauses and phrases in the same amount of speech time. However, Bidsy et al. (2007) found no evidence for an effect of pause frequency on charisma perception. Equally, it is reasonable to assume that shorter pauses also offer more opportunities for breathing. At the same time, this would also mean short time to breathe (which, when happening quickly, is also more likely audible). Fast breathing is more likely to be supported by chest breathing, but this equally positively affects other prosodic features as well (see Michalsky and Niebuhr, 2019; Barbosa and Niebuhr, 2020).

The final rhythm-related feature reviewed here are **emphatic accents**. Emphatic accents are extra-strong prominences in an utterance, and frequently go in hand with a pitch peak. Most importantly, though, they stand out because of lengthened onset consonants or long vowels—meaning segmental lengthening. Niebuhr et al. (2020a) found that the speech of SJ contained four times the amount of emphatic accents of all different subtypes compared to MZ, which increased the liveliness of SJ's speech. Both speakers used reinforcement accents (i.e., lengthened onset consonants) the most. The second most-used accent type for SJ was positive intensification (i.e., lengthened vowels) and for MZ accent chains (i.e., several emphatic accents in a row; Niebuhr et al., 2020a). Novák-Tót et al. (2017) found similar results in their analysis of the speech of SJ, Oprah Winfrey, and Ginny Rometty. The three speakers were rated similarly in terms of charisma by English-speaking raters. Winfrey had the most emphatic accents, followed by SJ (with a significant difference to Winfrey), and Rometty had the fewest (Novák-Tót et al., 2017). Novák-Tót et al. (2017) suggest that the significant difference in emphatic accent frequency in tandem with the non-significant charisma rating between SJ and Winfrey illustrates that female speakers have to put more acoustic effort into their performance to be perceived as just as charismatic as male speakers. The majority of emphatic accents were reinforcements for all three speakers, followed by positive intensifications for SJ and Rometty; SJ and Winfrey used significantly more accent chains than Rometty (Novák-Tót et al., 2017).

Table 3.3 provides an overview of the results regarding rhythm and phrasing features and their influence on charismatic speech.

3.2.4 Tempo

Rosenberg and Hirschberg (2009) did not find a statistically significant effect of speech rate variation in an experiment token for their political speakers. Each token consisted of several phrases. It is not clear in how far the variation of speech rate between the different phrases in a token was investigated, as variation tends to be seen as the key component of charismatic speech. Therefore, it is surprising that Rosenberg and Hirschberg did not find an effect in their study. They did, however, find that an overall faster speech rate resulted in higher charisma ratings which had—at the time of their study—not been mentioned before as a possible correlation with charisma. The authors explain the result by suggesting that “slower speech might convey hesitation and doubt” (Rosenberg and Hirschberg, 2009, p. 647) which would be a great disadvantage, if not the end, for a charismatic impression. They additionally suggest in a different study that there are cultural differences, and that a faster speech rate overall may be perceived as more charismatic, but tempo variation from phrase to phrase may not be as important in American English as in other languages (see Biadys et al., 2007).

Novák-Tót (2016) found that Oprah Winfrey, Ginny Rometty, and Meg Whitman had mean speech rates in the upper to middle range of reference values from the literature with female American English speakers. The mean of the speech rates of the three speakers fell between 4 and 5 syllables per second (syll/s; Novák-Tót, 2016). Additionally, they had significantly higher speech rate variability than the reference values (Novák-Tót, 2016), suggesting that—contrary to the findings of Biadys et al. (2007)—variation of speech rate between phrases sets charismatic speakers apart from the mass of speakers also for (at least some) American English speakers.

In case studies of MZ and SJ (e.g., Niebuhr et al., 2016a, 2016b; Niebuhr et al., 2020a), results show that MZ had a significantly higher speaking rate than SJ, upwards of 6 syll/s. However, “although [SJ’s] speaking rate level was lower than that of [MZ], it is still higher than mean rates found for ‘ordinary’ speakers of American English” (Niebuhr et al., 2020a, p. 21). It might very well be that MZ’s speechrate exceeds the limits of what is perceived as charismatic, and that the important thing is to have a speech rate that is higher than average, but not so high that it invites many articulatory reductions. Niebuhr et al. (2020a) call this an “effectiveness window” (p. 27) beyond which high speech rate is detrimental for charisma perception (see also Rosenberg and Hirschberg, 2009; Niebuhr et al., 2017).

Berger et al. (2017) and Berger (2017) digitally altered audio stimuli to investigate

Table 3.3: An overview^w of the previous literature on charismatic speech and the results on rhythm and phrasing features. This includes the rhythm measures %V and Varco_V, as well as pause duration and frequency, phrase duration, breathing (= BR), and emphatic accents (= Emph. acc.). The gender of the speakers included in each study is provided in parentheses (m = male, f = female). Furthermore, the genre of speech is also provided (P = politics, B = business).

| Study | Genre | %V | Varco_V | Pause dur. | Pause freq. | Phr. dur. | BR | Emph. acc. |
|-----------------------------|-------|------------|------------|-----------------------|------------------|-------------|----------------|-------------|
| Biadys et al., 2007 | P | - | - | less variation (m, f) | no effect (m, f) | - | - | - |
| D'Errico et al., 2013 | P | - | - | shorter (m) | - | - | - | - |
| Novák-Tóth et al., 2017 | B | - | - | - | - | - | - | more (m, f) |
| Michalsky and Niebuhr, 2019 | B | - | - | - | - | - | short, intense | - |
| Barbosa and Niebuhr, 2020 | B | - | - | - | - | - | short | - |
| Berger et al., 2020 | P | - | - | - | - | shorter (m) | - | - |
| Niebuhr et al., 2020a | B | higher (m) | higher (m) | shorter (m) | - | shorter (m) | - | more (m) |

Table 3.4: An overview of the previous literature on charismatic speech and the results on tempo-related features. This includes the speech rate and its variation. The gender of the speakers included in each study is provided in parentheses (m = male, f = female). Furthermore, the genre of speech is also provided (P = politics, B = business, L = laboratory).

| Study | Genre | Speech rate | Speech rate variation |
|--------------------------------|-------|--------------------------|-----------------------|
| Biadys et al., 2007 | P | faster (m, f) | no effect (m, f) |
| Rosenberg and Hirschberg, 2009 | P | faster (m, f) | no effect (m, f) |
| Novák-Tót, 2016 | B | faster (f) | larger (f) |
| Berger et al., 2017 | L | faster, not too fast (m) | - |
| Niebuhr et al., 2020a | B | faster, not too fast (m) | - |

the effect of differences in speech rate on the perception of charisma as well as the charisma-adjacent attributes *attractive*, *convincing*, and *motivating*. A slow speech rate was rated very negatively on all scales, while unchanged and increased speech rate were rated best—without significant differences on either of the scales. There was a tendency of a preference of unchanged speech rate on the charismatic and attractive scales which would suggest that speech rate should not be too high in order to be still perceived as charismatic (see Niebuhr et al., 2020a). Splitting up the data by speaker suggested that a reason for the tentative preference for unchanged speech rate likely comes from the manipulations that were created and the way the speakers themselves spoke. The unchanged speech rate of one speaker was just as fast as the increased speech rate of the other speaker and was rated just as positively. The original speech rate of the first speaker was therefore already extremely fast (6.484 syll/s), and the increased speech rate was even faster (7.484 syll/s) which likely a) did not sound as natural anymore, and b) was too fast to still be perceived as charismatic.

Table 3.4 provides an overview of the reviewed findings on the influence of tempo and speech rate on charismatic speech.

3.2.5 Phonation quality and loudness

A previous spectral study into charismatic speech in the broader business context suggests that “the louder and less breathy a speaker’s voice was, the better the speaker’s performance was rated by the listeners” (Niebuhr et al., 2018a, p. 361). This was measured using Long-Time Average Spectra (LTAS)² and significant results were found for two spectral features: the alpha ratio (see Frøkjær-Jensen and Prytz, 1976) which is also referred to as spectral slope and measured as the energy

²A spectrum consists of all frequency components of a complex speech signal (including F0 and higher harmonic frequencies; see Niebuhr et al., 2020b, for an overview of frequency components) and their amplitude/energy in the signal at one specific point within that signal. These spectra are created for many time points and then averaged for LTAS.

ratio between two frequency regions in the spectrum (0-1 kHz area and 1-5 kHz area), and the ratio of energy between the frequency areas of 1-5 kHz and 5-8 kHz (see Guzman et al., 2013). The alpha ratio was positively correlated with speakers' perceived leadership experience, which suggests a louder voice being perceived as more charismatic. The energy ratio between the 1-5 and 5-8 kHz frequency areas was negatively correlated, suggesting that less breathy voices were preferred by the listeners (Niebuhr et al., 2018a). Similarly, sections of speech with modal voice quality were rated as—among others—more likable and trustworthy compared to breathy, creaky and pressed voice quality for Czech speakers (Tylečková et al., 2017).

Niebuhr et al. (2018a) also investigated the so-called **Speaker's Formant** which is an energy peak in LTAS that occurs in the frequency area between 3 and 4 kHz, mainly for professional male voices, and makes voices sound more sonorous (sometimes also referred to as the **Actor's Formant**; see Bele, 2006; Master et al., 2008; Master et al., 2012). Niebuhr et al. (2018a) calculated two measurements for the difference in energy between two frequency ranges, namely the range where the first formant F1 usually occurs (300-800 Hz), and then the range of 2-3 kHz on the one hand, and 3-4 kHz on the other hand. They correlated the Speaker's Formant measures with the likelihood of perception experiment participants to invest in the speaker's company as an approximation of charisma. There were no significant patterns found for female speakers (even with an adjustment of the measurement values by a multiplication with a factor of 1.2, which was a way to account for spectral differences based on the female vocal tract, which is on average shorter than that of male speakers and would affect the energy in specific frequencies (Niebuhr et al., 2018a). There were significant negative correlations for male speakers, though, suggesting that the smaller the amplitude difference was between the regions of the Speaker's Formant and the area of F1, meaning the more pronounced the Speaker's Formant was, the more likely participants were to invest (Niebuhr et al., 2018a).

The spectral results of Niebuhr et al. (2018a) suggest that louder voices are perceived as more charismatic. Additionally, other findings suggest that the variability of intensity (as a measure of the acoustic energy in a signal) can also have an effect on charisma perception. Rosenberg and Hirschberg included the **standard deviation of intensity** as a loudness feature in their analyses to investigate the effect of the variability of loudness. The tokens in their study were between 2 and 28 seconds (s) long and included complete sentences, but in many cases comprised several phrases. Phrases were rated as more charismatic with a larger standard deviation of intensity (Rosenberg and Hirschberg, 2009; see also Berger et al., 2020 for similar findings with speakers JFK and RMN) which suggests that variation of intensity might be relevant for other attributes related to charisma, like enthusi-

Table 3.5: An overview of the previous literature on charismatic speech and the results on phonation quality and loudness features. This includes the alpha ratio (energy ratio between 0-1 kHz and 1-5 kHz areas of spectrum), the energy ratio between 1-5 kHz and 5-8 kHz areas of spectrum, and the speaker’s formant (= SF). Additionally, the findings on the variability of intensity as a loudness measure are included. The gender of the speakers included in each study is provided in parentheses (m = male, f = female). Berger et al. (2020) and Rosenberg and Hirschberg (2009) are from politics, the two other studies from business.

| Study | alpha ratio | 1-5/5-8 kHz ratio | SF | Variability of intensity |
|------------------------------|------------------------------|-----------------------------------|--------------------------------|--------------------------|
| Niebuhr et al., 2018a | higher (i.e., louder) (m, f) | lower (i.e., less breathy) (m, f) | more energy (m), no effect (f) | - |
| Rosenberg & Hirschberg, 2009 | - | - | - | larger (m, f) |
| Berger et al., 2020 | - | - | - | larger (m) |
| Niebuhr et al., 2020a | - | - | - | no difference (m) |

asm or passion (Rosenberg and Hirschberg, 2009) because of the quick changes that happen. Contrary to those findings, Niebuhr et al. (2020a) found no differences in intensity variation between the two speakers SJ and MZ.

Table 3.5 provides an overview of the reviewed literature and the findings therein regarding phonation quality and charismatic speech, as well as the influence of the variability of intensity.

3.2.6 Articulation space

Niebuhr (2017) used stimuli with the same content, but different degrees of articulatory reduction to find out how different reduction levels are perceived by listeners. When the reduction was consistently substantial, the speakers were perceived as overall less charismatic, for example in terms of being perceived as more stressed, less optimistic, or less sincere. Similarly, a later case study of MZ and SJ revealed that MZ has almost double the amount of place assimilations for consonants, significantly smaller differences between voiced and voiceless plosive closure times, and three times more devoiced plosives when compared to SJ (Niebuhr et al., 2020a; see also Niebuhr et al., 2018b; Niebuhr and Gonzalez, 2019), which also leads to more negative and less charismatic perceptions.

Another case study of SJ and MZ found that the two speakers also differ in their vowel spaces: MZ’s vowel space was smaller than SJ’s which means that MZ used vowels that were more similar to each other in quality and therefore less distinct than SJ (Niebuhr et al., 2020a; see also Niebuhr and Gonzalez, 2019). According to Niebuhr and Gonzalez (2019), on the dimension of the first formant (F1, relating to

how open the mouth is when producing a vowel; low values correspond to closed vowels, high values to open vowels), “SJ’s vowel space extends beyond that of MZ in both vertical directions. High and mid vowels are produced higher, low vowels are produced lower” (p. 349). On the dimension of the second formant (F2, relating to the back-front dimension: low values correspond to back vowels, high values to front vowels), SJ also has a larger vowel space than MZ, both for front and back vowels (Niebuhr and Gonzalez, 2019). Additionally, there is a difference between vowel spaces and the audience that is addressed: vowel qualities were more distinct when SJ and MZ addressed customers than when they addressed investors, and MZ did this to a stronger degree (Niebuhr and Gonzalez, 2019).

Niebuhr (2020) investigated the effect of the width and height of vowel spaces (F1 and F2 ranges) as well as their shape on charisma perception—which was determined by the scales *charismatic*, *captivating*, *passionate*, *trustworthy*, and *decided* in a perception experiment. Large and small vowel spaces, as well as horizontally (back-front axis) and vertically compressed vowel spaces (open-closed axis) were included. All correlations between the F1 and F2 ranges and the scales were positive and significant. Highest charisma ratings were reached by large vowel spaces, followed by the horizontally or vertically compressed vowel spaces, and the small vowel spaces had the lowest ratings (Niebuhr, 2020). A smaller F1 range (back to front compression) decreased ratings on the *passionate* and *captivating* scales, which Niebuhr (2020) relates to emotions, and a smaller F2 range was detrimental for the cognition scales *trustworthy* and *decided*. These findings suggest that clearer and more distinct vowel productions are perceived as more charismatic (Niebuhr, 2020). Additionally, the different formant dimensions may be relevant for different aspects of charisma, namely emotion as well as cognition/capability (Niebuhr, 2020).

A focus on articulation, how much reduction is used, and how this affects the way a speaker is perceived can be connected directly to the Effort Code (Gussenhoven, 2002; Chen et al., 2002). The Effort Code suggests that the more effort and energy is exerted on articulation, which mostly happens if what the speaker is saying is important to them, the pronunciation is less reduced, and therefore clearer. That means that “in terms of the Effort Code, the charisma effect of articulation is explained by a clearer pronunciation being an implicit signal of ‘I have something important and meaningful to say’ and/or ‘you, my listeners, are important to me’ ” (Michalsky and Niebuhr, 2019, p. 40). Extreme reduction or mumbling would convey that a speaker is indifferent about their message and/or audience (Niebuhr and Gonzalez, 2019). However, extreme hyperarticulation of segments can reduce the amount of sincerity and composure that is perceived (Michalsky and Niebuhr, 2019). This relates to the balancing act between hyper- and hypo speech suggested by Lindblom (1990). Niebuhr (2017), for example, found that a middle ground of

Table 3.6: An overview of the previous literature on charismatic speech and the results on features related to the articulation space. This includes the place assimilation of consonants, the closure times of voiced and voiceless plosives, the amount of devoiced plosives, and the vowel space area. The gender of the speakers included in each study is provided in parentheses (m = male, f = female). All studies investigate business speakers.

| Study | Assimilation | Difference plosive closures | Devoiced plosives | Vowel space |
|----------------------------|--------------|-----------------------------|-------------------|----------------------|
| Niebuhr et al., 2018b | fewer (m) | larger difference (m) | fewer (m) | - |
| Niebuhr and Gonzalez, 2019 | - | - | - | less centralized (m) |
| Niebuhr, 2020 | - | - | - | larger (m, f) |

moderate articulatory reduction increased the likelihood of a speaker being perceived as sincere or sociable compared to unreduced speech.

Table 3.6 provides an overview of the findings regarding the influence of articulatory reduction on charismatic speech.

3.2.7 Additional findings and applications

With the identified features of the speech of SJ as a charismatic, and MZ as a less charismatic speaker (see, for example, Niebuhr et al., 2016a, 2016b), Fischer et al. (2019) synthesized voices for robots who used the vocal features of the two speakers—among others, higher F0, larger pitch range, higher speech rate, more emphatic accents per minute, shorter pause durations, and more high-pitch accents per minute, like SJ, and the opposite for MZ-like prosody. They found that robots with SJ-like prosody were perceived as more enthusiastic, charming and passionate (for example), but not as significantly more persuasive than with MZ-like prosody. Additionally, the participants interacting with robots with SJ-like prosody were more likely to fill out a long instead of a short questionnaire, and they were more likely to follow suggestions for sightseeing stops (Fischer et al., 2019).

Likewise, navigational systems speaking with SJ-like prosody prompted participants in a different experiment to follow the directions more closely, even though participants were aware it was not the fastest route to the final destination. They trusted the system knew about problems on the route. With MZ-like prosody, this was not the case, and many participants stopped the experiment with the first wrong direction and suspected that the system was faulty (Niebuhr and Michalsky, 2019). That shows that charismatic speech and voices also have a major impact on and possible applications in technology.

Similarly, Mileva et al. (2020) found for the effect of the voice on voting behavior in student elections that voting behavior could not be consistently predicted by only one acoustic cue, but rather a combination of different—also multi-

modal—cues. However, they found that trustworthiness (another attribute that can be connected to charisma) was directly connected to the outcome of the election, and “trustworthiness as judged from the candidates’ voices was the best predictor of election success” (Mileva et al., 2020, p. 622).

3.3 Non-phonetic influences on charismatic speech

In this section, previous research findings on other influences on charismatic voices are reviewed. These other influences are aspects of the research that are not necessarily phonetic, but outside sources. However, they have an effect on charismatic voices (phonetically) and their perception.

The list of influences on charisma discussed below is not exhaustive. It has been found that speech act and syntax have an influence on charisma perception (Signorello et al., 2012a, 2012b; D’Errico et al., 2013). Likewise, the discourse style—for example, informal interview, formal address, or formal monologue—can also have an effect on the perception of charisma (Signorello and Demolin, 2013). Finally, the audience also has an influence (Signorello and Demolin, 2013; Mixdorff et al., 2018; Niebuhr et al., 2020a), also in terms of (imagined) audience gender (Gutnyk et al., 2019, 2020), and the raters’ age (Jokisch et al., 2018). The speaker’s age may have less of an impact on charismatic speech (see Jokisch et al., 2018), though this is rarely explicitly investigated. The following sections address selected non-phonetic influences in more detail, including the content and duration of the presented material, social influences like speaker gender and culture/origin, familiarity with the speakers, as well as audio compression.

3.3.1 Content

According to Rosenberg and Hirschberg (2005; 2009), the topic of the statements that were presented to the raters of their perception experiment did not have a statistical effect on the charisma ratings, even though the topics that occurred often elicited strong reactions. They investigated Democratic speakers running for the nomination to become the Democratic party’s Presidential candidate. As such, the material used for the perception experiments included a wide range of topics that can very easily stir up strong emotions and opinions like “post-war Iraq, health care, taxes, reason for running” for nomination as Presidential candidate (Rosenberg and Hirschberg, 2009, p. 645). One would expect an influence of topics like these on the ratings of charisma and many charisma-adjacent attributes, but that is not what Rosenberg and Hirschberg found.

However, the authors do not disclose whether or not the participants were asked to indicate their own political orientation. It is likely that content and audience

are intertwined, and depending on values and expectations of the audience, they might perceive and interpret the same content differently. It might therefore be that the majority of participants leaned towards a Democratic ideology anyway which would likely not influence a charisma rating negatively. Though, in this case, one could perhaps expect a positive correlation with the charisma ratings, as the content would amplify and support ideas and values already shared by the listeners, which also was not found. Two possible interpretations are that either participants reactions were so balanced that they canceled each other out, or that the topic or content really does not have an effect on the perception of charisma. The latter option seems unlikely, though, especially because many previous charisma investigations focus on the importance of content or at least the interplay of content and delivery for charisma (e.g., Holladay and Coombs, 1994; Caspi et al., 2019).

3.3.2 Duration of presented material

Studies by Rosenberg and Hirschberg (2005, 2009) showed that the effect of a positive correlation between the amount of words in a statement (on a lexico-syntactic level) and the charisma judgments approached significance. According to the authors, the perception of charisma was positively correlated with the amount of speech material that was presented to the experiment participants. Likewise, phrases with more words in them received a higher charisma rating (Rosenberg and Hirschberg, 2009).

Something similar was observed by Jokisch et al. (2018). They investigated German politicians and had participants rate the charisma of speakers' voices in stimuli of different lengths. The duration of the stimuli had a significant effect on the ratings. The ratings were significantly higher for stimuli with a duration between 21 and 25 seconds compared especially to the shortest stimuli with durations between 1 and 6 seconds. However, stimuli that were longer than 25 seconds were perceived as less charismatic again, perhaps suggesting a sweet spot of charisma perception with a duration of between 21 and 25 seconds of speech (Jokisch et al., 2018). According to Caspi et al. (2019), charisma perception of weak and strong delivery starts to drift apart at around 20 or 30 seconds of presented material (see also Section 2.4), so it is possible that listeners need this time to register if a delivery is perceived as strong or weak. In their study, the content seems to have an effect much later (Caspi et al., 2019). Another possibility to explain this "sweet spot" could be that participants get bored as perhaps the stimulus duration got too long in an experiment setting.

These findings suggest that in experimental set-ups, stimuli with a length between 21 and 25 seconds might get the highest charisma ratings from listeners. It might be that before that point, the listeners only make a snapshot first impression

judgment of the speakers. That may be less favorable in terms of charisma proper, but might be insightful for spontaneous first impressions. If the stimuli are between 21 and 25 seconds long, speakers might be more familiar with the voice and judge it more highly. Afterwards, perhaps, the content might become more relevant as enough material is presented to give a deeper insight into what the speaker is actually conveying (see Caspi et al., 2019, for findings on the timing of the influence of a mismatch between the content and the delivery strength).

3.3.3 Speaker gender

Jokisch et al. (2018) analyzed gender effects of German political speakers and the effects on the perception of the speakers. The sample consisted of 14 German politicians (7 male, 7 female) between the ages of 31 and 68 (mean age around 50 for both male and female speakers; see Jokisch et al., 2018). Different excerpts were chosen for each speaker from videos of debates published on YouTube and played to 20 raters (18 male, 2 female; Jokisch et al., 2018). The authors acknowledge the gender bias in the perception sample. The participants were asked to rate the voices on the degree of charisma on a 5-point scale ranging from “no charisma”, translated into a charisma score of 1, to “very strong”, corresponding to a charisma score of 5 (see Table 2 in Jokisch et al., 2018). Male speakers received significantly higher charisma ratings than female speakers. Six of seven male speakers ranked in the top half of the sample, six of seven female speakers in the bottom half (Jokisch et al., 2018; see also, e.g., Niebuhr et al., 2018a; Niebuhr and Wrzeszcz, 2019; Niebuhr, 2020).

Niebuhr et al. (2019) investigated male and female presenters in an entrepreneurial task and the effect of training on their presentation performance. They found that female speakers started the training course with lower prosodic charisma than male speakers. However, training also had a gender-specific effect in that the female speakers benefited from training a lot quicker and were able to catch up with male participants in as little as four hours spent in training sessions. This also suggests that women were able to learn more than men, since they started from a lower baseline (Niebuhr et al., 2019).

Gutnyk et al. (2019) investigated several features of charismatic speech in sales pitches of speakers from four different languages, and—most importantly—they compare male and female speakers. They found that women had a higher pitch level (the median of F0 values in the 90th percentile of data), more variability of pitch (standard deviation), slower speech rate, as well as breathier phonation quality than male speakers—though differences in pitch level and phonation quality between male and female speakers can be expected. Female speakers also paused significantly less than male speakers (Gutnyk et al., 2019).

3.3.4 Culture

With globalization and the internet comes the question of how speakers from different cultures perceive voices. Phonetic cues for a concept like charisma may be used in similar ways in some cultures, but are likely perceived or produced differently in different parts of the world.

Signorello et al. (2012b) ran a perception experiment with French and Italian listeners on de-lexicalized stimuli of one Italian speaker before and after impairment from a stroke. They tested the effect of intonation contour without the influence of content, segments and voice quality, keeping all other aspects of the methodology constant from previous experiments (Signorello et al., 2012a). The results show that French listeners combine similar attributes as before: “as *charming, who propose, timorous, confident, pleasant, introverted* in the [pre-stroke condition] and as *inadequate, spontaneous, active, leader* in the [post-stroke condition]” (Signorello et al., 2012b, italics in original). Italian listeners, on the other hand, perceived the de-lexicalized stimuli completely differently from the French listeners, namely “as *boring, indifferent and unimportant* in the [pre-stroke condition] and as *attractive, visionary, sexy, cold, passionate, seductive* in the [post-stroke condition]” (Signorello et al., 2012b, p. 347, italics in original). This result does not quite fit the hypothesis that the pre-stroke voice is perceived as more charismatic than the post-stroke voice. That suggests that there seems to be a difference in charisma perception between the French and Italian cultures. One also has to keep in mind that the results represent statistical correlations between attributes and intonation contours. Attributes like *passionate* and *cold* do not seem to go together rationally which is unfortunately not elaborated on further by Signorello et al. (2012b). However, they also mention that their results were preliminary and needed further analysis.

D’Errico et al. (2013) found a cultural effect of pause duration and pitch. They had Italian listeners (with no knowledge of French) rating French stimuli, and French listeners (with no knowledge of Italian) rating Italian stimuli on 67 charisma-adjacent attributes. The study specifically investigated pause durations, pitch level, and speech rate. In terms of pause duration, the study revealed that “short pauses affect the perception of a ‘dominant’, ‘passionate’ and ‘seductive’ dimension of charisma more than long ones” (D’Errico et al., 2013, p. 556). The short pauses were rated more highly with these attributes by Italian participants. Additionally, a leader using short pauses, higher pitch, and higher speech rate was perceived “as ‘prudent’, ‘wise’ and ‘altruist’” by Italian raters (D’Errico et al., 2013, p. 556), while French raters associated these leader attributes with long pauses, lower pitch, and lower speech rate. The results suggest that Italian listeners trust a speaker with shorter pauses, higher pitch and higher speech rate, which D’Errico et al. (2013) attribute to a preference for extroverted personalities, and that French

listeners resonate more with a speaker with longer pauses, lower pitch and lower speech rate, which might be more strongly associated with an introverted personality.

While many studies suggest that variability is key for the perception of charisma, Biadisy et al. (2007) suggest cultural differences in their comparison of American English and Palestinian Arabic. They found, for example, that American English speakers were perceived as more charismatic with faster speech rates overall, but Palestinian Arabic speakers were perceived as more charismatic with variation between very high and low speech rates from phrase to phrase (Biadisy et al., 2007). Similarly, more variation in pause duration was perceived as more charismatic for Arabic, but relatively consistent pause durations were rated more positively for American English speakers. While there were no differences between the two cultures in terms of mean F0 (both overall and for the pitch peaks per intermediate phrase) in that higher F0 was perceived as more charismatic, standard deviation of F0 and maximum F0 was significantly and positively correlated only in Arabic. However, the stimuli of both languages had in common that stimuli that has F0 values in higher areas of a speaker's individual pitch range were perceived as more charismatic than stimuli produced in a speaker's lower ranges (Biadisy et al., 2007). The effect of the standard deviation, however, suggests to the authors that pitch range may be a more important cue of charismatic speech for listeners of Palestinian Arabic stimuli than for American English listeners. Finally, Biadisy et al. (2007) also found that the more !H* and L+H* (downstepped high accent tones, as well as late pitch accents) were used in both languages, the more charismatic a token was perceived.

Gutnyk et al. (2019) found significant differences in several acoustic features for sales pitches of speakers with four different first languages (German, Mandarin Chinese, Ukrainian, and Spanish) in their study. They found, for example, that German speakers had the highest pitch level, Spanish speakers the lowest, and the other two languages were in between. Ukrainian had the highest speech rate, and Mandarin Chinese the lowest, with the other two languages again in between. They also found that the Ukrainian and Spanish speakers in their sample used a significantly larger amount of pauses than German speakers (Gutnyk et al., 2019).

3.3.5 Familiarity

In their 2005 study, Rosenberg and Hirschberg investigated speech material from nine potential Democratic candidates from the 2004 nomination for presidency. They found that familiar speakers were rated as more charismatic than unfamiliar or unrecognized speakers (see also Biadisy et al., 2007; Rosenberg and Hirschberg, 2009). The authors remark that this result can have two opposite implications: Does

the recognition of a speaker increase the chance of them being perceived as charismatic, or does the charisma of the person make them more recognizable? Especially the second suggestion can, potentially, also be flipped, because it seems likely that a less charismatic person might be easily recognized as such.

Lavan et al. (2016) found that listeners who were familiar with a speaker could identify these speakers more easily when listening to both backward- and forward-played speech, and had more difficulties trying to identify speakers they were unfamiliar with. The listeners were familiar with the speakers from professional settings or lectures which may have influenced the identifications.

3.3.6 Audio compression

Finally, as a more technical influence on charismatic speech, Siegert and Niebuhr (2021a, 2021b) found that audio compression, for example when speech signals are transmitted through the internet, can affect prosodic features in such a way that speakers are perceived as less charismatic. The degree of how much specific features (like pitch level, minimum F0 or phonation quality measurements, for example) are affected by data compression differs depending on the compression method (or codec, see Siegert and Niebuhr, 2021b). They also found that this effect is stronger for female speakers, which could either be a result of audio compression methods being optimized for male rather than female speakers, or that listeners were less harsh in their ratings with male than female speakers (Siegert and Niebuhr, 2021a).

3.4 Summary and observations

Rosenberg and Hirschberg draw the conclusion that for most listeners of their study, “charismatic speakers are also perceived as enthusiastic, charming, persuasive, and convincing” (Rosenberg and Hirschberg, 2009, p. 653). They summarize the acoustic properties of charismatic speech as including a higher speech rate, together with higher and more variable pitch as well as more variable loudness, which were more likely to be perceived as charismatic—at least in the context of political speeches. And these features come together in lively, expressive speech, suggesting a connection between expressiveness and charisma (Rosenberg and Hirschberg, 2009).

The previous research has furthermore suggested that variation is one of the main aspects of charismatic speech (e.g., Niebuhr et al., 2016a). However, it is also clear that there are not only a couple of features that create a charismatic impression. Rather, many features work together (Niebuhr et al., 2016a), and not every

speaker necessarily needs to use them all in the same manner to be perceived as charismatic.

However, there are features and characteristics that appear in the research literature again and again: higher pitch level for men, lower pitch level for women; a larger pitch range; more high and late pitch accents; falling and plateau final pitch contours rather than rising; more variable rhythm; shorter pauses; short breathing; many emphatic accents (and preferably variety thereof); few disfluencies; variable, but generally faster speech rate; mostly modal phonation quality (in particular, less breathy voice); and moderate articulatory reductions.

Features connected to increased vocal effort and emphasis (larger pitch range, higher and later pitch accents, higher pitch level, more emphatic accents) can be understood as some of the stronger charisma cues (see Niebuhr et al., 2020a), together with (among others) intensity variability, speech rate, and rhythmic variability as well as phonation quality. Niebuhr et al. (2020a) further consider pause and phrase duration among the weaker charisma cues. These weightings are based on research into business speakers, in particular SJ and MZ as case studies.

Furthermore, previous research so far found no extensive influence of content on charisma perception. Presented material with a duration between 21 and 25 seconds may be rated as most charismatic (Jokisch et al., 2018). Male speakers tend to be perceived as more charismatic than female speakers (Jokisch et al., 2018; Niebuhr et al., 2018a; Niebuhr and Wrzeszcz, 2019; Niebuhr, 2020). It is also likely that there are differences in charisma perception between different (regional) varieties of one language. However, this has not been a focus of research yet, so there are no studies available that explicitly compare charismatic speech of speakers from different regional varieties of the same language. Finally, familiar or recognized speakers were found to be rated more highly in terms of charisma and charisma-adjacent attributes.

Something else should be addressed at this point: all findings that were presented here are based on vastly different sample sizes. Most studies on charismatic speech are more considered case studies with one to three speakers (for example, Signorello et al., 2012a, 2012b; D'Errico et al., 2013; Niebuhr et al., 2016a, 2020a; Novák-Tót et al., 2017; Niebuhr and Michalsky, 2019; Niebuhr and Gonzalez, 2019; Berger et al., 2020; though this list is by no means exhaustive). Others have slightly larger sample sizes around ten speakers (e.g., Rosenberg and Hirschberg, 2005, 2009; Biadys et al., 2007, 2008; Niebuhr et al., 2018a). Only few studies include larger speaker samples, like the study by Niebuhr and Skarnitzl (2019; 51 speakers) or the cross-cultural studies by Gutnyk et al. (Gutnyk et al., 2019, 2020; 80 and 120 speakers, respectively). These vastly different sample sizes can affect the results and ability to generalize. Especially for the case studies, the results are very specific to those speakers in the context of the chosen material, while studies with

larger sample sizes have the opportunity to find more general patterns. This also means that while currently these findings are used as descriptors of charismatic speech, the understanding can and will change as more research on more speakers, but also more diverse speech genres, gets collected. And in the end also every case study adds to the understanding of what charismatic speech can mean and continue to move research into new or old directions, since findings from small samples are a starting point that can then be expanded upon and revisited with larger samples once the case studies have shown possible topics.

The same is the case for varying participant sample sizes in perception experiments, as the bigger a group is, the more general can the results be interpreted. Additionally, the different studies used different rating tasks (e.g., Likert scales or indirect ratings via imaginary investments). These differences can also affect the results and lead to less general interpretations. While this is something that cannot be addressed by research immediately, collecting data from larger and especially more varied speaker and rater samples (in terms of sizes, genders, origins, speech genres, etc.) can ultimately lead to a broader understanding of charismatic speech.

Chapter 4

YouTube: Lights, camera, business

4.1 A history of YouTube

4.1.1 Beginnings and growth of a platform

The video sharing and social networking platform YouTube was founded in 2005 and sold to Google in 2006 (Arthurs et al., 2018). It is consistently in second place of the most visited websites of the world, only behind its parent company Google (Arthurs et al., 2018; Bärthl, 2018; Netcraft, 2023). Since YouTube is owned by Google—which as a major company is oriented towards collecting profits off all daughter companies—it is surprising that about ten years after being sold to Google, YouTube still did not manage to turn a profit (Bishop, 2018). Rather, YouTube “was still regarded by its CEO Susan Wojcicki as in an ‘investment stage’ of development” (Arthurs et al., 2018, p. 7, referring to Rao, 2016). This may have changed by now, as the profit apparently doubled in 2021 compared to the year before, likely also influenced by the global Covid-19 pandemic starting in late 2019 (Hollister, 2021).

By now, YouTube is a global platform, and almost 80 percent of its visitors originate from places other than the United States of America (Kyncl and Peyvan, 2017, see also Arthurs et al., 2018). According to Bärthl (2018), ever since the purchase by Google in 2006, the number of channels that actually create content—and were not only registered to consume content—grew by around 20 percent every year (based on data from 2008 to 2018). However, the actual number of users, channels, and videos on the platform is only estimated as YouTube only provides vague and implicit public numbers (Bärthl, 2018). Some channels easily contain hundreds or even thousands of videos, meaning that YouTube is a treasure trove of speech material for research regarding language change over time, as well as snapshots of speaking patterns (Lee, 2017).

YouTube is a platform and not strictly a streaming service such as Netflix, Disney Plus, Sky, HBO Max, or Amazon Prime Video. The difference between them is that those streaming services produce or purchase their own media content, while YouTube as a platform does not. As a video sharing platform, YouTube has “a unique

role as a repository of popular culture” (Arthurs et al., 2018, p. 3), an ever-growing archive of information and documents from years worth of worldwide uploads. At the same time, it bridges a gap between industry and popular culture that allows YouTube and YouTubers to mediate between audience interaction, popular culture, and brands that financially sponsor the content creation.

4.1.2 Monetization and business on YouTube

On the platform, user-generated content and monetization are closely tied together as monetization is based on how much viewers interact and engage with the videos (e.g., how much the videos are viewed, how much of time of advertisements is viewed and not skipped, etc.; see Arthurs et al., 2018). Some researchers also call YouTube, as a daughter company of Google, “a profit-oriented company that produces audiences as commodities for advertisers” (Bishop, 2018, p. 70). Depending on the interaction of content creators and viewers, advertisements for different brands are shown before and in the middle of videos. This gives exposure to companies who pay for advertising. YouTube has implemented strict guidelines that are supposed to remove and demonetize material that is potentially harmful to the community, and YouTubers have to adhere to these guidelines in order to earn a share of the revenue (Bishop, 2018). In fact, YouTubers get paid 45 percent of ad revenue (55 percent goes to YouTube) in “cost per thousand views”, which means that advertisers pay YouTube and content creators a sum of money which is given out per 1,000 views of a specific advertisement (Hou, 2019, p. 541). Additionally, advertisers might pay YouTubers directly to sponsor specific content. In the UK, the video creators are required by law to make this obvious by including the term “ad” as a tag in their video descriptions (Arthurs et al., 2018, see also Zhang, 2018) and titles (Hou, 2019), though many also add it on screen and say it verbally.

The traditional type of videos on YouTube—or as Arthurs et al. (2018) call them, videos “native to the platform” (p. 5)—are videos produced by amateurs. However, after growing their audiences, YouTubers often start teaming up with advertisers and sponsors—essentially investors—and turn a hobby into a career and further expand.

That means, YouTubers can be seen as entrepreneurs, and they are the face of their own brand. More specifically,

creators are more like entrepreneurs than celebrities. They start out by drawing support from friends and family. As they grow and develop, they expand that circle, relying mostly on word of mouth and social networking. Eventually they reach a size where they’re able to generate revenue [...] that they can reinvest in their nascent business to grow their audience. But as with a start-up, that reinvestment is primarily

in their product—their videos—rather than in marketing or PR. (Kyncl and Peyvan, 2017, p. 133)

Therefore, they still rely on word of mouth to grow, even when they might have already reached large audiences. In this way, they can turn their initial hobby into a career, running content companies (Kyncl and Peyvan, 2017) and might branch out even further into other areas of entertainment and other industries.

Nowadays, channels are frequently started with the goal in mind to make a career on YouTube or even gaining celebrity status (Arthurs et al., 2018). This was a fairly recent development. Vloggers and other content creators who started when advertising opportunities were not available rather happened upon their career path and business. These content creators are also referred to as “entrepreneurial vloggers” (Hou, 2019, p. 540). While they do not tend to make professional film-like productions on the platform (though as channels grow, the video and production quality also tend to improve), they use their online presence to gather revenue, unlike amateur producers (Burgess and Green, 2009a). Additionally, despite being characterized as entrepreneurs, they are actively involved in the community on YouTube and considered “authentic participants” thereof because “they use communicative and aesthetic conventions that are continuous with the practices of the YouTube community” (Burgess and Green, 2009a, p. 104).

4.1.3 Video style: vlog

One of the signature styles of videos on YouTube is the vlog—the term itself is a portmanteau of *video* and *blog*. A definition of a vlog and the activity of vlogging is as follows:

Videoblogging, or “vlogging,” is a dominant form of user-created content, and it is fundamental to YouTube’s sense of community. Typically structured primarily around a monologue delivered directly to camera, vlogs are characteristically produced with little more than a webcam and some witty editing. The subject matter ranges from reasoned political debate to the mundane details of everyday life and impassioned rants about YouTube itself. (Burgess and Green, 2009a, p. 94)

There are several different types of vlog topics, ranging from lifestyle to gaming to fashion, beauty, and many more (see Arthurs et al., 2018), but all have in common that the YouTubers show their own emotions, opinions, and reactions, and interact directly with their viewers (Morris and Anderson, 2015). At the same time, vlogging can be seen as a more authentic medium than traditional media outlets like radio and television, especially because the vloggers are ordinary people. They usually make their videos by themselves, as themselves, in their own environments (like a bedroom or office), and they address their audience directly rather

than broadcasting from impersonal studio sets and speaking in a standard variety, which can be perceived as inauthentic (Morris and Anderson, 2015). YouTubers are mostly also aware of general characteristics of their audience which allows them to tailor content specifically for this audience (see Section 4.1.5 for an overview of the available demographic statistics on the platform).

Addressing the audience directly also encourages feedback and interaction (Burgess and Green, 2009a), which is often also referred to as “participatory culture” (Burgess and Green, 2009b, p. 10): it is not only about self-promotion, but also social networking. Vlogs (and other genres like tutorials and gaming) are furthermore “down to earth, rather than unreachable for normal audiences” (Zhang and Bi, 2018, p. 86) which may also serve as an incentive for viewers to become content creators themselves. The audience is also directly involved in the communication: they can like and dislike the videos (i.e., click a thumbs-up or thumbs-down button to indicate if they enjoyed the video), comment and enter conversations with other users, but also influence future content by giving suggestions (Zhang, 2018).

Vlogs and vlogging channels allow for “a convergence between user-generated and advertiser-friendly business models” (Arthurs et al., 2018, p. 4). They are therefore bridging the gap between entertainment and business.

4.1.4 Becoming successful on YouTube

It is very difficult to become successful on YouTube. In 2016, an estimated minimum of 20 percent of all videos, and almost 90 percent of video views were uploaded and gathered by “the top 3% most viewed channels” (Bärtl, 2018, p. 26). Compared to the number of videos on the platform, the overwhelming majority are not viewed in relevant numbers, and if videos are viewed, they collect most attention in the first few days after the publication of the video (Bärtl, 2018). In early 2022, about 29,000 channels had over 1 million subscribers, and the vast majority of channels (around an estimated 30 million) have one hundred or fewer subscribers, though this number is likely much larger since social statistics providers (e.g., Socialblade.com) tend to not count channels with less than five subscribers (Funk, 2020).

The difficulty for content creators is that the algorithm behind the system that recommends videos to potential new viewers is “black-boxed” and therefore there is no official information about which variables are actually included in the algorithm’s decision-making process (Bishop, 2018, 71f., see also Arthurs et al., 2018). Many changes have been reverse-engineered, and one relevant element seems to be to increase the amount of time a user spends on the platform (see Chi, 2021 for an overview).

4.1.5 YouTube statistics and audience demographics

The currency on YouTube are its statistics—the number of likes, views, subscribers, and comments—and they are often seen as equivalent or descriptive of a content creator’s reputation and business value. The climate on YouTube is therefore also called a “like economy” (Gerlitz and Helmond, 2013, p. 1349; see also Arthurs et al., 2018). A YouTuber’s statistics-based reputation can then be used “to seek external outcomes such as sponsorship deals and advertising revenue and, for a few, paid work in the traditional media or the wider promotional ecosystem” (Arthurs et al., 2018, p. 9), that is, to work directly for the sponsors. The second element of YouTube currency is attention, since having the attention of a viewer is the only way to sell what is being advertised (Kyncl and Peyvan, 2017; see also Raun, 2018).

In terms of using YouTube metrics as facsimiles of popularity, Ha (2018a) suggests using the subscriber count rather than views. While views fluctuate depending on the video (one viral video might blow the view count of a channel out of proportion), “[subscription] represents regular viewership, interest in and commitment to the channel from the users” (Ha, 2018a, p. 137). Subscriptions also provide demographic statistics about the viewers (origin, gender, age, etc.) that YouTubers can take into account when creating content or tailoring advertising (Ha, 2018a). However, the number of likes and comments can predict the number of views—the more likes/comments, the more views—“because the more the audience like the video, the more they will repeat watching it (boosting the number of views) and perhaps sharing it to their friends to encourage them to watch it” (Ha, 2018a, 142f.).

However, high like and view numbers do not necessarily equate persuasion of the viewer by the video, but rather that the audience needs to be understood in order to understand the metrics (Ha, 2018b). According to Ha (2018b), most audience members are completely passive users. For example, a majority of users in a sample never reads or posts comments (Wen, 2018). However, viewers tend to watch diverse video genres, for example not only vlogs, but also news, comedy, tutorials, TV shows, etc. (Fisher and Ha, 2018). In general, Ha (2018b) suggests that there are five audience types plus “professional media companies and marketers” (p. 7) that work together for YouTube’s success:

the prosumers [(i.e. users who are at the same time producers *and* consumers)] who provide the truly diverse user-generated content on YouTube, the active commentators who express their views on the videos they watch and influence others’ perceptions of the video, the sharers who help spread the videos, the passive viewers who support the YouTubers with their regular viewership and the ad hoc viewers who boost the viewership of certain videos to a new height and help those niche videos to get viewership. (Ha, 2018b, p. 7)

In terms of demographics, a study by Abuljadail (2018) found that about 80 per-

cent of male study participants indicated that they regularly use YouTube, but almost 90 percent of women indicated the same, which may suggest that more women use YouTube. This has to be treated cautiously, though, as the sample was also female-biased. According to Shepherd (2023), roughly 56 percent of YouTube users are classified as male, and 44 percent as female, suggesting the ratio is much more balanced (though not completely representative either, as non-binary gender identities are not included in these statistics), and viewers of all ages are present on the platform.

Overall, only 69 percent of viewers are subscribed to YouTube channels, while 31 percent seem to “select channels to watch as needed or follow other algorithms such as YouTube’s recommended videos” (Fisher and Ha, 2018, p. 36). Most of the subscribed users (35 percent) are subscribed to five or less channels, while 29 percent of subscribed viewers have 25 or more channel subscriptions (Fisher and Ha, 2018). Additionally, it is likely that if an audience member is subscribed to a channel, they also use and watch YouTube on a day-to-day basis (Ha, 2018a).

Since YouTubers have access to the demographic information of their subscribers like gender, age and rough origin, it is likely to assume that they are actively aware of a relevant part of their audience, even though the majority of the audience is usually not subscribed. It is therefore also likely that YouTubers tend to adjust their content (and perhaps also their speech) to this vaguely defined group of viewers, though this is difficult to prove without interviews with professional YouTubers. Additionally, content creator Daniel Howell mentions in a panel discussion that YouTubers start by creating content that they want to create for themselves and by doing that gather an audience; and once they have an established audience they also have an established type of content which is tied to the audience, but still leaves room for personal experiments since the main goal of a content creator should be to remain true to themselves (Edinburgh TV Festival, 2017).

4.2 YouTubers as leaders: Charisma online

YouTubers were chosen as the subject of this study because they represent a middle ground between entertainers, celebrities, and business people (though these different roles are likely not balanced but can also not be quantified for specific YouTubers as the entirety of their businesses is not always known). As such, they would be classified as what Tur et al. (2022) refer to as informal leaders—leaders who have no formal authority over their followers (see also Section 2.3.4). Charisma (aside from topic, appearance, and body language) can create a connection between content creator and audience. That means that YouTubers could feel more like friends, i.e. approachable and authentic. At the same time, though, they entertain an audience (which is mostly also perceived that way by the audience), and

they have the opportunity to earn a living from that connection, either via ad revenue or profit percentages from purchases of followers of sponsored content which makes them entrepreneurs, as was also discussed in Section 4.1. Furthermore, according to Burgess and Green (2009a), YouTube revolutionized and more or less removed the separation between producer and consumer of traditional media, but at the same time opened up “dynamic and emergent relations between market and non-market, social and economic activity”.

4.2.1 YouTube and authority

Mixdorff et al. (2018) write that “[attracting] attention as well as gaining and persuading followers without having any formal authority is the essence of charisma” (p. 814). This definition fits the content creators included in this dissertation. They all started publishing videos before YouTube was a mainstream platform that provided a way for YouTubers to earn money with their creations. Their channels grew over time, corresponding to the “attracting attention” and “gaining followers” parts of the definition. Yet, content creators are rarely trained entertainers or business people, and still managed to build businesses based off their YouTube success—the reason for calling them ‘YouTube entrepreneurs’ in the title of this project. They do not attract attention and followers by having authority over them, but by trying to be role models and inspirations—sometimes their videos are even a means of escape from real life for their viewers. YouTubers can be seen as informal leaders, which sets them apart from formal leaders such as, for instance, politicians (Tur et al., 2022; see also Section 2.3.4).

This lack of authority also means that part of charisma and its associated attributes—for example, motivating or persuading followers to act a certain way—is perhaps not at the immediate forefront on YouTube. This aspect of charisma is without a doubt activated when the content creators promote their merchandise or products, and ask their followers to like a video, comment on it and/or subscribe, but otherwise mainly the part of charisma responsible for attracting attention is used. Therefore, there might be differences in how charismatic speech is perceived on YouTube compared to other areas like politics and business.

4.2.2 YouTube, authenticity, and intimacy

Authenticity is another attribute which is relevant for YouTubers. YouTubers can be classified as micro-celebrities, and as such they—according to Raun (2018)—need to “signal accessibility, availability, presence, and connectedness—and maybe most importantly authenticity—all of which presuppose and rely on some form of intimacy” (p. 99f.). Vloggers therefore usually also include sections in their videos that

are clearly meant to be interpreted as completely authentic—like bloopers, pets running around, imperfect lighting, framing or focus, etc. (Bishop, 2018).

Intimacy is also the major difference between micro-celebrities like YouTubers and mainstream celebrities: unlike mainstream celebrities like Hollywood stars, micro-celebrities like YouTubers are expected to share a large quantity of their lives and can only protect a small part of their privacy as they will otherwise lose attention and with it their status and success (Raun, 2018). Vloggers therefore use a degree of intimacy to connect to and engage with followers (Raun, 2018, see also Hou, 2019). This “ordinariness, intimacy, and equality by social media celebrities creates a sense of authenticity characterizing their videos” (Hou, 2019, p. 536). However, the authenticity is, at least partly, an act or performance: “authenticity on YouTube does not refer to a reflection of reality without mediation; instead, it is a specific means and content of representation” (Hou, 2019, p. 536).

It has to be kept in mind, though, that YouTubers are usually portraying a persona, which is a version of themselves, but strongly exaggerated. For example, Swedish YouTuber *PewDiePie* “says that he is ‘himself’ in the videos but with ‘100 percent energy’” (Lee, 2017, p. 29). Nowadays, YouTubers and vloggers tend to make this gap clearer (Lee, 2017), though it is not always directly expressed. Additionally, the portrayed intimacy and thereby authenticity is done for entertainment purposes, which most viewers understand. Hou (2019) calls this “staged authenticity” (p. 548) where the viewers are invited into the home and private lives of the content creators in their vlogs to show their authentic self, but the YouTuber is always in control of what and how much they share—and it is rarely their entire life, but rather some snapshots.

4.2.3 YouTube and enthusiasm

Not only are perceived authenticity and intimacy crucial in vlogs, the YouTuber also needs to appear as enthusiastic as possible to engage with the audience. This seems to be most important in the first few seconds of a video to get the audience interested and keep them watching. According to internet personalities Daniel Howell and Phil Lester, it is very important for YouTubers to be very open and energetic in the first 20 seconds. They write:

At first it might feel strange talking to no one, so imagine you are talking to a friend and you’ll come across natural to the audience! Try to be enthusiastic so you command people’s attention. The first 20 seconds or so are the most important so be energetic and to the point to draw your audience in! (Howell and Lester, 2015, p. 135)

YouTuber Alfie Deyes also mentions the importance of the first 10 seconds of a video to be captivating in order to keep the audience watching (Alfie Deyes Vlogs,

2020, April 19). In these 10 to 20 seconds, YouTubers have to convince the viewers to stay aboard and not click off to another video. In this short amount of time, persuasion cannot be achieved solely by content, but rather through appearance, body language, gestures, overall state of mind, and, of course, the voice (Reh et al., 2017; Caspi et al., 2019; see also Section 2.4).

4.2.4 YouTube and charisma

There are not many studies that deal with charisma and YouTube directly. Conde et al. (2020) in their study of the charisma of Spanish YouTubers call YouTubers “audience leaders” and say that both the content of the videos and the way this content is presented (e.g., based on the language) catches the attention of an audience—specifically subscribers—and leads to interactions between YouTubers and audience members. They equate the interaction between viewers and YouTubers as evidence for the YouTubers charisma and leadership capacity.

According to Cocker and Cronin (2017, the second available study investigating YouTube and charisma), the celebrity status of YouTubers falls somewhere between traditional celebrities (who have “omnipresent” fame; p. 457) and people who are renowned because of specific accomplishments, their personality or appearance—and this is within a certain social group which concentrates the fame on a much smaller circle of followers and hinges on interactions between YouTubers and followers. This interaction between leaders and followers is also part of many charisma definitions (see Chapter 2). On YouTube, followers actively choose people to heighten to (micro-)celebrity status, and these people usually are initially normal viewers and producers themselves, but somehow “develop a charismatic-like appeal that attracts others to follow and cohere around them” (Cocker and Cronin, 2017, p. 458). Personality is initially more prevalent in grabbing an audience on YouTube than skill or talent (though of course a certain degree of talent with camera work, editing, and story-telling is always needed), but most importantly the audience takes part in creating the personality that the YouTuber chooses to express (Cocker and Cronin, 2017).

An idea behind vlogging and YouTube is to create communities of likeminded people, which is often done by “ritualized calls to action or calls of support among [...] community members” like consistent greetings across videos, nicknames for the followers which especially early subscribers tend to expect when clicking on a new video (Cocker and Cronin, 2017, p. 463). For example, content creator Daniel Howell used to begin all his videos with the phrase “Hello Internet!”, and Lilly Singh used to refer to her loyal fans as “Team Super” (see also Mingione, 2014 for more examples), which long-term viewers come to expect. The knowledge that viewers and YouTubers exist on the same platform and the possibility to inter-

act—either positively or negatively—in the comments “are critical in building and maintaining charismatic authority and a sense of community” (Cocker and Cronin, 2017, p. 463). The community is often engaged by coming together in charitable efforts, calls to action, and fundraisers (see, for example, Andersen-Peters, 2016).

Additionally, especially the YouTubers who started their channels early on are seen as revolutionizing the entertainment space and suddenly offering alternatives to traditional media. Being at the forefront of revolutionary changes is also part of the traditional charisma definition (see Cocker and Cronin, 2017). This leads followers to be impressed by the YouTubers and their achievements: watching their videos may have had a profound and positive impact on viewers’ lives or they see a content creator as “much more than a woman who sits in her room on a Friday and talks to the camera [...] [but rather] a business owner, a brand, a marketer, a working mother, a creator and so much more” (Cocker and Cronin, 2017, p. 464, who cite comments of unnamed followers of YouTuber Louise Pentland). They go on to say that those comments include new ideas, advancing forward and helping to reinvent (a part of) society (Cocker and Cronin, 2017).

However, the charismatic effect of YouTubers is also not sufficient indefinitely. As their channels grow more professional (in terms of production quality, sponsorships, advertising, etc.) the perception of them as authentic leaders can dip, and especially early subscribers can stop following. With the professionalization often comes an audience-perceived sense of “selling out”, despite followers usually being (implicitly and explicitly) made aware that making videos is a job and no longer just a hobby (see Cocker and Cronin, 2017). According to Roux (2008), followers have so-called “marketplace metacognition” which means that there is an “awareness individuals have about persuasion techniques, their relevance and effectiveness in convincing them, and their own susceptibility to these tactics” (p. 467).

Charisma on YouTube is after all perhaps best described as a charisma-based community, containing both a micro-celebrity/YouTuber and the followers/fans/audience. The YouTuber tends to be aware of their status and persona and tends to use it actively to promote themselves and their content. The followers/fans/audience take an active role in the process of creating a micro-celebrity by endorsing the personality of the YouTuber and “[promoting] the longevity of the central personality” (Cocker and Cronin, 2017, p. 468).

4.3 YouTube voice

The voice of YouTubers has been extensively discussed in the traditional media as well as in social media after an article appeared in *The Atlantic* in December 2015 (Beck, 2015; see also Green, 2015, Hacker News, 2015, Dredge, 2016, Hagi, 2017,

Jennings, 2021). Many YouTubers employ similar speaking styles that the original article describes as “bouncy”. The author of the article cites personal communication with linguistics professor Naomi Baron who suggested features of the so-called “YouTube voice” which according to Lee (2017) has the purpose “to be performative and to attract more viewers” (p. 28).

4.3.1 The features of “YouTube voice”

For the *Atlantic* article, Baron identified five features that she attributes to “YouTube voice”. The first feature is “overstressed vowels”; in other words, hyperarticulation of typically schwa sounds in order to emphasize the word. The next component is mentioned as “sneaky extra vowels between consonants”, meaning an epenthetic vowel that creates emphasis on a word by adding another syllable. “Long vowels” is the next feature. Vowel lengthening “is a common way of emphasizing words” (Beck, 2015, see Section 7.2 for an introduction to positive intensification accents, the type of emphasis mentioned here) and is often used functionally. This lengthening can be very noticeable, but it can also be subtle, just a slightly longer vowel duration than normal which can lead (according to Beck, 2015) to the bounciness of YouTuber’s voices. Presentational coach David JP Phillips also mentions long vowels as part of YouTuber *Markiplier*’s speaking style (David JP Phillips, 2020). Consonant lengthening is also mentioned as a linguistic component of YouTube voice, especially word- or syllable initially. This is also a known type of emphatic accents—reinforcements—that is also briefly introduced in Section 7.2. The final feature that is mentioned in connection with YouTube voice is increased aspiration with voiceless consonants. As Beck (2015) writes: “There’s normally an aspiration on the K, even if you say it normally, but if you huff and puff a little more, that makes the word stand out”. In other words, YouTubers frequently use a longer and stronger aspiration of plosives for emphasis.

The general gist of the articles dealing with “YouTube voice” is that this speaking style consists of different emphasis strategies that are also used elsewhere, but with a higher frequency on YouTube (Beck, 2015). Using emphasis helps to gain and (hopefully) keep the attention of an audience, especially if the YouTuber only sits in front of the camera talking without any additional content that creates diversity in the video. For example, previous research suggests that emphatic speaking styles show that the speakers are involved with an interaction, and emphasis strategies are (among others) connected to “conversational story-telling” and in particular marking important passages (Selting, 1994, p. 384). Selting (1994) also found that emphasis is fairly consistently perceived by listeners, suggesting that these strategies also trigger the attention of listeners.

In addition to different ways of emphasizing individual words, YouTu-

bers—especially in their “talking alone to the camera” sections—often use variation of speech tempo (i.e., speech rate or articulation rate). Varying the tempo between the individual utterances or within an utterance by employing some of the emphasis techniques mentioned above that are based on lengthening creates an expressive, interesting and engaging melody that can help with keeping the audience watching. Overly expressive gestures can also be found in some YouTube videos. In general, these features in combination seem to appear most frequently

in videos where the people are just talking to the camera as themselves, with no acting, no props, no action. And in videos where people monologue for a minute, and then break away into a sketch or a scene [...], the tics [(i.e., the vocal behaviours used for emphasis and variation)], if they're there, seem far less pronounced than when the person was speaking directly to the camera. It's a 'talking to the audience' voice. (Beck, 2015)

4.3.2 Comparisons with other speaker groups

In the *Atlantic* article, linguist Naomi Baron compares YouTube voice with infotainment newscasting like late night shows. Another linguist, Mark Liberman, compares YouTubers' speaking style with the style of “carnival barkers”, only with a somewhat muted intensity (Beck, 2015). YouTubers have to keep the attention of their audience, but unlike an audience at a carnival, the YouTube audience comes to a specific video by choice which calls for less extreme vocal methods than carnival barkers. Liberman further calls YouTube voice “intellectual used-car-salesman voice” (Beck, 2015). These vocal tactics are a trend that is used especially in—but is not limited to—the genre of YouTube vlogs. The usage of these features divides listeners as some are attracted and pulled into a video by speakers using them, and many are annoyed and actively turned away (for example, see the discussion thread in Hacker News, 2015).

But while emphasis strategies might be shared between YouTubers and other speaker groups, and more importantly other entertainers like late show hosts, not all strategies seem to work across platforms. On his channel *The Film Theorists*, YouTuber Matthew Patrick investigated why the YouTube channels of most late night shows were losing subscribers during quarantine in 2020 after they were forced to move their shows from TV to YouTube. Patrick came to the conclusion that the late night hosts did not accommodate to the speaking style that is customary on YouTube. They were talking with their TV voices, especially using long pauses to elicit laughter when something is (meant to be) funny—which works fine if you have a live studio audience, but appears awkward when you are by yourself talking to a camera like a ‘regular’ YouTuber. As Patrick puts it in his video, “the internet runs at a mile a minute, and pauses in this universe are used to show that

something is awkward, not that something is funny. You let it sit and hang because it's uncomfortable" (The Film Theorists, 2020, min. 7:10ff.). Instead, pauses on YouTube are short or often replaced by jumpcuts that make the video a bit choppy, but even more fast-paced. *Jumpcuts* are cuts where parts of the video are cut out but part of a sound is followed by the next sound, usually the beginning of a new utterance. The transition between one utterance and the next becomes very abrupt and noticeable (both visually and audibly) with this technique.

4.3.3 Influence of video type on YouTube voice

Research suggests that the type of a YouTuber's video changes the way they speak. Lee (2017) investigated the vowels in four different types of videos of one speaker—Phil Lester (*AmazingPhil*). Lee studied formant differences between a solo vlog, a collaborative vlog, a gaming video and a live video. These four styles of video can be placed on a spontaneity continuum with a live video being most spontaneous as the speaker reacts to comments and questions submitted in real time by the audience, and the solo vlog being least spontaneous because it is more planned out and there are no external inputs to react to (Lee, 2017).

Lee analyzed two features that differentiate Northern English varieties (Lester's "home variety") from more standard varieties like Standard Southern British English (SSBE): the so-called TRAP/BATH split and the FOOT/STRUT merger—in both cases, the vowels are pronounced the same in the North but differently in the South. The study found that Lester "shows more evidence of the FOOT/STRUT merger in more spontaneous contexts [(live and gaming videos)]. This is reversed in less spontaneous contexts [(solo and collaborative vlogs)], manifested through more SSBE-like F1 values for FOOT/STRUT" (Lee, 2017, p. 35). That suggests that Lester speaks slightly more "Northern" or regional in more interactive videos and that he perhaps pays more attention to the way he is speaking in the less spontaneous vlogs—maybe to be as intelligible and accessible as possible in the videos that gather the largest and therefore more diverse audiences (Lee, 2017). Lee's final conclusion is that

Lester style-shifts based on the context of the video—not because of audience design or a speaker designed performative register, but because of increased attention paid to speech in more scripted contexts. (Lee, 2017, p. 37)

Audience design would mean that the speaker reacts and adjusts to the way they perceive their audience (Lee, 2017, based on Bell, 1992). This cannot account for the style-shifting as it is assumed that the audiences of the different videos are not so different from each other, in fact the audience likely overlaps as live and gaming videos can be seen as supplementary to the vlogs (Lee, 2017). Speaker design

means a performative register of speech that is similar or aligned with the audience, though this is also unlikely as the audience on YouTube is largely anonymous (Lee, 2017, see also Coupland, 1984). Rather than these two aspects, Lee suggests that Labov's model of attention paid to speech can explain the style-shifting. The attention-paid-to-speech model suggests that the less attention a speaker pays to their speech, the more informal or casual the speech style becomes; in comparison, the more attention is paid to speech, the more formal and prestigious is the resulting speaking style (Labov, 1972; Lee, 2017).

4.3.4 Vocal effort and charisma (on YouTube)

The previous sections on "YouTube voice" show that this specific style of speaking relies heavily on emphasis strategies (Beck, 2015), but also on orienting vocal productions towards the audience. For many of the features mentioned above, a connection to several phonetic theories has been made, like the Frequency Code (Ohala, 1984), the Effort Code (Gussenhoven, 2002, 2016; Chen et al., 2002), the H&H theory (Lindblom, 1990), as well as some of Labov's principles of linguistic methodology (Labov, 1972). These theories can not only be connected to features of "YouTube voice", but also to charismatic speech.

The Frequency Code is by now also viewed critically (e.g., Winter et al., 2021). It says in general that higher F₀/higher voices tend to be perceived as smaller, submissive (and feminine, though this interpretation is not considered to be relevant or culturally appropriate in the current study), while lower F₀/lower voices are perceived as larger and dominant (see Ohala, 1984). This Code is connected to the size of vocal folds, larynx, and vocal tract: larger sizes here tend to correlate with lower voices, though the perceived size can also be deceiving. The only aspect of charismatic speech found in the previous literature where an application of the Frequency Code might be of value is the high pitch level that is perceived as more charismatic. Niebuhr et al. (2016a) suggest that this high level could be seen as submissiveness, or more specifically, a lack of authority, and together with other features be turned into charisma. However, since pitch level is strongly influenced by other pitch features like pitch range, pitch variability, and the height of pitch accents, a direct interpretation in terms of the Frequency Code seems to be less appropriate.

Other theories, especially the Effort Code and the H&H theory, seem more appropriate for charismatic speech in general and charismatic speech on YouTube in particular. According to Beck (2015), the main features of "YouTube voice" are connected to emphasis. Producing schwas (central vowels) as other, more peripheral vowel qualities, strong aspirations of voiceless plosives, vowel and consonant lengthening, as well as epenthetic vowels can all be seen as hyperspeech. Hyper-

speech is output-oriented and focuses on segments being as distinctive from each other as possible to make understanding as easy as possible for the listeners (Lindblom, 1990). However, in order to not lose authenticity and sincerity, and to remain with an approachable impression, content creators on YouTube also cannot move too strongly in the hyperspeech direction, but still have to use phonetic reductions to sound natural. That would be hypospeech, and—like every speaker—YouTubers need to move and vary between the two extremes on the continuum. It is likely that content creators will show a moderate amount of reduction to appear as sincere and relaxed, but clearly understood (see also Niebuhr, 2017; Michalsky and Niebuhr, 2019).

Hyperarticulation of vowels and consonants require increased articulatory effort of the speaker. Similarly, other emphasis strategies like high and late pitch peaks (and connected to this a larger pitch range) and variation of speech rate from phrase to phrase also require more effort from the speaker. These features are partly also suggested to be characteristics of “YouTuber voice” (Beck, 2015), and they are seen as characteristics of charismatic speech (see discussion above in Section 3.2). The Effort Code can explain this, since Gussenhoven (2002) states that the more effort a speaker invests in their speech production, the more precise are the articulations, but also pitch movements get wider. This is mainly interpreted as emphasis in order to make a message abundantly clear to listeners (Gussenhoven, 2002). It is about prominence, which can be achieved either by higher pitch peaks, or later pitch peaks (or stronger segmental articulations). For example, a high peak is usually later than a low peak, as the speaker’s vocal folds need to reach the needed speed, which takes some time. A later peak can therefore mimic the height and therefore prominence of a peak (Gussenhoven, 2002; see also Chen et al., 2002).

Finally, we have to assume that three of Labov’s principles of linguistic methodology (1972) can also be applied to charismatic speech, and in general speech on YouTube. The first is the “principle of style shifting” which suggests that all speakers use different speaking styles depending on the situation. This study investigates a specific speaking style, which is the style found in YouTube vlogs. This principle also operates under the assumption that specific listeners only experience part of what a speaker is able to do linguistically (Labov, 1972) which means—in the context of YouTube—that we only see the part and style of the content creators they want us to see. Secondly, and perhaps most importantly, there is the “principle of attention” which says that speaking styles are ordered on a continuum from formal at one end to casual at the other end depending on how much attention a speaker pays to their speech. If the speaker pays much attention and speaks carefully, this automatically results in a more formal speaking style (see Labov, 1972). On YouTube, the speaking style in vlogs would be considered to be placed towards the casual end of the continuum, but not at the extreme as they are planned

(though not scripted) and mostly edited (see also Lee, 2017). Finally, the “principle of formality” suggests that once a speaker is observed (and this can merely be the presence of a recording device) the speaker pays more attention to their speech and already moves the speaking style slightly away from casual towards formal (Labov, 1972). Since content creators are consciously talking into a camera lens and microphone, fully aware of the potential size of their audience, it is likely that this also has an effect on the speaking style and with it the precision and effort used for articulation.

4.4 Summary

YouTube as a video-sharing platform is based on participatory culture, which means it depends on communication and interaction between content creators and viewers. These interactions can take place actively through comments, subscriptions, and leaving likes, or passively through simply watching the content. YouTubers often also directly react to interactions like comments, for example by basing a future video on a viewer’s suggestion and crediting them for the idea. The content creators also call viewers to action, i.e. they try to elicit interaction.

Content creators like vloggers are also building and leading communities that come together not only to enjoy the YouTuber’s content, but frequently also to join together for charitable efforts. To build a community, YouTubers often make use of ritualized greetings and nicknames for their particular audience (see Cocker and Cronin, 2017; Mingione, 2014). This community can be referred to as a charismatic community where content creator and audience work together to create an engaging persona (see Cocker and Cronin, 2017; Belk, 2013).

Important attributes on YouTube—especially in vlogs—are authenticity, intimacy, enthusiasm, and a lack of authority over the viewers. However, authenticity and intimacy tend to be staged and only reflect the small portion of a vlogger’s life that they *choose* to share. The format of a vlog where one person is speaking directly to the audience creates a feeling of intimacy, whilst simultaneously encouraging interaction within the online community.

In terms of phonetic characteristics of vloggers—or “YouTube voice”—there are a few features that are mentioned in the media and in the literature. “YouTube voice” uses different strategies for emphasizing words (long vowels and consonants, strongly aspirated plosives), speech tempo variation (see, e.g., Beck, 2015), short pauses (The Film Theorists, 2020), and more standard varieties in less spontaneous contexts like vlogs (Lee, 2017).

Chapter 5

Research questions and hypotheses

Based on the findings of previous research on charisma, charismatic speech and speaker perception, and YouTube (see Chapters 2, 3, and 4), this thesis investigates several elements of charismatic speech and its acoustic-prosodic characteristics and how these characteristics are perceived for a sample of English-speaking YouTubers from North America and England (see Chapter 6 for a data overview). This section introduces the research questions and some general hypotheses for the entire project. This chapter only introduces general, overarching hypotheses. Each of the empirical chapters (Chapters 8 through 10) then includes specific hypotheses referring to the relevant acoustic-prosodic features investigated in that particular experiment. The overarching hypotheses and the research questions are guiding the entire investigation and are the basis for the specific hypotheses.

The first research question RQ1 (see below) is concerned with the specific acoustic parameters and how they should be configured in the YouTube context for a positive perception of five investigated attributes: *charismatic* for a direct charisma rating, and *authentic*, *enthusiastic*, *likable*, and *persuasive* as charisma-adjacent attributes. These attributes have been used in previous studies on charisma and charismatic speech (e.g., Rosenberg and Hirschberg, 2009; Grabo et al., 2017; see also Section 2.1.3), and they are relevant for interactions and communication on YouTube (see Section 4.2). In the experiments, the attributes are paired with audio stimuli and rated based by participants on the voices. The audio stimuli for the experiments are based on digital manipulations or modifications of the original stimuli. For the investigation of the so-called “prosodic manipulations”, four acoustic parameters that have been shown to be relevant for charismatic speech (see Sections 3.2.1 and 3.2.4) are manipulated: the pitch level, pitch range, and speech rate are increased and decreased, and the final contour direction is changed (possibilities: rising, falling, plateau). They are presented with the unmodified stimuli to the participants (see Section 8.3 for the procedure behind the stimulus creation and the experiment design). For the investigation of the so-called “pause manipulations”, pause durations in other stimuli were lengthened or shortened, and breathing noises were added or removed (see Section 9.3). The first research question will be extensively investigated in the those two chapters (Chapter 8 and Chapter 9),

though findings from the third experimental chapter (Chapter 10) will also add to this. The general hypothesis for the most likely acoustic feature characteristics to evoke the perception of charisma is provided in H1 below. Possible differences in ratings of the five attributes depending on the manipulation are analyzed in order to see if specific manipulations trigger different aspects of charisma.

RQ1: *How should acoustic parameters be configured to be perceived as charismatic (both in terms of charisma directly and charisma-adjacent attributes) in the context of YouTube vlogs?*

H1: *Stimuli with larger pitch ranges, higher pitch level, medium speech rates, and non-rising phrase-final pitch contours on the one hand, and audible breathing and shorter pauses on the other hand are perceived as more charismatic.*

The first two experiment chapters therefore investigate the influence of specific acoustic-prosodic features in a very controlled manner through manipulation. Chapter 10 then combines the results from the perception experiments with the acoustic measurements of the unmodified stimuli. The acoustic measurements are correlated with the mean ratings of the different attributes, and the acoustic measurements are also ranked by value for each speaker for visual correlations with the ratings (see Section 10.3.3). This method can show if ratings and acoustics are connected in the sample, and how they might be connected. A selection of acoustic-prosodic features that have been shown to be relevant for charismatic speech is investigated (pitch level, pitch variability, pitch range, maximum and minimum F0, pitch peak timing, prominences in general and emphatic accents in particular, speech rate, and phrase duration; see Chapter 3 for more information on the connection of these and other features to charismatic speech; see Chapter 10 for the reasoning behind choosing these features and excluding others). This part of the project focuses on the second research question which investigates if there is a connection between higher perception ratings (in terms of charisma directly and the four charisma-adjacent attributes) and the acoustic feature values known to be indicators of charismatic speech in the stimuli they rated (see RQ2). It is predicted that the higher a speaker/a stimulus is rated, the more the acoustic feature values align with what is known about charismatic speech (see H2). The specific predictions for correlation directions are included in Section 10.2.

RQ2: *Do speakers in the sample who receive higher ratings for charisma and/or charisma-adjacent attributes in the perception studies employ acoustic features in ways that have been reported in other charisma literature?*

H2: *The ratings from the experiments correlate with the acoustic feature values known to be used in charismatic speech and related attributes.*

In the perception experiments, the participants' familiarity with the speakers was also collected in conjunction with the direct charisma ratings, which allows the

investigation to look into the connection between familiarity and charisma perception. The participants were asked to indicate how familiar they were with the speaker before they came into the experiment session (possible response options: “I know the speaker”, “They seem familiar”, “Unsure” to “I do not know the speaker”). They were specifically instructed to not indicate if they recognized the speaker from previous stimuli. The relationship between charisma and the familiarity with the speakers (as suggested by, e.g., Rosenberg and Hirschberg, 2005, 2009; see Section 3.3.5 for an overview) is part of the third research question of this thesis (see RQ3 below). It is predicted that more familiar speakers would be perceived as more charismatic (see H3 below).

RQ3: *Is there a connection between charisma ratings and familiarity with the speakers in the sample?*

H3: *The more familiar a speaker is to the listeners, the more charismatic they are perceived.*

The empirical chapters additionally investigate the influence of speaker gender and speaker origin. At the time of writing, there are no phonetic studies on charismatic speech available that research British English varieties in combination with charisma. All results that investigate English refer to American English varieties (see Chapter 3). Likewise, there are only few studies that investigate charismatic speech and female speakers, though the number is growing, and generally suggests that male speakers tend to receive higher charisma ratings than female speakers (e.g., Jokisch et al., 2018; Niebuhr et al., 2019; see Section 3.3.3). Therefore, this current project has the opportunity to compare acoustic features and their charisma perception for both speakers from North America and England as well as male and female speakers. These factors will be considered in the interpretation of the results.

Part II

General Methodology

Chapter 6

Data selection

6.1 Speaker demographics

The present study includes ten English-speaking YouTubers. Five of them are from North America (NAM), more specifically from the United States of America (US, $N=4$) and Canada (CA, $N=1$), and five speakers are from England (ENG). Five speakers are male (two from NAM, three from ENG), and five are female (three from NAM, two from ENG). All speakers were between 24 and 32 years old when the respective video used for the analyses was published. All have been active on YouTube at least since 2015. Table 6.1 lists the speakers with their ages, gender¹, and country of origin. The abbreviations in the speaker column will be used throughout the study.

Speakers from two broad regional varieties of English were chosen in order to explore possible influences of regional variety on charisma perception. Several studies have shown that charisma perception differs cross-linguistically with different languages (e.g., Biadisy et al., 2007, 2008; Signorello et al., 2012a, 2012b; D'Errico et al., 2013; Gutnyk et al., 2019), and that might also be the case with regional varieties of the same language. In a general sense, England has more diverse regional varieties than North America, stemming from the shorter period of time that English was spoken on the continent (around 300 years in North America versus around 1,500 years in England, see Trudgill, 2000). Additionally, it has been suggested that YouTubers tend to speak in a less or non-regional variety in their videos in order to be understood by a wide range of listeners with different language backgrounds (Bishop, 2018). The speech of YouTubers might also be less regional in vlogs, intended to reach a larger audience, than in a live show directly interacting with a smaller group of followers—meaning the context has an influence on how much regionality is used (Lee, 2017, see also Section 4.3.3).

With this information in mind, the current study will only address differences between two broader varieties of English (i.e., North American English and En-

¹This project uses the terms “male” and “female” to refer to socially constructed and performed identities that may not correspond to biological sex. The project makes no assumption of the biological sex of the speakers.

Table 6.1: Age, gender, country, and general origin information of the YouTubers in the sample. Age represents the age of the speakers at the time of video publication. Abbreviations: ENG = England, NAM = North America.

| Speaker | Age | Gender | Country | Origin |
|---------|-----|--------|---------|--------|
| LP | 31 | female | England | ENG |
| ZS | 27 | female | England | ENG |
| AD | 24 | male | England | ENG |
| DH | 26 | male | England | ENG |
| PL | 31 | male | England | ENG |
| CB | 31 | female | US | NAM |
| LS | 29 | female | Canada | NAM |
| SP | 31 | female | US | NAM |
| MF | 29 | male | US | NAM |
| MP | 32 | male | US | NAM |

glish from England). It is assumed that the speakers have adjusted to a more standardized variety (e.g., Lee, 2017; Bishop, 2018). However, while there is much information on charismatic speech of American speakers (which allows for comparison with the present sample; see, among others, Rosenberg and Hirschberg, 2005, 2009; Novák-Tót et al., 2017; Niebuhr et al., 2020a), as well as speakers from other languages (e.g., Signorello et al., 2012a, Signorello et al., 2012b and D’Errico et al., 2013 for Italian and French; Jokisch et al., 2018 for German; Biadys et al., 2007, Biadys et al., 2008 for Palestinian Arabic and Swedish; Gutnyk et al., 2019 for German, Ukrainian, Mandarin Chinese, and Spanish), there is currently no research available that looks at features of charismatic speech in varieties of British English (see Chapter 3 for an overview). Therefore, this project is a broader starting point which can be narrowed in the future. Other varieties of English (Australian, New Zealand, Ireland, India, etc.) are not included in the present project to have a limited number of varieties. Like British English, there is no research on other varieties of English and charismatic speech available. Therefore, the varieties were restricted to one previously researched variety (American English) and one not previously researched variety (British English from England).

6.2 Speaker selection

According to Funk (2020), there are over 51 million channels on YouTube, with most of them having fewer than 10 subscribers. In order to narrow down from 51 million channels, selection criteria were developed to choose the channels/speakers for the project.

All of the YouTubers in the sample had to have well above one million subscribers on the channel that was chosen. However, there was also supposed to be a range of subscriber numbers between all speakers in the sample in order to be able to investigate possible correlations between the acoustic features and the subscriber

count. Ha (2018b) suggests subscriber count as an approximation of popularity on YouTube, as it is an active choice of an audience member to decide to follow the publications of a specific YouTuber and showing interest in seeing their future work. Additionally, subscriptions also help content creators to tailor their content to their audience, as they receive demographic information about their subscribers (Ha, 2018b; see also Section 4.1.5).

The content creators in the sample also had to have started making videos on YouTube before being a “YouTuber” was even a possible career. All YouTubers in the sample have videos on their channels talking about why they started their channels—to have something to do, to connect with people worldwide, to show their skills while unemployed, to have fun (see, for example, Alfie Deyes, 2013; Markiplier, 2013; Daniel Howell, 2013; AmazingPhil, 2013; Lilly Singh, 2013; Zoella, 2013; The Game Theorists, 2013). And at some point, the hobby turned into career. As the channels grew, they started gaining revenue from advertisements and other opportunities (see Section 4.1.2).

All channels in the sample are what Funk (2020) calls “dinosaur” channels, the smallest group of all channels on the platform, but they can be considered superstars, at least in the context of the platform. He writes:

As of January 2022, there are around 29,000 YouTube channels out there having over 1 million subscribers. They are the dinosaurs of the YouTube space! Massive in size. Most specimens started ages ago. And it took them very long to grow this size. (Funk, 2020, paragraph 4; date of update to 2022 unknown)

Finally, all speakers needed to have additional businesses outside of the platform that developed as a result of the YouTubers’ success as content creators on YouTube. That—on top of the mere presence of a monetized channel—made them into entrepreneurs, be it sometimes “accidental entrepreneurs” (The Game Theorists, 2022, min. 15:21; see also Chapter 4 for further information). The YouTubers in the sample all have several of the following ventures: books, acting, TV and/or radio presenting, comedy shows, games, apps, clothing brands, production, merchandising and media consulting companies, and so on. The information on the speakers is displayed in Table 6.2.

Ten speakers were chosen in order to gain insights into both an equal amount of male and female speakers, as well as speakers from North America and England. Therefore, the project overall focuses more on groups than individual differences, although some individual observations are additionally included. This also offers the opportunity for more general interpretations compared to a case study with only one or two speakers.

It is likely that speaking style and prosody change over the years with growing

Table 6.2: An overview of the speakers and their channels in the sample, including the year of channel creation, the number of subscribers (= Subs.), accessible videos and views on the channel, as well as other channels run or created by the speakers. The subscriber, video and view numbers were collected November 22, 2022. (m = million, b = billion subscribers/views)

| Speaker | Channel | Channel created | Subs. | Videos | Views | Other channels |
|-----------|--------------------------|-----------------|-------|--------|-------|---|
| LP | <i>Louise Pentland</i> | 2010 | 2.22m | 532 | 191m | <i>SprinkleofChatter</i> |
| ZS | <i>Zoe Sugg</i> | 2012 | 4.93m | 635 | 976m | <i>Zoella</i> |
| AD | <i>Alfie Deyes Vlogs</i> | 2010 | 3.64m | 1,549 | 1.1b | <i>Alfie Deyes</i> <i>PointlessBlogGames</i> |
| DH | <i>Daniel Howell</i> | 2006 | 6.19m | 142 | 686m | <i>DanAndPhilGAMES</i> <i>danisnotinteresting</i> |
| PL | <i>AmazingPhil</i> | 2006 | 3.93m | 333 | 662m | <i>DanAndPhilGAMES</i> <i>LessAmazingPhil</i> |
| CB | <i>Colleen Ballinger</i> | 2006 | 8.71m | 1,088 | 1.9b | <i>Miranda Sings</i> <i>Colleen Vlogs</i> |
| LS | <i>Lilly Singh Vlogs</i> | 2011 | 2.74m | 1,543 | 454m | <i>Lilly Singh</i> |
| SP, MP | <i>GTLive</i> | 2015 | 2.96m | 1,140 | 659m | <i>The Game Theorists</i> <i>The Film Theorists</i> <i>The Food Theorists</i> <i>The Style Theorists</i> |
| MF | <i>Markiplier</i> | 2012 | 34m | 5,351 | 19b | — |

experience in public (and YouTube) speaking as well as a deeper knowledge of the audience and their interests, but also each content creator’s interests. Since the focus is not strictly on individual differences but rather group differences, it was not deemed necessary to include more than one video per speaker.

Additionally, one has to keep in mind that it is not the private person’s voice that is analyzed in this investigation. It is rather the voice of the public persona put forward by the speakers on YouTube. The voices are their own, but likely exaggerated compared to private life for entertainment (see Lee, 2017; Hou, 2019). Findings of the study regarding charisma and four of its related attributes (in this investigation, *authentic*, *enthusiastic*, *likeable*, and *persuasive*) can therefore only be related to the public persona, not the person.

6.3 Video selection

6.3.1 Video overview

As can be seen in Table 6.2 above, the number of videos on a channel ranges greatly between around 140 videos to well over 5,000 videos per channel. Each speaker appeared in one video each, with the exception of two speakers who appeared in the same video together. Therefore, nine videos were chosen for analysis, one from each represented channel. Table 6.3 lists the speakers and the used videos. For

Table 6.3: An overview of the speakers (= S) and videos included in this study with information on publication dates, as well as the reference and the timepoint within the video where the analysis starts (in seconds, rounded). The view count of the video is also provided (rounded to Thousands; data collected on November 22, 2022; the raw numbers as well as Likes of the video and Subscribers of the channel can be found in Table A.1 in Appendix A).

| S | Video title | Published | Reference | Start | Views |
|-----------|---|--------------|-------------------------|----------|-----------|
| LP | <i>I'm SO sorry IWD2017</i> | Mar 8, 2017 | Louise Pentland, 2017 | 0.00s | 427,000 |
| ZS | <i>2018 plans & re-united with my bestie</i> | Jan 29, 2018 | Zoe Sugg, 2018 | 155.90s | 2,437,000 |
| AD | <i>Am I happy making YouTube videos?</i> | Jan 18, 2018 | Alfie Deyes Vlogs, 2018 | 613.69s | 444,000 |
| DH | <i>Daniel & Depression</i> | Oct 11, 2017 | Daniel Howell, 2017 | 0.00s | 3,731,000 |
| PL | <i>Why I went to hospital</i> | Nov 15, 2018 | AmazingPhil, 2018 | 0.00s | 1,576,000 |
| CB | <i>My experience with Netflix</i> | Dec 31, 2017 | Colleen Ballinger, 2017 | 0.00s | 1,717,000 |
| LS | <i>We need to have an honest talk</i> | Dec 9, 2017 | Lilly Singh Vlogs, 2017 | 0.00s | 689,000 |
| SP, MP | <i>GTeaLive: Europe PASSED Article 13! Your memes are banned?</i> | Mar 27, 2019 | GTLive, 2019 | 1154.77s | 274,000 |
| MF | <i>Let's be completely honest</i> | Aug 19, 2018 | Markiplier, 2018 | 0.00s | 2,237,000 |

different reasons (see Section 6.3.3 below), the annotation and analysis of three of the videos could not start at the beginning of the videos, but a few minutes later. In order to be able to reconstruct and replicate the study, the starting times of the annotations used for analysis within each video are included here.

The view and like numbers of a video were not used as criteria for inclusion or exclusion of a video into the project. The statistical measures provided by YouTube (subscriber, view, and like counts) were collected and can be found—in both raw and normalized format—in Table A.1 in Appendix A.

6.3.2 Inclusion criteria

The videos were chosen with several criteria in mind. All videos were published between 2017 and 2019. The end of the publication year range was determined by the annotation process of the project. The start was determined so as to have a defined two-year period in which the videos were published. That ensured a similar state-of-the-art of recording equipment, although the exact recording equipment (camera, microphone, editing software and used processes) is unknown.

The videos had to be at least 10 minutes long to provide sufficient speech material for the analyses. In the end, only five minutes of material were analyzed per speaker. However, the minimum inclusion duration ensured that all five minutes

could be used from the same video, in case some sections had to be excluded (see Section 6.3.3).

The videos in the sample were (semi-)spontaneous vlog-style videos where the YouTuber speaks directly to the camera in a monologue (see also Section 4.1.3 for more details on vlogging as a video genre). It is unknown to what extent the videos are planned out. It is clear from the speakers' verbal behavior when watching the videos that the YouTubers speak freely and do not know what they are saying by heart, but the speech is also not produced completely spontaneously. They clearly know the points they want to address. Possibilities are that they prepared bullet points beforehand, or that they used a teleprompter, but this can only be hypothesized. The impression that they know exactly what to say while still appearing to be spontaneous may also be a case of subsequent editing and perhaps re-organizing of the video structure. None of the videos list an outside editor. It is therefore assumed that the YouTubers edited the videos themselves and also made the decisions of what to leave in or out of the final video themselves. The editing and presumed planning of the video make it impossible to categorize the speech style in the videos as completely spontaneous, but rather semi-spontaneous.

The topics of the videos ranged from mental health and health in general to business, and thoughts on YouTube as a platform and the engagement with an audience. In all videos, the audience was directly addressed—for example, in the form of questions, reactions to comments, or addressing concerns or wishes of the YouTubers. The content was not controlled for because the vlog-style of the video was deemed most important. This decision was also made in light of the results from political speech data (Rosenberg and Hirschberg, 2009) which showed that there was “no statistically significant impact [of topic] on subjects' ratings of charisma” (p. 645). Content is affected by aligning (or not aligning) ideals and values of listeners/viewers, though (see Section 3.3.1). However, since viewers on YouTube interact with a video by choice, this likely has less of an impact.

All videos that were chosen were filmed in one location and the speaker remained more or less stationary in that location. That way, there was no change in room acoustics and no noise because of movement. It is another reason why only one video per speaker was chosen: the recording equipment, recording environment and editing methods did not vary within the video. Variations within these factors likely would have influenced the acoustic analyses.

6.3.3 Exclusion criteria

There were also criteria for excluding sections of the chosen videos from the subsequent analyses. First of all, heavily scripted sequences were excluded as the speaking style differed strongly from the semi-spontaneous style of the rest of the vlogs.

Heavily scripted in this case means that the vlog cut away to sequences acted out in different locations, often involving props and sound effects. These mostly occurred in the DH's video.

Videos that had music playing in the background for the whole duration of the video were excluded, as background music severely impacts the acoustic analyses. Two of the videos (by DH and PL) still had short sections with background music and/or sound effects. These sections were also excluded from the analyses.

Many vlog-style videos feature the YouTuber moving around the house or outside, filming while moving. While this is part of the vlog genre, the sections where the speaker was mobile were excluded from the analyses. That way, constant changes in the environment and room acoustics were avoided as likely interactions with the sound quality of the videos and subsequently the acoustic analyses. This was the case for speakers AD and ZS, who filmed on the move in the beginning, but had long stationary sections later on in their videos. Thus, the analysis of the speech material does not start at the beginning of the video (see Table 6.3).

Finally, only monologue sections were used for analyses. Speakers MP and SP appeared in the same video and there are sections of the video where they talk to each other or overlap each other. Sections where each speaker was not addressing the audience directly, but talking to the other person in the room, or overlapping speech between the two were excluded. Thus, the analysis of the speech material starts later into the video and not at the beginning, as is shown in Table 6.3.

Chapter 7

Data treatment

7.1 Data preparation

The videos that were analyzed were downloaded from YouTube. The video files were converted into wav files using the program *FormatFactory* (FreeTime, 2021). The intensity was normalized to -3 dB in Audacity (AudacityTeam, 2017) to account for the different recording settings and equipment of the YouTubers.

English transcripts of the videos were also downloaded from each video. Some of the videos had transcripts manually prepared by viewers in different languages, but most transcripts were auto-generated by YouTube’s speech recognition software. These transcripts—viewer-made and auto-generated—are used on the platform as optional closed captions. All transcripts were checked for errors and corrected. The correction was necessary for the auto-generated transcripts in particular, as there were many problems in speech recognition.

Following their clean-up, the transcripts were uploaded as text files with the corresponding wav files of the videos to the online segmentation service WebMAUS (Kisler et al., 2017). WebMAUS takes the transcript and the audio and automatically creates and annotates intervals on three annotation levels (or tiers) using the Speech Assessment Methods Phonetic Alphabet (SAMPA). An annotation file (TextGrid) is created that can be downloaded. The first two tiers contain interval labels of individual words—one with orthographic transcription (“ORT”), the other with SAMPA transcription (“KAN”). The third tier contains SAMPA-annotated, individual segments (“MAU”). The audio files were also run through a Praat script, *Pause Detector* (de Jong and Wempe, 2009), which segments the signal based on sounding and silent stretches (i.e., IPUs and pauses), labels syllable nuclei, and saves a TextGrid file to work with further. The TextGrids from both applications were merged and re-ordered in order to work in one file.

7.2 Annotation for acoustic measurements

Annotations were made in Praat (Boersma and Weenink, 2018, version 6.0.37). Figure 7.1 shows the structure of the TextGrid used for acoustic measurements. Table

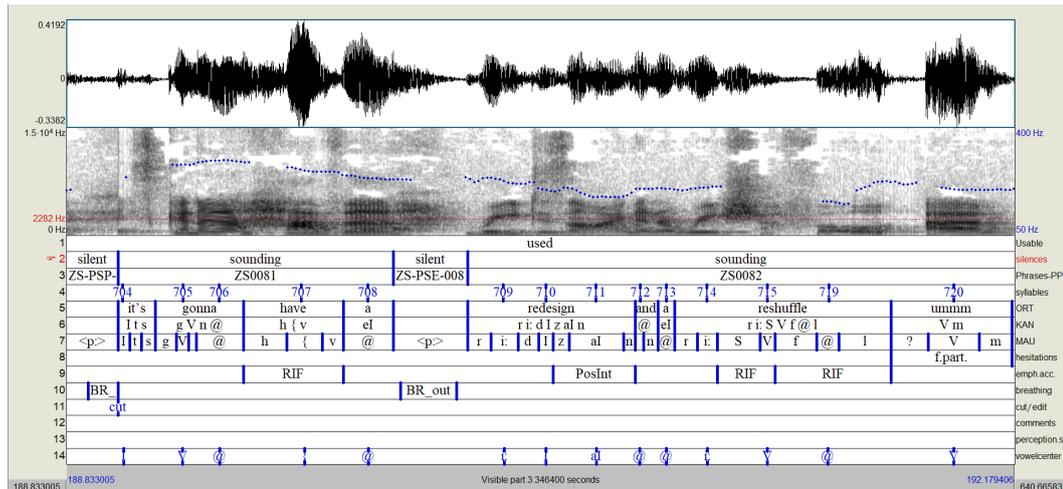


Figure 7.1: Example of the annotation tiers used for measurements and analyses.

7.1 provides an overview of the different annotation tiers. Subsequently, only tiers are explained in detail that are relevant for the following analyses and procedures. They are referred to by number. The other tiers that were annotated in the corpus but are not addressed in the experiments are briefly explained in Table 7.1.

On the first tier, portions of the audio that were eligible for analysis were marked for efficient analysis. Portions of the audio were not included in the analysis if a) there was music or sound effects in combination with the speaker’s voice, b) there were two speakers talking at the same time, or c) the video changed from a somewhat spontaneous style to scripted content (see Section 6.3.3). Sections to be used in the analyses were labeled “used”, and the acoustic analyses were only carried out within these sections.

Tier 2 is the phrase tier that was created by the *Pause Detector* script using the labels “sounding” for a phrase and “silent” for a pause. The boundaries were manually corrected and placed on positive zero crossings to avoid jumps in the measurements. In this project, phrases are defined as a cohesive stretch of speech with an overall F0 declination and phenomena like final lengthening or creaky voice. Most of the time, this fell together with pauses or cuts (i.e., IPUs), but sometimes phrase boundaries were also set in places without pauses where other features (declination, final lengthening, creaky voice) were prevalent (Peters, 2015), or a cut occurred between phrases.

Tier 3 is another phrase tier (referred to as the “Phrases-PP” tier). The phrase boundaries are identical to those of the *Pause Detector* phrase tier. The “Phrases-PP” tier was used for the measurements using ProsodyPro (Xu, 2013, version 5.7.8.1) and other scripts in order to receive phrase-level acoustic measurements (for further information on the measurements, see Section 7.4 below). The intervals on this tier were manually labeled. The labels for the phrases consist of the initials of

Table 7.1: Overview of the annotation tiers for the acoustic measurements with tier number, tier label, tier type and an explanation with possible labels.

| Tier | Tier label | Tier type | Explanation |
|------|---------------------|-----------|---|
| 1 | Usable | interval | Label “used” if section is included in analysis; Label “not used” if section is not included |
| 2 | Silences | interval | Label “sounding” for phrases; Label “silent” for pauses; created by <i>Pause Detector</i> script (de Jong and Wempe, 2009) |
| 3 | Phrases-PP | interval | Label for phrases: speaker abbreviation and running number; label for pauses: speaker abbreviation, pause type abbreviation, separate running number Pause type labels: PSP = silent pause at syntactic boundary; PSD = silent pause before/after discontinuity; PSH = silent pause for hesitation; PSE = silent pause for emphasis |
| 4 | Syllables | point | Labels: running number in vowel/syllabic consonant center of produced syllables; created by <i>Pause Detector</i> script (de Jong and Wempe, 2009) and manually corrected |
| 5 | ORT | interval | Orthographic word segmentation; created by WebMAUS (Kisler et al., 2017) |
| 6 | KANN | interval | Canonical SAMPA annotation on word level; created by WebMAUS (Kisler et al., 2017) |
| 7 | MAU | interval | SAMPA annotation on segment level; created by WebMAUS (Kisler et al., 2017) Additional manual tags on auditory and visual basis for voiceless plosives as strongly aspirated (<code>_sa</code>), or without burst (<code>_nb</code>); glottal stops as realizations of other voiceless plosives (e.g., <code>?_t</code>) |
| 8 | hesitations | interval | Classification of discontinuities: <code>f.part.</code> = filler particle (e.g., “ummm”, “uh”); <code>break-off</code> = word/phrase is not finished; <code>repair</code> = a speaking error is repaired; <code>repetition</code> = a section is repeated as a repair attempt; <code>icon.part.</code> = e.g. “ugh”, combinations thereof |
| 9 | Emph.acc. | interval | Types of emphatic accents: <code>RIF</code> = reinforcement (long onset consonant); <code>PosInt</code> = positive intensification (long vowel); <code>AC</code> = accent chain (several emphatic accents in a row); <code>RFR</code> = rising-falling-rising contour; <code>WR</code> = word repetition; combinations thereof |
| 10 | Breathing | interval | Types of breathing: <code>BR_in</code> = audible inbreath; <code>SBR_in</code> = strong audible inbreath; <code>BR_in_nasal</code> = inbreath with audible nasal component; <code>BR_in_cut</code> = breathing cut off in editing process; <code>BR_out</code> = audible outbreath; <code>SBR_out</code> = strong audible outbreath; <code>LAU</code> = laughter; <code>COU</code> = coughing |
| 11 | Cut/edit | point | Point where a cut was made in the edit, label: “cut” |
| 12 | Comments | point | Any comments and questions that may have come up during annotation |
| 13 | Perception.sections | interval | Intervals that were used for the perception experiments |
| 14 | Vowelcenter | point | SAMPA annotation of vowels (only monophthongs) in the vowel center |
| 15 | Phrases-PP-exp | interval | Copy of tier 3; intervals adjusted in accordance with tier 13 |

the speaker (e.g., DH for Daniel Howell) followed by a running four-digit number. The pause labels included the speaker initials followed by a code for the pause type (see Table 7.1 for further information) followed by a running number that was independent of both the pause type and the phrase number.

On tier 9, emphatic accents are annotated as intervals. Emphatic accents are strong segmental accents used to highlight parts of a phrase for emphasis. Five different types of emphatic accents were annotated (see Table 7.1 for the labels and short definitions; see Kohler, 2006; Niebuhr, 2010; Niebuhr et al., 2016b; Berger et al., 2020, for further information).

On tier 10, instances of breathing are annotated. Instances of laughter and coughing are excluded from analyses, but annotated (see Table 7.1 for further information).

Tier 13 was used to mark off sections of the speech material that were used for the experiments.

7.3 DIMA annotation for English data

The intonation contours of the speakers were also annotated using DIMA, a relatively new annotation system (*Deutsche Intonation—Modellierung und Annotation; German intonation—modelling and annotation*, see, e.g., Kügler et al., 2015; Kügler and Baumann, 2019a). DIMA is a consensus system created for German speech data that is meant to combine elements from all different systems that are used in German intonation research to unify the annotations but still allow translation into the other systems. While DIMA is created for German, it is meant to be developed for application for other languages as well. This system was chosen for this study because of an existing affiliation with the DIMA creator group, a general similarity between the two Western Germanic languages German and English in terms of their intonation patterns but different realizations (Grabe, 1997; Grabe et al., 2000), and it was seen as an opportunity to test the DIMA system on a language other than German, since only one other study has (to the knowledge of the author) applied DIMA annotations to English (Niebuhr et al., 2018b; also in the context of charismatic speech).

The DIMA system uses a word and syllable level as reference points for the annotation. The word tier from each acoustic measurements annotation file (tier 5, see Table 7.1) was extracted and used as reference for the DIMA annotation. The original segment tier (tier 7) was manually converted to a syllable tier.

DIMA itself consists of four annotation tiers in addition to the word and syllable tiers: a phrase tier, a tone tier, a prominence tier, and a tier for comments and uncertainties. The following explanation of the annotation workflow is based on Kügler & Baumann (2019a) and Kügler et al. (2015), unless otherwise specified. Table 7.2

provides an overview of the different annotation tiers, labels, and annotation steps. The tone and prominence tiers are relevant for the remainder of the thesis and are introduced in detail below; for more information on the other tiers, see Table 7.2 as well as Kügler et al. (2015, 2019), and Kügler and Baumann (2019a, 2019b).

The **prominences** in each phrase are labeled on the third DIMA tier. There are three degrees or levels of prominence. The label “1” is used for weak prominences that can be triggered either by the rhythm of the phrase or tonal movements. “2” is the label for a strong prominence and it mostly falls together with an accent tone, meaning a pitch peak or a valley. Finally, the “3” is the label for an extra strong prominence which is usually an emphatic realization of a full accent with strong prominence (Kügler and Baumann, 2019b).

The **tone** labels are annotated on the second DIMA tier. There are two general tone labels: H (a high tone) and L (a low tone). First, all boundary tones are labeled as either “L” (the default label) or “H”. Then, accent tones (H* or L*, the asterisk marks them as the accent tones) are placed in the prominent syllable of a word—either on the F0 maximum or minimum in the prominent syllable, or in the center of the vowel if a) the accent tone is part of a downstep contour and the “pitch peak” is actually a plateau, or b) the F0 maximum or minimum occurs outside the prominent syllable (additionally marked with a diacritic). Accent tones are connected to a prominence of at least a level 1. Finally, additional H and L tones are labeled at tonal targets (F0 maxima and minima) that occur before and after an accent tone but are not necessarily connected to a prominence.

In general, there were no major issues with using DIMA for English speech materials, even though the system was not originally designed with English intonation in mind.

Interrater agreement of the DIMA annotations was calculated using Cohen’s Kappa (Cohen, 1960). One trained annotator labelled 30 seconds of randomly chosen, but consecutive speech material from two speakers (i.e., one minute total). A script was written by the author to gather the labels of the author (SB), and check if the second annotator (KF) placed a label within a surrounding interval (and vice versa). For each speaker, the mean vowel duration across their full five minutes of speech material was calculated. The interval for the interrater agreement was the mean vowel duration before and after the label of SB on each annotation level. Cohen’s Kappa was calculated in R Studio (RStudio Team, 2023) for a) the presence or absence of a label in the analysis interval; b) for the annotation labels with diacritics (“! ^ &”); and c) for the annotation labels without diacritics.

The agreement was overall poor, with the exception of phrase labels without diacritics (see Table 7.3). This could be because of different familiarity with the speaker’s speaking styles between the two annotators, or it could be a result of different annotation styles. After all, intonation annotation also has some degree of

Table 7.2: The inventory of labels and the general usage of each label. The levels are ordered according to process work-flow of the annotation. The table is adapted from Kügler and Baumann (2019a). A + marks a deviation from the official annotation guidelines.

| Step | Tier | Label | General usage |
|--------------------------|----------------------------|------------------|--|
| 1 | Phrase (tier 1) | % | Beginning and end of a major prosodic phrase |
| | | - | Beginning and end of a minor prosodic phrase |
| | Diacritics | l ^ | Phrasal downstep and upstep, respectively; in reference to preceding phrase (can refer to either narrower/wider pitch range or higher/lower overall pitch level) |
| | | & | Disfluencies integrated into a phrase or creating a break in the phrase |
| 2 | Prominence (tier 3) | 1 | Weak prominence, does not have to occur together with a (clear) F0 movement |
| | | 2 | Strong prominence, mostly occurs together with an accent tone |
| | | 3 | Extra strong prominence (emphatic accents), triggered by extreme F0 movements, often with segmental hyperarticulation |
| | | (3) ⁺ | Extra strong prominence without F0 movement, but with segmental hyperarticulation |
| 3 | Tone (tier 2) | HL | Boundary tones |
| | | H*L* | Accent tones (phonological association with prominence-bearing syllable) |
| | | HL | Non-accent tones (maxima and minima between accent tones) |
| | Diacritics | l ^ | Tonal downstep and upstep, respectively; in reference to preceding tone of identical quality |
| | | < > | Accent tone (maximum/minimum) associated with prominent syllable occurs in following/preceding syllable |
| Comments (tier 4) | Text | | (a) Comments: phenomena that cannot be labeled otherwise using the inventory; |
| | | | (b) Different types of disfluencies (e.g., hesitational particles, break-offs, etc.) |
| | | ? + (alt. label) | (c) Types of emphatic accents (RIF, PosInt, RFR, WR, AC; see Table 7.1) ⁺ |
| | | | Uncertainties with an alternative annotation on the comments tier; the “?” is labeled on either of the other tiers with the uncertain label as well |

Table 7.3: The Cohen’s Kappa (κ) values of the interrater agreement analysis of the DIMA annotations overall and for the separate DIMA tiers phrase, tone, and prominence. Cohen’s Kappa was calculated for the presence/absence of labels, labels with diacritics (upstep, downstep, disfluency) and without. The calculations were made for the full data set (= All), but also for each of the two speakers (AD and SP) included in the interrater sample. κ is rounded to two decimal spaces. The ⁺ marks moderate, the * substantial agreement (see Landis and Koch, 1977).

| Level | Presence/absence | | | With diacritics | | | Without diacritics | | |
|------------|------------------|-------|-------|-----------------|-------|-------|--------------------|-------|-------------------|
| | All | AD | SP | All | AD | SP | All | AD | SP |
| Overall | -0.13 | -0.14 | -0.12 | -0.02 | 0.01 | 0.00 | 0.19 | 0.22 | 0.16 |
| Phrase | -0.05 | -0.04 | -0.05 | -0.11 | -0.04 | -0.09 | 0.49 ⁺ | 0.61* | 0.47 ⁺ |
| Tone | -0.11 | -0.14 | -0.07 | 0.02 | 0.00 | -0.02 | 0.03 | -0.03 | 0.03 |
| Prominence | -0.21 | -0.18 | -0.23 | 0.03 | 0.04 | -0.15 | 0.03 | 0.04 | -0.15 |

subjectivity to it, and is therefore prone to annotator differences. SB often tended to annotate up to double the amount of labels than KF, mostly on the tone level, but also on the prominence level. While the agreement is poor, analyses are still carried out with the annotations of SB, even though some of the results will be discussed also with the interrater agreement in mind.

7.4 Measurements and scripts

The majority of the acoustic and prosodic measurements that particularly come into play in the analysis in Chapter 10 were obtained by using the ProsodyPro script (Xu, 2013, version 5.7.8.1). Both standard measurements as well as Bio-informational Dimensions (BID) measurements were used for analysis. Two measurements were added into the script by the author: the standard deviation of intensity (in dB) as well as the standard deviation of pitch in both Hz and semitones (st). The script was run separately for each of the male and the female speakers in the sample¹. The F0 range in ProsodyPro was set to 75 to 600 Hz (default) for female speakers, and to 60 to 600 Hz for male speakers. The maximum formant was left at the default 5000 Hz for male speakers, but increased to 5500 Hz for female speakers, as per the suggestion in the script.

In order to minimize the chance of ProsodyPro measuring octave errors rather than the actual values of the speaker’s pitch contour, the pitch of each speaker was manually checked and sections with obvious octave errors were unvoiced and re-synthesized, and subsequently used for the analysis. This procedure is done in Praat by turning a sound file into a Pitch object (pitch range settings: 60 to 800 Hz for all speakers). Pitch candidates of Praat’s pitch tracking algorithm that were far

¹While the project refers to gender as a social construct, the phonetic analyses require the definition of pitch ranges, which were chosen on an auditory basis that would correlate more with biological sex.

Table 7.4: An overview of the measurements taken for the analyses and the scripts used for measuring. SB = scripts written by the author.

| Area | Feature | Script |
|-----------------------------|---------------------------|------------------------|
| Pitch | mean F0 | ProsodyPro (Xu, 2013) |
| | median pitch | ProsodyPro |
| | minimum F0 | ProsodyPro |
| | maximum F0 | ProsodyPro |
| | excursion size | ProsodyPro |
| | pitch variability | addition to ProsodyPro |
| Intonation | pitch peak timing | SB |
| | final contour direction | SB |
| | emphatic accent frequency | SB |
| | prominence ratio | SB |
| Tempo & duration | speech rate | SB |
| | speech rate variability | SB |
| | pause duration | ProsodyPro |
| | phrase duration | ProsodyPro |
| | breathing duration | SB |

outside the range of the rest of the data were selected and then unvoiced². Once all sections of interested are checked and, in many cases, unvoiced, the Pitch object is converted into a Pitch tier. The original sound is turned into a Manipulation object³, and the Pitch tier in the Manipulation object is replaced by the new one. Finally, a re-synthesis is created from the Manipulation object using the “Get resynthesis (Overlap-add)” function.

The measurements used for analysis in the subsequent experiment chapters (Chapter 8 to Chapter 10) are summarized in Table 7.4. Further information on the measurements can be found in the respective experiment chapters.

²“Selection” → “Unvoice”

³“Manipulate” → “To Manipulation”; settings: 60 Hz as the pitch floor and 800 Hz as the pitch ceiling.

Part III
Studies

Chapter 8

Perception I: Prosodic manipulations

8.1 Introduction

This chapter, together with Chapter 9, works towards answering the first research question that was introduced earlier (see Chapter 5), and is repeated below as RQ1. This research question focuses on the influence of acoustic feature characteristics on charisma ratings. Additionally, this chapter addresses the third research question of the project (see Chapter 5), which is repeated as RQ3 below and focuses on the relationship between charisma and familiarity.

RQ1: *How should acoustic parameters be configured to be perceived as charismatic (both in terms of charisma directly and charisma-adjacent attributes) in the context of YouTube vlogs?*

RQ3: *Is there a connection between charisma ratings and familiarity with the speakers in the sample?*

In this part of the investigation, a perception experiment (with five rating attributes: *charismatic, authentic, enthusiastic, likable, and persuasive*) is conducted that uses stimuli of the ten speakers which are manipulated in terms of pitch level, pitch range, speech rate, and final contour direction. The four features investigated here have been shown to be connected to charisma perception. The results from previous research have mainly focused on male speakers and speakers from North America, so this project also investigates possible differences between the speaker groups, as this study also includes female speakers and speakers from England.

8.2 Hypotheses

A general hypothesis for RQ1 and this part of the investigation was put forward in Chapter 5, and it is recounted in part below as H1.

H1: *Stimuli with larger pitch ranges, higher pitch level, medium speech rates, and non-rising phrase-final pitch contours [...] are perceived as more charismatic.*

This general hypothesis can be separated into several more specific hypotheses that are directly connected to the experiment presented in this chapter. These specific hypotheses are marked with a subscript P_1 , where the P1 signifies that they belong to the first perception experiment.

Specifically, as has been suggested by previous research (see Chapter 3) and as might be expected for the realm of YouTube (Chapter 4), it is expected that features connected to more vocal effort (e.g., listener-oriented speech through more articulatory precision; see Section 4.3.4 for an overview) are, for example, perceived as more charismatic, but on YouTube less authentic and likable (see H_{P1_1}). That also means that the ratings of the acoustic manipulations are expected to differ between the attributes. Table 8.1 shows the predictions for each attribute regarding the four acoustic features and which of their specific characteristics are expected to be perceived more positively. Characteristics are likely similar between *charismatic*, *persuasive*, and *enthusiastic* ratings, as well as between *authentic* and *likable*. In particular, a previous study on vowel spaces and the same speaker sample (Berger et al., 2023) suggests that charisma and authenticity may be connected to different acoustic characteristics on YouTube.

H_{P1_1} : *There are differences in acoustic feature characteristics between the different attributes, mainly in that features connected to more vocal effort (in particular a larger pitch range, but also higher pitch level) are likely more strongly connected to higher charisma, enthusiasm, and persuasiveness ratings, but are perceived less positively for authenticity and likability in the context of YouTube.*

Most of the previous phonetic studies of charismatic speech and charisma perception have made reference to male speakers. Fairly little is known about female speakers in this context. What is expected from previous research is that the male speakers are likely perceived as generally more charismatic and persuasive than female speakers (H_{P1_2}). This was also found by, for example, Jokisch et al. (2018) and Niebuhr et al. (2019). For the other three attributes, no direction of a possible rating difference between male and female speakers is predicted (H_{P1_3}), mainly because this is not addressed in detail in previous research.

H_{P1_2} : *Male speakers are expected to be perceived as more charismatic and persuasive than female speakers.*

H_{P1_3} : *No direction of ratings the authentic, enthusiastic, and likable attributes depending on speaker gender is predicted.*

It has been shown in previous research that male speakers with higher pitch level are perceived as more charismatic (e.g., D’Errico et al., 2013; Mixdorff et al., 2018; Niebuhr et al., 2018a). This is therefore also expected in the current investigation

Table 8.1: An overview of the predictions for which feature characteristics are likely to elicit more positive ratings for each of the attributes.

| | Pitch level | Pitch range | Final contour | Speech rate |
|---------------------|-------------|-------------|---------------|----------------|
| Authentic | no effect | original | rising | no effect |
| Enthusiastic | higher | larger | rising | medium or high |
| Likable | original | original | rising | medium or high |
| Persuasive | higher | larger | non-rising | higher |
| Charismatic | higher | larger | non-rising | medium or high |

(H_{P14}). For pitch level, female speakers are not analyzed in the current sample due to stimulus exclusions. The stimuli were created through digital manipulation from the original as an intermediate stimulus to stimuli with increased as well as decreased pitch level which were piloted for naturalness (see Section 8.3.1). For the female speakers, three of the five stimuli with decreased pitch level and three of the five stimuli with increased pitch level were deemed too unnatural-sounding and were therefore excluded, which meant that entire speaker groups were no longer represented, rendering statistical analyses problematic. For the other three features (pitch range, speech rate, and final contour direction), no difference between male and female speakers has been reported, and is therefore also not predicted here (H_{P15}). While no differences have been reported in the context of charismatic speech, it is important to check since this study offers the opportunity to investigate the impact of controlled acoustic differences on the perception of both male and female speakers.

H_{P14}: *Male speakers are expected to receive higher ratings of charisma and charisma-adjacent attributes with increased F0.*

H_{P15}: *No direction for possible speaker gender differences is predicted for pitch range, speech rate, and final contour direction.*

No studies have so far been published that investigate charismatic speech of speakers from England. So far, studies into charismatic speech have focused on North American English speakers, or speakers of non-English languages. A comparison of the two speaker groups will therefore offer first new insights into differences between varieties of English and their perception. Since there is no evidence of different charisma perception from speakers of different English varieties from previous research, no direction differences of ratings and the characteristics of the acoustic features are predicted between speakers from England and speakers from North America (H_{P16} and H_{P17}). It is included in the study in an exploratory manner because the data set has both speakers from North America and England, and is therefore the first study to be able to compare results from speakers from both general origins. It is likely to find differences between the two cultures (also be-

cause all raters originate from the British Isles, see Section 8.3.5), but the direction is unclear.

HP₁₆: *Rating differences for the attributes between speakers from North America and England are likely, but the direction of these differences cannot be predicted.*

HP₁₇: *Rating differences between speakers from North America and England based on the characteristics of the acoustic features are likely, but the direction of these differences cannot be predicted.*

Finally, this chapter also addresses the third main hypothesis of this dissertation, and therefore also RQ3. It is predicted that familiar or recognized speakers are perceived as more charismatic than unrecognized speakers (H3; see also, for example, Rosenberg and Hirschberg, 2009; Lavan et al., 2016; Jokisch et al., 2018). This hypothesis refers to all acoustic features and all specific manipulations. This hypothesis is repeated here from Chapter 5.

H3: *The more familiar a speaker is to the listeners, the more charismatic they are perceived.*

8.3 Methods

8.3.1 Stimulus selection and creation

The stimuli in this part of the investigation were short stimuli (1.9 to 2.8 seconds) comprising one phrase bordered either by pauses or a pause and a minor phrase boundary. These stimuli were manipulated in terms of four acoustic-prosodic features: overall pitch level, pitch range, speech rate, and the final contour direction. These features were increased or decreased using the TD-PSOLA resynthesis algorithm (*Time-Domain Pitch-Synchronous Overlap-and-Add*, see Charpentier and Stella, 1986; Moulines and Charpentier, 1990) that is integrated into Praat. The overlap-and-add resynthesis technique adjusts the signal in overlapping segments by stretching or shrinking the signal (see Henderson and Skarnitzl, 2022 for a schematic overview). Using manipulations has the advantage that the content and the rest of the acoustic features stays constant, and differences in ratings of different attributes by listeners can be directly connected to specific acoustic characteristics.

There were some criteria that were followed in the selection process of the stimuli that would then be digitally altered. Content-wise, none of the stimuli contained dramatic superlatives or negative words or interpretations. This approach was chosen to stay away from content that is too emotionally charged since words like “depression” (Daniel Howell, 2017, e.g. min. 0:13) or phrases like “You’re telling

everyone else that you are inferior” (Louise Pentland, 2017, min. 3:14) can sway listeners into having sympathy or dislike (depending on the person) for the speaker based purely on content which in turn will automatically affect and bias the ratings of the voices as well. It was not possible to have completely neutral content. Example 1 below lists the different stimuli that were chosen for the experiment.

- (1) a. *I mean, I make so many videos, I upload so much content.* (AD; Alfie Deyes Vlogs, 2018, min. 13:03)
- b. *I got together with my brother 'cause my brother's a great writer.* (CB; Colleen Ballinger, 2017, min. 1:53)
- c. *And just being able to tell someone.* (DH; Daniel Howell, 2017, min. 6:17)
- d. *And I was one of them and that was cool to me.* (LP; Louise Pentland, 2017, min. 4:08)
- e. *How much time we're gonna spend on YouTube.* (LS; Lilly Singh Vlogs, 2017, min. 1:21)
- f. *And that is an incredible feeling.* (MF; Markiplier, 2018, min. 0:41)
- g. *It's kinda like a ticking clock now of two years.* (MP; GTLive, 2019, min. 24:45)
- h. *Unless I did die and I just haven't figured out I'm a ghost yet.* (PL; Amazing-Phil, 2018, min. 0:22)
- i. *Unless you're a big media player.* (SP; GTLive, 2019, min. 22:05)
- j. *Make the most out of every avenue that I have.* (ZS; Zoe Sugg, 2018, min. 3:36)

All stimuli that were chosen had either a falling boundary tone or a plateau. Stimuli did not have cuts in order to retain natural phrases and therefore a semi-spontaneous speaking style. The speech rate of the stimuli (specifically, the articulation rate, as speech rate here is defined as the number of syllables per seconds without including pauses) was not controlled for, but the natural speech rate was taken as the original speech rate. The different speakers were divided into three groups depending on their speech rate (see below). Controlling the speech rate was not possible on top of all other criteria for stimulus selection, so working with the natural speech rate and adjusting the subsequent manipulations was more efficient, especially for the synthesis of natural-sounding manipulated stimuli. Acoustic measurements of the stimuli are included in Section 8.3.2.

The original stimuli (ORIG) were also resynthesized without additional changes. The re-synthesis process often leaves audible artefacts in the new audio file depending on the sound quality of the audio file (Henderson and Skarnitzl, 2022). By also running the unchanged stimuli through the algorithm and using the new file for the experiments, there is no difference in artefacts to the manipulated stimuli that could influence the listeners. A small degree of artefacts could not be avoided.

Four acoustic-prosodic parameters were manipulated: pitch level, pitch range, speech rate, and final contour direction. Each parameter was present in three ver-

sions—a high version, an intermediate version (usually the original), and a low version. The different manipulations were not crossed in this experiment, so there was always only one change in the acoustic-prosodic make-up of the stimuli. There were nine stimuli per speaker (90 individual stimuli in total). The process behind each of the manipulations is detailed below.

A manipulation object was created in Praat (Time step: 0.01s; Minimum pitch: 60 Hz; Maximum pitch: 600 Hz) for each change. Pitch points indicating octave errors were deleted to not influence the manipulations by affecting the mean F0 measures or the deviations from the mean unnaturally.

In order to manipulate the **pitch level** with Praat, the entire utterance is selected in the manipulation object. Using the “Shift pitch frequencies” function, the pitch frequencies are increased by 3 st for the stimuli with a pitch level that is higher than the original (HF0) or decreased by 3 st (LF0) for stimuli with a lower pitch level than the original.

For the manipulation of the **pitch range**, the PitchTier is first extracted from the manipulation object, selected and the “Formula” function (via “Modify”) is chosen. The formula (Berger, 2017, Berger et al., 2017)¹ used is given in Example 2 below. In this formula, the overall mean F0 for the phrase (in Hz) is subtracted from each pitch point in the signal (= *self*). The result is then multiplied by a factor and finally added onto the mean F0 value again. The factor was 1.5 for an increase of pitch range compared to the original stimulus (HF0R) and 0.5 for a decrease of pitch range (LF0R). The method of only multiplying the difference between each pitch point and the mean has the advantage that the change in pitch has a greater impact on pitch values further away from the mean and therefore affects the extremes more severely than values close to the mean. This method resulted in stimuli with natural-sounding pitch ranges. Finally, the original PitchTier in the manipulation object was replaced by the new PitchTier, and the new sound was re-synthesized.

$$(2) \quad ((self - mean) \cdot factor) + mean$$

For the **final contour direction**, the pitch points were manually shifted in the manipulation object. All stimuli were naturally non-rising, but some had a final falling contour and some had a natural plateau. The original stimulus was therefore either tagged as falling or plateau, and the other category was the one that was created by manipulation. All stimuli were also changed to have a final rising contour. The changes happened after the final pitch accent of the utterance so as to not change the meaning of the pitch accent by accidentally switching it from a H* to a L* pitch accent. The manipulations are based on the direction of the contour (rising, falling, or plateau), not primarily on the ending height of the boundary tone itself. The

¹The formula was created in 2016 for Berger (2017) by Evelin Graupe.

falling contours ended up with values between -2.2 and -16.7 st; the rising contours between 1.6 and 14.8 st; the plateau contours had changes between the minimum values for the rises and falls. There were some measuring issues for some stimuli that did not coincide with the auditory impression of the stimuli (see Table B.1 in Appendix B). However, it was deemed most important that the resulting stimuli sounded natural, but as distinctive as possible. There were no standardized values for the size of the rise and the fall, but within each speaker, the categories were auditorily distinguishable.

The **speech rate** is manipulated by adding a duration point at the beginning of the sound file in the manipulation object. The duration point has an effect on the relative duration of the file. The relative duration is changed by inputting a factor that either slows down the speed of the phrase or increases the speed. The factor is calculated with the equation in Example 3 below, where the original speech rate (number syllables divided by phrase duration) is the speech rate before changes, and the target speech rate is the speech rate the stimulus is supposed to have.

$$(3) \quad \frac{\text{original speech rate}}{\text{target speech rate}}$$

There were three categories for the speech rate manipulations: low speech rate (below 5.6 syll/s), medium speech rate (5.6 to 6.6 syll/s), and high speech rate (above 6.6 syll/s). The speakers' stimuli were classed into one of those categories. That way, the manipulated speech rates are comparable because they are in the same three groups within the same extreme values, but the natural speech rate is retained and comparable.

For stimuli that were classified as having a low speech rate naturally, the speech rate was increased by one syllable per second for the medium level, and by two syllables per second for the high speech rate level. Stimuli that fell into the medium speech rate category were slowed down and increased by one syllable per second, respectively. Stimuli that already had a high speech rate were slowed down by one syllable per second to reach the medium level, and by two syllables per second for the low speech rate level.

Following a naturalness pilot with seven participants (2 male, speakers of German as a first language/L1, trained linguists and phoneticians), ten stimuli were excluded, leaving 80 stimuli for the final experiment. The participants of the pilot were presented with the audios and a goodness scale: *How realistic/natural does the voice sound on a scale from 1 (= very unrealistic/unnatural) to 4 (= very realistic/natural)?* The purpose of the pilot was not to judge the naturalness in terms of the sound quality, but rather the realism of the intonation contours. Would it be possible for speakers to speak like the recording in the real world? Participants were made aware of this purpose. Ten stimuli were deemed as too unnatural-sounding, since

Table 8.2: An overview of the different prosody-related manipulations of stimuli used in the perception experiments in the present study. Abbreviations of the manipulations, their explanations as well as the underlying change are provided. Note that one of the three speech rate manipulations is always included in the original per speaker.

| Abbreviation | Explanation | Change/Factor/Range |
|--------------|---|---------------------|
| ORIG | unchanged stimulus | — |
| HF0 | increased pitch level | +3 st |
| LF0 | decreased pitch level | – 3 st |
| HF0R | widened pitch range | *1.5 |
| LF0R | narrowed pitch range | *0.5 |
| HSR | high speech rate | > 6.6 syll/s |
| MSR | medium speech rate | 5.6 – 6.6 syll/s |
| LSR | low speech rate | < 5.6 syll/s |
| rising | rising final contour | rise |
| non-rising | falling or plateau, depending on original | fall or plateau |

they only achieved a mean rating of 2.0 or below, and were therefore excluded from the analysis.

Table 8.2 provides an overview of the prosody-related manipulations, including abbreviations, explanation, and the change, factor or range that is part of each manipulation. This chapter only investigates these four features. The rest of the acoustic make-up of the stimuli—while important for the full picture of the perception of charisma and its related attributes—will not be included here, but come into play in Chapter 10. Differences in rating between the manipulations are seen as indicators for differences in charisma perception since the rest of the stimuli and their acoustic make-up is kept constant.

8.3.2 Stimulus measurements

Measurements of the features were taken for each of the stimuli in Praat (Boersma and Weenink, 2018, version 6.0.37). Mean pitch and pitch range (i.e., excursion size) were measured with ProsodyPro (Xu, 2013, version 5.7.8.1). The mean pitch of the LF0 stimulus of PL was manually re-measured after removing octave errors from the signal. The ProsodyPro pitch range measurements for the ORIG and HF0R stimuli of speaker DH were inconsistent. A visual inspection suggests that the measurement for the ORIG stimulus is correct and the one for the HF0R stimulus is too low (which is also suggested by the auditory inspection of the stimulus). This was re-measured manually following the method used by the ProsodyPro script. As a control, the ORIG stimulus was also re-measured manually, but this value was equal to the one measured by ProsodyPro. The final contour was measured manually as the F0 of the final measurable point minus the F0 of the tone point

before (usually, an accent tone). This method resulted in the difference in F0 or excursion between the two points in semitones. The main aspect of categorization was the auditory impression of the stimuli, though. The speech rate was measured by a Praat script written by the author. The absolute measurements are included in Table B.1 in Appendix B, together with the other measurements. Note that the falling and plateau contours of speaker ZS could not be measured as Praat provided no pitch track.

8.3.3 Experiment procedure

Due to travel restrictions during the Covid-19 pandemic, the experiment sessions were carried out with a remote, but still face-to-face method using a video conferencing system. The experiments could not be run as a web-based study since the audio data could not be uploaded to another website like LimeSurvey (Limesurvey GmbH, 2017) for copyright reasons.

Therefore, Experiment Multiple Forced Choice (ExperimentMFC, version 7, Boersma, 2013), an in-built experiment-programming tool in Praat (Boersma and Weenink, 2018) was used for the perception experiments as an offline option because it allowed the free coding of Likert scales and the playing of audio files, whilst also enabling two ratings in one trial. For the basic experiment file and instructions on how to fill it, see Boersma (2016) as well as Beck (2014) and Mayer (2017). The details of the experiment designs are included below, an example of the experiment files is included in Appendix C. In ExperimentMFC, participants hear an audio stimulus and can then rate a specific statement by clicking on the button corresponding to their respective response.

Zoom (Zoom Video Communications, 2021, versions 5.4.2 to 5.12.2, December 2020 to October 2022) was chosen as the video conferencing system. It offered all necessary functions for the experiments (screen and audio sharing; remote control for a specific desktop window) and was so widely accessible that it was likely that most prospective participants would use Zoom in their daily lives or at least were familiar with it.

The issue with Zoom was data security, as data collected via Zoom are stored on servers in the US where European data protection policies do not apply. After conferring with the data security expert at Kiel University, the use of Zoom was allowed—especially because none of the other systems that were considered more safe offered the functions needed for the experiments (D. Geißler, p.c., July 2020). However, in order to keep all personal information off of Zoom, metadata about the participants (gender identity, age, origin, previous experiment experience, etc.) were collected before the experiment sessions on LimeSurvey (Limesurvey GmbH, 2017, versions 3.25.6 to 3.28.35, December 2020 to October 2022), a survey website

for which Kiel University has a license, where the results of the survey are also stored according to European data protection policies. There is no remaining photographic record of the calls. All results and metadata were made anonymous.

In the experiment session, the author was on the Zoom call with the audio and camera turned on. During the experiment itself, the author's microphone was muted but could be unmuted at any time in order to answer questions from the participants. The author wore headphones for the entire time to hear if a question was asked. The camera remained turned on during the experiment, but the author moved off camera and sat at a second table to the side to ensure the participants that their responses were not being watched or monitored in real time in order to avoid possible effects of monitoring (see, e.g., Donadeli and Strapasson, 2015). The participants were informed that they needed to have their microphone turned on in the beginning of the session to talk through the instructions and in the end for the debriefing after the experiment was finished. They were allowed to mute themselves during the experiment, but were informed that the author was able to hear them if they decided not to mute themselves. They were able to ask questions between the experiment parts and when they had finished the experiment. They were also asked to use headphones during the experiment. The type of headphones (in-ear or on-ear) was not restricted, but it was noted down in the beginning of a session should this information become relevant in the future. Cable-connection or Bluetooth-connection of the headphones was also not restricted. Each participant could decide if they wanted to have their camera turned on or off.

Once the video call was started on the day of the experiment session, the instructions were explained to the participants in detail. They were able to ask questions. The full instructions can be found in Appendix D. They were sent to the participants in advance but explained face-to-face extensively. This took about 15 minutes between the beginning of the session and the start of the actual experiment. After all instructions were clear, the screen sharing was started and the participants were given remote control over the Praat experiment window. Then, the experiment started with a soundcheck where participants could adjust their audio to a comfortable level, before the actual experiment began.

The Zoom call with the participants was run from the student Zoom account via the Kiel University log-in. The camera was the in-built webcam of a laptop. The background in the camera frame was either blurred by Zoom or solid white to reduce the risk of distractions for the participants.

8.3.4 Experiment design

The experiment was split in two parts. In Part 1 of the experiment, the participants rated statements regarding the authenticity, likability, enthusiasm, and persuasive-

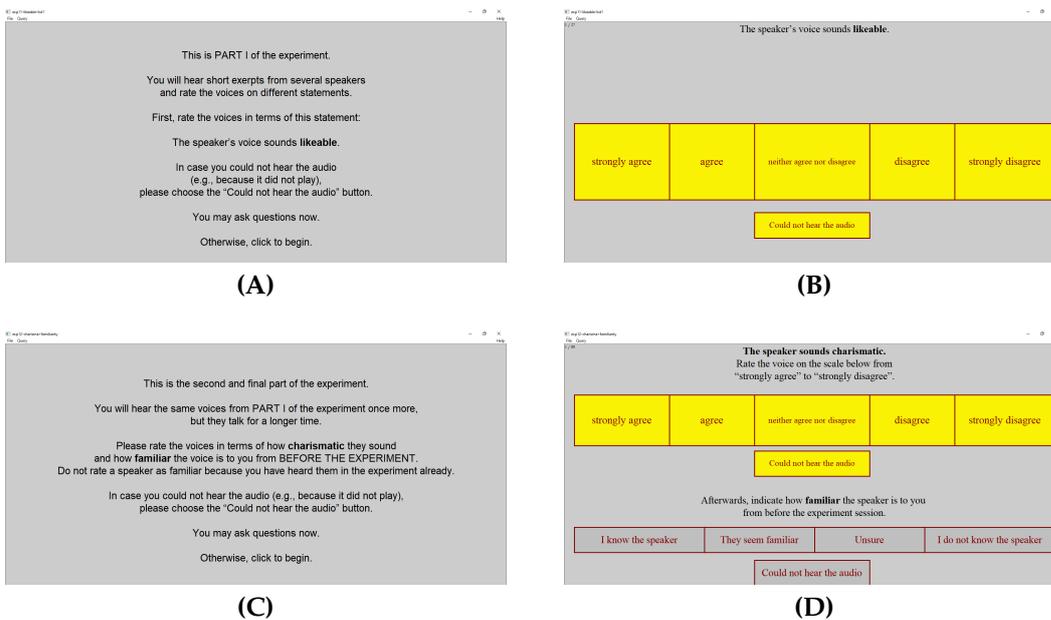


Figure 8.1: (A) Instruction and (B) rating screen in experiment Part 1; (C) instruction and (D) rating screen in experiment Part 2.

ness of the speakers. Afterwards, they immediately started Part 2 of the experiment where the perceived charisma and familiarity of the speakers were rated. Between the experiment parts, instruction screens were shown (Figure 8.1A and C) and participants could ask questions.

All experiment slides included possible responses for the current statement and audio stimulus the participants had to rate. All stimuli for one statement were rated, then the experiment moved to the next statement. This blockwise stimulus presentation was chosen to reduce the chance of misjudgments because of a missed change of attribute. That was seen as a desirable advantage, despite the risk of increasing fatigue because of the repetition. A grey screen alerted the participants to the switch.

The experiment was designed with 5-point Likert scales ranging from “strongly agree” (5), “agree” (4), “neither agree nor disagree” (3), “disagree” (2) to “strongly disagree” (1)—always in that order from left to right which could introduce a bias, but also reduces the risk of random clicks because the labels were not read and processed correctly by the raters. The numbers in parentheses indicate the point value given to the ratings which are—aside from the categorical values—also used in the analyses. The 5-point Likert scales were chosen to include a neutral option for participants’ ratings. While that may pull average response results towards a neutral middle more easily, having the neutral answer also reduces the stress of a forced choice on participants. This was deemed beneficial considering a long and potentially fatiguing experiment. The stimuli were rated by the listeners on four

different Likert scales in Part 1 of the experiment, see Example 4 below, and on charisma directly in Part 2 of the experiment, see Example 5.

- (4) a. *The speaker's voice sounds authentic.*
b. *The speaker's voice sounds enthusiastic.*
c. *The speaker's voice sounds likable.*
d. *The speaker's voice sounds persuasive.*
- (5) a. *The speaker's voice sounds charismatic.*

The attributes *enthusiastic* and *persuasive* were chosen because they were also included in previous studies of charismatic voices (see Rosenberg and Hirschberg, 2005, 2009; Signorello et al., 2012a, 2012b). The attribute *likable* was similar to the attributes *friendly* and *charming* from Rosenberg and Hirschberg's studies (2005, 2009), and its inclusion in some charisma definitions (Grabo et al., 2017). The term *charming* was also used in a previous pilot study with English L1 speakers, but turned out to be too abstract for many of the participants. *Likable* was deemed a more immediately graspable personal attribute. The term *authentic* was also used in a pilot study, and was carried over to the experiment because a major selling point of many YouTubers is that they claim to be authentically themselves, especially in their vlogs (see, e.g., Bishop, 2018, see also Section 4.2.2). In the pilot, a participant mentioned that they could not rate a YouTuber as authentic by default because "they always want to sell something" (paraphrased). This apparent contradiction made authenticity an interesting attribute to include. At the same time, authenticity is closely related to genuineness and honesty, which are also seen as connected to charisma (Potts, 2009; Bastardo, 2020). All four of the attributes are considered to play into charisma, but are also relevant for success on YouTube (see Chapter 4). They therefore serve as both personality ratings and indirect charisma ratings.

On top of the direct charisma rating in Part 2, participants were asked to indicate how familiar they were with each speaker from the social media context prior to the experiment session. The response options for the familiarity check were "I know the speaker" (4), "They seem familiar" (3), "Unsure" (2), and "I do not know the speaker" (1). The numbers in parentheses indicate the numerical value given to the ratings, and both the categorical and numerical values are used in the analyses. A familiarity check was included as previous research found that recognized speakers were rated as more charismatic by listeners than unrecognized speakers (e.g. Rosenberg and Hirschberg, 2009; Jokisch et al., 2018).

One attribute was presented after the other to minimize the potential for confusion over which attribute participants were rating. A grey screen informed them when the attribute was changing. Within each attribute block and across all of Part 2, the stimuli were randomized by the Praat experiment algorithm using the <Per-

Table 8.3: Experimental lists for Part 1 of the experiment for the prosodic manipulations. Each line corresponds to one stimulus. Each stimulus is labeled by the one manipulation that created the stimulus. (M = manipulated)

| Stimulus | List 1 | List 2 | List 3 | List 4 |
|-------------------------------|--------------|--------------|--------------|--------------|
| ORIG LF0 Rising | likable | authentic | enthusiastic | persuasive |
| ORIG HF0 Non-rising (M) | persuasive | likable | authentic | enthusiastic |
| ORIG LF0R SR large (M) | enthusiastic | persuasive | likable | authentic |
| ORIG HF0R SR small (M) | authentic | enthusiastic | persuasive | likable |

muteBalancedNoDoublets> function. The instructions for both experimental parts as well as examples of the rating screens are displayed in Figure 8.1. An additional response button labeled “Could not hear the audio” was added at the bottom of the rating screen which participants could use in case the internet connection was not as stable as needed to transmit a stimulus completely. That way these responses could be disregarded, but the rest of the responses remained valid.

The experiment was a between-subjects design in order to reduce the duration of the experiment and to avoid fatigue while gathering as much information as possible. There were two groups to begin with—Group 1 rated short stimuli in Part 1 of the experiment and long stimuli in Part 2, while Group 2 rated the opposite. Each group was split into four lists for Part 1 so that each participant rated each stimulus, but only on one Likert scale. The stimuli were randomized within each attribute, and all participants rated all stimuli on charisma and familiarity in Part 2 of the experiment. Five participants were recruited per list in each group, leading to the 40 participants that were recruited (see Section 8.3.5 for details on participant recruitment and the participant demographics). The experimental lists with the Likert scales and stimuli are included in Table 8.3 for the prosodic manipulations that are investigated in this chapter.

8.3.5 Participant recruitment and demographics

Participants between the ages of 18 and 44 were chosen for the experiment. This age range was chosen to include the major part of the YouTube demographic, although of course people of all ages are active in all YouTube communities. SocialShepherd, a website for YouTube statistics, states the following: “In the UK, around 24% of YouTube users were between 16-24 years old. 44% of users were between 25-44

years old” (Shepherd, 2023, last accessed January 31, 2023). This covers almost 70 percent of YouTube users in the UK, but it also shows that there is a large number of users younger and older than this age bracket.

Prospective participants had to speak English as L1 which was explained in the call for participants as having spoken English at home as a child, although having grown up in a bilingual household was also acceptable. While initially only participants were asked to sign up if they grew up in the country of England in order to have a defined area of speech, the call was adjusted to include English L1 speakers of any origin currently living in Europe. This adjustment was made after a first round of participant recruitment proved to be very slow. The adjustment of the call widened the prospective participant pool in the hopes of speeding up the recruitment process. A final requirement was that the participants had no known hearing impairments as the study relied on auditory perception.

The call for participants was distributed through different channels. A collaboration with the University of York was started in order to recruit many participants living in one area who were very likely all in the targeted age range. The study was added to a newly implemented participant recruitment system (SONA system) at the Department of Language and Linguistic Sciences. Students were able to sign into the system and sign up for a specific time slot.

The call was also disseminated to researchers at universities, institutions, and acquaintances in England and Germany. Prospective participants from these recruitment methods were asked to sign up for a time slot of their choice on a poll on the scheduling tool of the DFN (*Deutsches Forschungsnetz*, engl. *German Research Network*) and leave a comment with their first name and email address (visible only to the author) so they could be contacted via email for more information.

All participants who finished an experiment session were indirectly compensated for their time and effort. After all experiment sessions were finished, the author donated £1.00 per finished experiment session to a non-profit organization working in the UK. POPYRUS² is an organization that stands for the prevention of young suicide, whose age range is the age range of the participants. Additionally, linguistics students at the University of York also received course credit for participating which is part of the collaboration with the university.

Even though the call for participants was distributed via several channels, the sign-up rate was extremely slow. In eight months, only 19 participants could be recruited. Five of them had to be excluded for various reasons (age, L1, technical difficulties). Two additional participants were recruited with this method in October 2022, bringing the final number of participants recruited with the Papyrus method to 16. Therefore, the recruitment method was subsequently adjusted.

²Website of the organization: <https://www.papyrus-uk.org/>

The rest of the participants were recruited via the paid participant recruitment platform Prolific (Prolific, 2022, April to October 2022). A vetting study was set up where participants were sent to the metadata survey that the other participants also filled out. It was only slightly adjusted in that it now included a field for the Prolific ID that is used for anonymity as well as for reaching out to the participants, and a completion link was added in the end to send the participants back to Prolific after filling out the survey. This vetting study was sent to 150 participants in two rounds, who received a compensation of £0.75 for their participation. Several digital pre-screening attributes were chosen on Prolific in order to limit the possible eligible participants. The study was only shown to Prolific account holders between the age of 18 and 35, with nationality UK, US, Ireland or Australia, but currently living in Western or Central Europe with English as their first language. They were also supposed to have normal or corrected to normal vision, and no hearing difficulties or cochlear implants. The study was only shown to account holders who indicated that they would be interested in a video call interview, and regularly use a computer.

The responses of the vetting study were manually reviewed and the participants for the interview were selected randomly out of those who fit all requirements (especially regarding age, English as L1, and no hearing difficulties). Messages were sent out to the selected participants, and they were sent to the DFN scheduling tool to select a timeslot for the experiment.

The perception experiments were then conducted the same as before. At the end of each week, the participants of that week were sent another study on Prolific. That survey checked their Prolific ID again, and asked for a brief comment on what they understood as “charisma” or “charismatic personality” in order to gather additional information. This short survey was used as a way to compensate them for the participation in the perception experiment proper. Compensation was £7.50 for the 60-minute long experiment.

Compensation differed greatly between the different recruitment methods. With the Papyrus method, they were compensated indirectly with a £1.00 donation, while some student participants additionally got course credit. With the Prolific method, participants were paid significantly more—£7.50 for the full experiment session plus £0.75 for the vetting study. This unequal compensation may have affected their behavior. However, the recruitment net was cast so wide that it is unlikely that differences in participants’ motivations are systematic. Additionally, the ratio of cancellations to finished experiments was higher with the Prolific method (see Figure 8.2), which would not suggest an increased pull of compensation.

In total, 24 participants were recruited via Prolific, on top of the 16 participants recruited elsewhere. After the prolonged and difficult recruitment process, gathering a sufficient amount of 40 participants for statistical analysis was deemed the

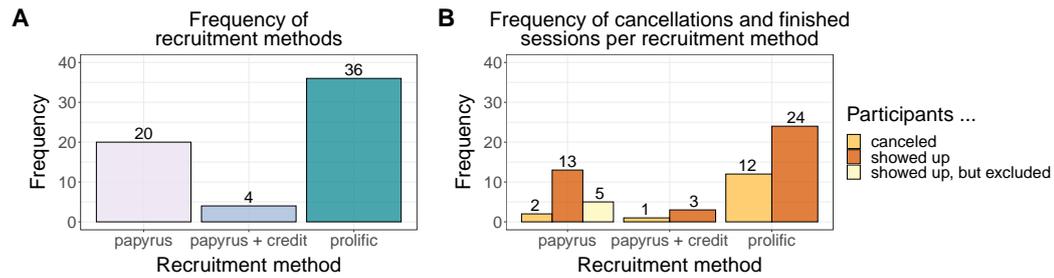


Figure 8.2: A) shows the frequency of participants who signed up with the different recruitment methods. B) shows the frequency of finished experiments (= showed up), but also cancellations and exclusions from the analyses.

priority, leaving a gender-balanced sample for future research. In the end, 22 participants identified as female, 16 as male, one as non-binary, and one preferred not to answer. Participants were between 18 and 36 years of age ($M = 28.73$ years; $SD = 5.23$ years). All participants lived in the UK, Ireland, or Germany at the time of the interview. They all spoke British English varieties having grown up in the UK or Ireland for the majority of their lives. North American speakers did not respond to the recruitment calls and could therefore not be included in the investigation. All participants reported normal hearing and normal or corrected to normal vision. Some additional information on the participants is also available since the participants were asked to fill out a personal information survey before the experiment session. Ten participants used in-ear and 26 on-ear headphones; the information is unavailable for four participants. Six of the 40 participants noted singing experience, and 17 playing an instrument. Seven individuals had previously taken part in a perception experiment, while 32 had not (the information was unavailable for one participant). Finally, 33 of 40 participants reported following internet personalities on social media, and 27 participants reported watching vlogs on YouTube. All participants accepted the data protection agreement.

In this study with the stimuli with prosodic manipulations, 20 participants rated the stimuli on the charisma-adjacent attributes, and 20 other participants rated them on charisma directly and familiarity. Table 8.4 provides an overview of the relevant demographics of the participants. Note that the participants rating the charisma-adjacent attributes in this study then rated the stimuli in the second study in terms of charisma and familiarity, and vice versa (i.e., long stimuli; see Chapter 9).

8.3.6 Statistical analyses

Some responses had to be excluded from the analyses. For the results regarding the charisma-adjacent attributes (2,200 possible responses), three responses were excluded because the participants indicated they had not heard the audio, and

Table 8.4: An overview of the participant demographics in the two participant groups. Part 1 refers to the part of the experiment with ratings of charisma-adjacent attributes, and Part 2 to the direct charisma and familiarity ratings. Participants either rated short stimuli or long stimuli for each part. (pnta = prefer not to answer; nb = non-binary)

| | Group 1 Part 1 (short); Part 2 (long) | Group 2 Part 1 (long); Part 2 (short) |
|-----------|---|---|
| Gender ID | f = 11; m = 8; pnta = 1 | f = 11; m = 8; nb = 1 |
| Age | $M = 27.5$; $SD = 5.07$ | $M = 29.95$; $SD = 5.23$ |
| Age range | 18 – 36 | 21 – 35 |

four responses were excluded because the participant had clicked their response before the stimuli stopped playing, leaving 2,193 responses in the data set. Similarly, for the direct charisma and familiarity part of the experiment (1,600 possible responses), there were two responses excluded each because of no audio or a response before the end of the stimulus, leaving 1,596 responses in this data set. These small numbers of exclusions are negligible for the experiment.

The statistical analyses in this chapter were calculated in R Studio (RStudio Team, 2023, version 2023.6.1.524; R version 4.2.2). Linear mixed models (LMMs) were used via the `lmer()` function of the `lme4` package (Bates et al., 2023) to analyze the responses for different subsets based on the manipulated features.

For the charisma-adjacent attributes, four models were calculated for each of the attributes, one per acoustic feature (pitch level, pitch range, final contour shape, and speech rate). That way, the unchanged stimulus could always be directly compared to the two manipulations, and the unchanged stimulus was not rated repeatedly, but occurred only once per attribute for each experiment participant.

The models are knowledge-based models, meaning that the variables are included that are relevant for the investigation, and the random effects structure is built according to expectations. The rating response is the dependent variable. The *Manipulation* (three levels per model), speaker gender (*Gender*, levels: male and female), and speaker origin (*Origin*, levels: ENG and NAM) are the independent variables. Additionally, interactions between *Manipulation* and *Gender* as well as *Manipulation* and *Origin* are included. A three-way interaction was not relevant for the research question and the experimental design at hand, so it was not included in the model.

It was assumed that the individual speaker (*Speaker*) and individual listener (*Participant*) were two variables to have a random effect on the ratings, as both speakers and experiment participants were pulled randomly from the population and would vary in a replication study and should therefore be included as random effects in the models with their own intercepts (see, e.g., Baayen, 2008; Winter, 2019). For *Participant*, it was also assumed that each participant would react differently to a female or male voice, or a speaker from England or North America. Therefore,

Gender and *Origin* were included as random slopes for *Participant* as well. The full model that was used throughout this analysis is shown in Example 6 below.

(6) `lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 | Speaker) + (1 + Gender | Participant) + (1 + Origin | Participant), data)`

For the direct charisma ratings, *Familiarity* was additionally included as a main effect, as well as two-way interactions with *Manipulation*, *Gender*, as well as *Origin*, and as a random slope for *Participant*. It was included as a random slope because it was assumed that participants who knew a speaker could either like the person and therefore give them a higher charisma rating, or dislike the person which would likely result in lower ratings. Since this is not available information in the current experiment design, it was included as a random influence. The R code for the LMMs of the charismatic ratings is shown in Example 7.

(7) `lmer(response ~ Manipulation + Gender + Origin + Familiarity + Manipulation:Gender + Manipulation:Origin + Manipulation:Familiarity + Familiarity:Gender + Familiarity:Origin + (1 | Speaker) + (1 + Gender | Participant) + (1 + Origin | Participant) + (1 + Familiarity | Participant), data)`

These models were chosen for all analyses dealing with the charisma-adjacent attributes on the one hand, and the direct charisma ratings on the other hand, even though it might not be the mathematically best-fitting model in all cases. This way, the analyses are comparable, also with the models from the following chapter. Some models returned a singular model fit. Barr et al. (2013) argue that overfitting (which is what the singular fit implies) tends to be far less of a problem than underfitting a model. That is why for this study, in order to have comparable models, the structure was kept constant. For all models that either did not converge or showed singular fit, the `allFit()` function from the `lme4` package (Bates et al., 2023) was used to check if using a different optimizer could help the model converge (see Clark, 2020). The full models, including changes in optimizer, are collected in Appendix E.

For pitch level and its manipulations, only the male speakers in the sample are investigated. The LF0 stimuli of three of the five female speakers had to be excluded, as was the case for the HF0 stimuli (see Section 8.3.1). That meant that the sample was getting too small to gain reasonable insights. *Gender* was therefore removed from the model structure for the pitch level analyses.

All models showed acceptable variations of normal distribution of residuals. This was tested using the `mcp.fnc()` function of the `LMERConvenienceFunctions` package (Tremblay and Ransijn, 2020). Additionally, the independent variables were not colinear, which was tested using the `vif()` function of the `car` package (Fox et al., 2023). The Estimated Marginal Means (EMMs) of the different groups

were compared using the `emmeans()` function from the `emmeans` package (Lenth, 2023) with Tukey adjustment.

In addition to the LMMs, the correlation between the direct charisma ratings and the familiarity ratings was calculated with Pearson correlations (r) using the `cor.test()` function. Even though the responses and familiarity ratings were not normally distributed, the sample was deemed large enough to use the parametric Pearson correlation nonetheless since all subsets of the data that were analyzed had between 300 and 600 data points (Levshina, 2015). Visuals were created using the `ggplot()` function provided by the `ggplot2` package (Wickham et al., 2014).

The significance level is set at $\alpha = .05$. Nonetheless, non-significant trends (p -values that can be rounded to .1, i.e., higher than .05 and lower than .15) are also reported since they can also provide insights for future studies.

8.4 Results I: Charisma-adjacent attribute ratings

In this first results section, the results regarding the charisma-adjacent attributes *authentic*, *enthusiastic*, *likable*, and *persuasive* are presented. They are organized by acoustic feature that was manipulated. Each feature section presents the rating results for each of the four attributes. The results are presented both for a comparison of male and female speakers, as well as for speakers from England and from North America. Both descriptive results from the figures and quantitative results from the LMMs are included.

8.4.1 Pitch level

Note that only the five male speakers are included in the analysis of the pitch level manipulations and their effect on the charisma-adjacent attributes. The female speakers were excluded here, since stimuli had to be excluded from the experiment after a naturalness pilot (see also Section 8.3.1). That also means that the analyses dealing with pitch level manipulations make reference to two male speakers from North America and three male speakers from England.

For the *authentic* ratings, the LMM revealed a significant main effect of *Manipulation* (see Table 8.5) suggesting that HF0 stimuli were generally lower rated in terms of authenticity than ORIG stimuli. Pairwise comparisons did not reveal significant group differences. The EMMs were lower for HF0 than for ORIG and LF0 (HF0 = 3.03; ORIG and LF0 = 3.14). The figures do not show much difference in variation and rating for the male speakers (Figure 8.3A). The median of the LF0 stimuli is higher than the medians of the ORIG and HF0 stimuli though, perhaps suggesting a slight preference for a lower pitch to perceive male speakers as authentic. For speakers from England, it seems like LF0 and ORIG stimuli have a tendency to be

Table 8.5: The output of the LMM analysis for pitch level and the *authentic* ratings. The intercept is the ORIG stimulus for male speakers from England. The independent variables are *Manipulation* (three levels: ORIG, LF0, HF0), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|-------------|-------|------|---------|-----------|---|
| (Intercept) | 3.40 | 0.24 | 14.06 | <.001 | * |
| LF0 | -0.07 | 0.31 | -0.23 | .821 | |
| HF0 | -0.81 | 0.31 | -2.64 | .009 | * |
| NAM | -0.52 | 0.43 | -1.21 | .271 | |
| LF0 × NAM | 0.15 | 0.54 | 0.28 | .783 | |
| HF0 × NAM | 1.40 | 0.54 | 2.58 | .011 | |

Signif. codes: * .05, . .1

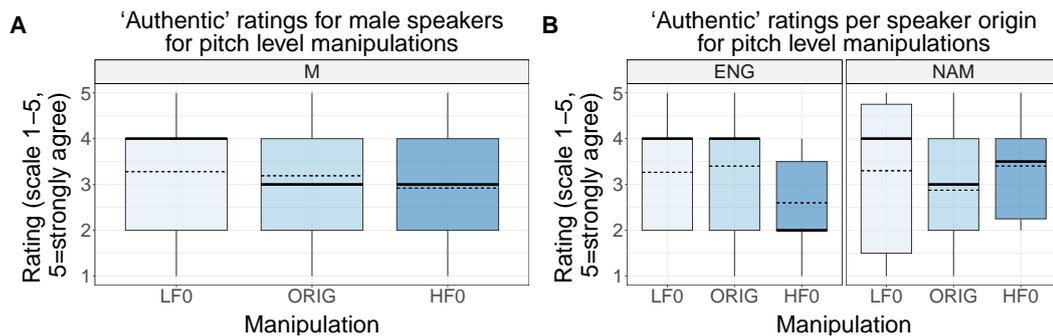


Figure 8.3: The results of the *authentic* ratings of the stimuli with pitch level manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

perceived as more authentic than HF0 stimuli (left panel in Figure 8.3B). A preference for LF0 when rating authenticity is visible from the higher median response for the North American speakers compared to ORIG and HF0 stimuli (right panel in Figure 8.3B), but this goes in hand with incredible wide variation, which is not the case for HF0.

For the *enthusiastic* ratings, the LMM showed a significant main effect of *Manipulation* as well as a significant main effect of *Origin* (see Table 8.6). LF0 stimuli were significantly lower rated than ORIG stimuli. The pairwise comparisons did not reveal significant contrasts, but the mean of LF0 was lower than those of both ORIG and HF0 stimuli, and while there was no significant difference between ORIG and HF0, the mean of HF0 was still higher (LF0 = 3.06; ORIG = 3.47; HF0 = 3.54). Speakers from North America were rated as less enthusiastic than speakers from England, which was also confirmed by a non-significant trend in the post-hoc pairwise comparisons (*estimate* = 0.52, *SE* = 0.23, *t-ratio* = 2.26, *p* = .06). The descriptive results corroborate the results from the LMM: male speakers were perceived as most enthusiastic with ORIG or HF0 stimuli (Figure 8.4A). For speakers from England, HF0 seems to be preferred, though also ORIG and LF0 stimuli have high medians, but less variation towards the highest rating option (left panel in Figure 8.4B). For speakers from North America, there was no preference when rat-

Table 8.6: The output of the LMM analysis for pitch level and the *enthusiastic* ratings. The intercept is the ORIG stimulus for male speakers from England. The independent variables are *Manipulation* (three levels: ORIG, LF0, HF0), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|-------------|-------|------|---------|-----------|---|
| (Intercept) | 3.78 | 0.13 | 30.07 | <.001 | * |
| LF0 | -0.64 | 0.29 | -2.25 | .027 | * |
| HF0 | 0.15 | 0.28 | 0.53 | .601 | |
| NAM | -0.63 | 0.22 | -2.94 | .007 | * |
| LF0 × NAM | 0.49 | 0.46 | 1.05 | .297 | |
| HF0 × NAM | -0.14 | 0.46 | -0.30 | .763 | |

Signif. codes: * .05, . .1

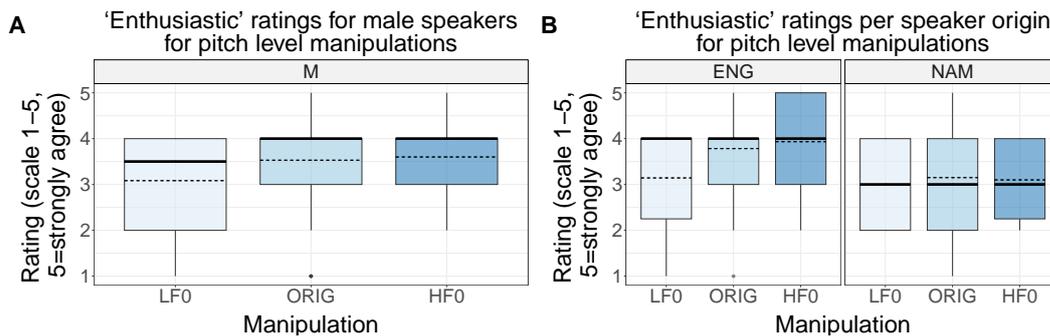


Figure 8.4: The results of the *enthusiastic* ratings of the stimuli with pitch level manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

ing enthusiasm (right panel in Figure 8.4B), though the ORIG stimuli may polarize more with variation to the extreme ratings, while the HF0 and LF0 stimuli were rated more neutrally.

The LMM for the *likable* ratings did not reveal significant interactions or main effects (all p -values $\geq .3$; see Table F.1 in Appendix F). Similarly, the visuals do not offer discernible patterns for male speakers (Figure 8.5A). For speakers from England (left panel in Figure 8.5B), LF0 might be slightly preferred, but there is no obvious pattern. For North American speakers, the LF0 stimuli seem to be rated as least likable, and the HF0 stimuli as most likable: the variation reaches the positive extreme, but not the negative extreme (right panel in Figure 8.5B).

The LMM of the *persuasive* ratings also did not reveal significant interactions or main effects (all p -values $\geq .4$; see Table F.2 in Appendix F). Visually, *persuasive* ratings for male speakers (Figure 8.6A) range across the board for ORIG and HF0, and LF0 seems to be rated most persuasive, both in terms of median and small variation. For speakers from England (left panel in Figure 8.6B), the medians suggest that LF0 and ORIG get generally neutral ratings, while HF0 tends to be perceived as less persuasive. LF0 is perceived as most persuasive for speakers from North America, with a high median and almost no variation in ratings, and ORIG seems to be similarly rated, but with more rating variation. HF0 seems to polarize more

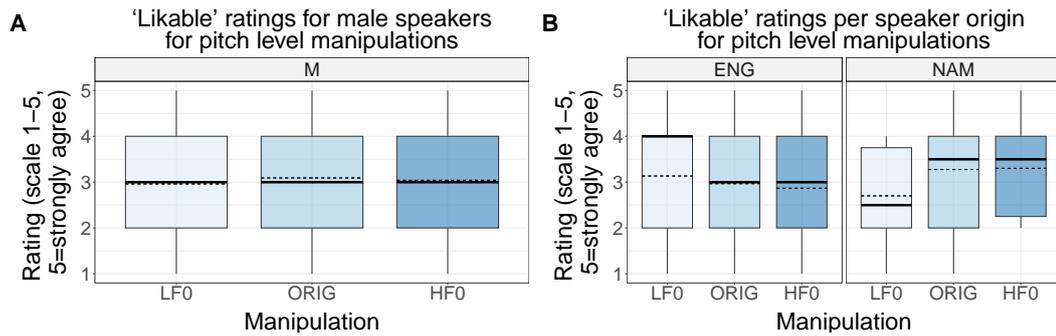


Figure 8.5: The results of the *likable* ratings of the stimuli with pitch level manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

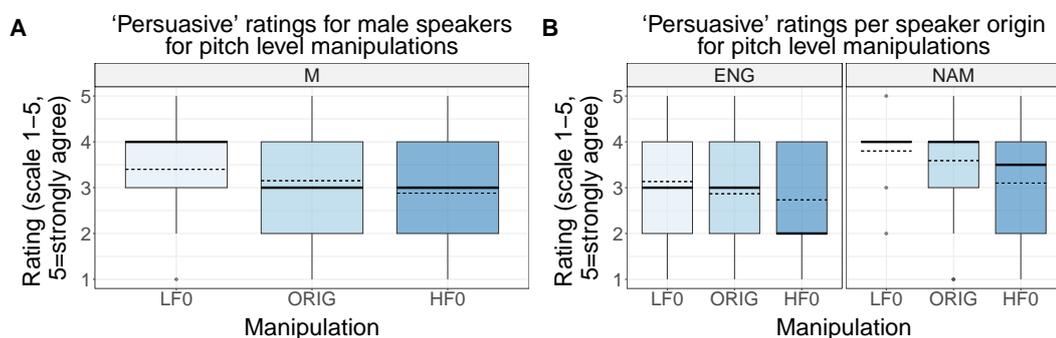


Figure 8.6: The results of the *persuasive* ratings of the stimuli with pitch level manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

for North American speakers as the variation reaches both extreme responses (right panel in Figure 8.6B).

8.4.2 Pitch range

The LMM of the *authentic* ratings revealed no significant interactions or main effects for either of the groups—*Manipulation*, *Gender*, and *Origin* (see Table F.3 in Appendix F). When looking at the *authentic* ratings and pitch range manipulations visually, it seems like female speakers with an increased pitch range (HF0R) are perceived as more authentic than the other manipulations, mainly because of a smaller amount of rating response variation (see left panel in Figure 8.7A). For male speakers, the median of HF0R is slightly higher than the ORIG or decreased pitch range (LF0R) ratings, but otherwise, there is no visual difference between the groups (right panel in Figure 8.7A). Similarly, the visual inspection suggests that HF0R received slightly higher authentic ratings than LF0R for speakers from England, and the HF0R median is also higher for North American speakers than the other manipulations (see Figure 8.7B).

For the *enthusiastic* ratings, the LMM showed a significant main effect of *Ori-*

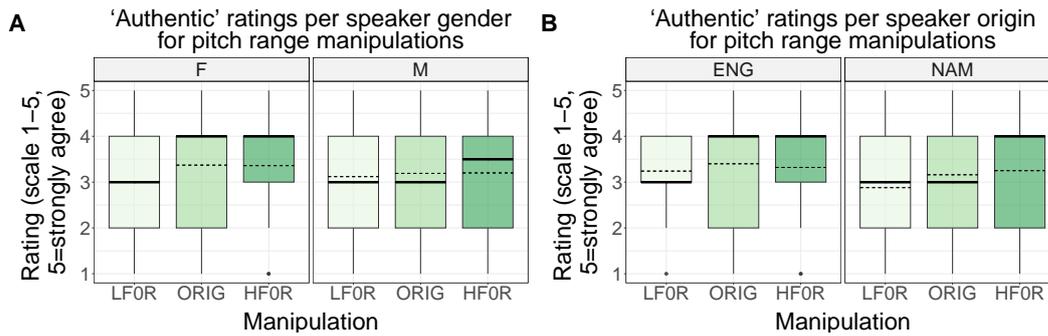


Figure 8.7: The results of the *authentic* ratings of the stimuli with pitch range manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

gin in that speakers from North America received significantly lower enthusiastic ratings than speakers from England (see Table 8.7). This was also confirmed by a post-hoc pairwise comparison ($estimate = 0.7$, $SE = 0.27$, $t\text{-ratio} = 2.57$, $p = .03$). Additionally, there were non-significant trends for *Manipulation* and *Gender*. They suggest that LF0R stimuli tended to be perceived as less enthusiastic than ORIG stimuli, which was confirmed by the EMMs (LF0R = 2.98; ORIG = 3.23; HF0R = 3.05), but not by significant pairwise comparisons. The non-significant trend of *Gender* suggests that male speakers tended to be rated as more enthusiastic than female speakers which was supported by a trend in the pairwise comparisons ($estimate = -0.48$, $SE = 0.26$, $t\text{-ratio} = -1.86$, $p = .09$). Visually, female speakers were rated as generally less enthusiastic than male speakers, especially when LF0R is concerned. For female speakers, LF0R has a similar median rating as ORIG and HF0R, but less variation above the neutral response (left panel in Figure 8.8A). For male speakers, the ratings for LF0R and ORIG center between “agree” (4) and “neither agree nor disagree” (3, the neutral response) and are therefore perceived as a) more enthusiastic than the female speakers, and b) slightly more enthusiastic than the HF0R manipulation which have larger variation (right panel in Figure 8.8A). Speakers from England are generally perceived as more enthusiastic than North American speakers, regardless of manipulation (Figure 8.8).

The LMM for the *likable* ratings did not reveal significant interactions or main effects (all p -values $\geq .2$; see Table F.4 in Appendix F). There are also no visual patterns emerging for male and female speakers, and neither for speakers from England or North America (Figure 8.9). Variation stretches from “agree” to “disagree” with some ratings further into the extremes, and the median (and mean) is around the neutral answer (3). One exception is the rating of HF0R for male speakers where the variation does not extend further up than “agree” (4), suggesting slightly lower likable ratings than the other manipulations. The other exception is the rating of

Table 8.7: The output of the LMM analysis for pitch range and the *enthusiastic* ratings. The intercept is the ORIG stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: ORIG, LF0R, HF0R), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|-------------|-------|------|---------|-----------|---|
| (Intercept) | 3.40 | 0.25 | 13.70 | <.001 | * |
| LF0R | -0.50 | 0.30 | -1.65 | .101 | . |
| HF0R | -0.10 | 0.31 | -0.31 | .756 | . |
| male | 0.44 | 0.25 | 1.74 | .114 | . |
| NAM | -0.76 | 0.26 | -2.91 | .014 | * |
| LF0R × male | 0.32 | 0.30 | 1.05 | .296 | . |
| HF0R × male | -0.19 | 0.33 | -0.56 | .573 | . |
| LF0R × NAM | 0.31 | 0.32 | 0.95 | .342 | . |
| HF0R × NAM | -0.12 | 0.35 | -0.35 | .730 | . |

Signif. codes: * .05, . .1

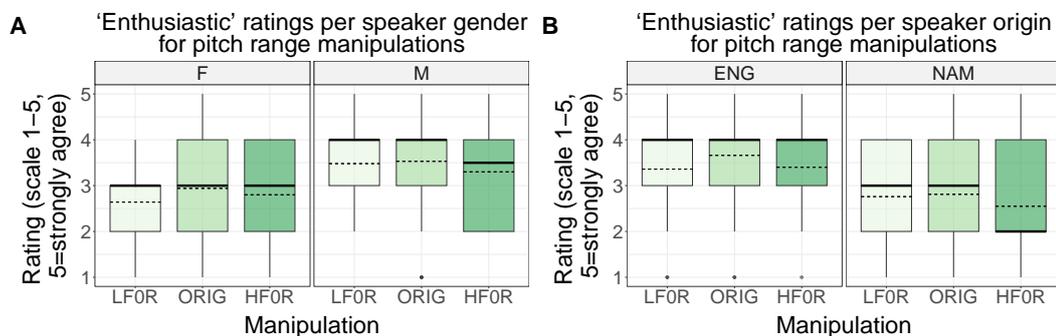


Figure 8.8: The results of the *enthusiastic* ratings of the stimuli with pitch range manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

ORIG stimuli of speakers from England where the median is higher (on “agree” rather than the neutral response) compared to the other stimuli.

In terms of *persuasive* ratings, the LMM again revealed no significant interactions or main effects. There was a non-significant trend, though, that suggests that HF0R stimuli were perceived as more persuasive than ORIG stimuli, but only for North American speakers ($p = .09$; see Table F.5 in Appendix F). Visually, HF0R seems to be perceived as most persuasive for female speakers (median on 4 and variation not reaching the bottom extreme), while ORIG has the same median, but larger variation (left panel in Figure 8.10A). For male speakers, the medians suggest that LF0R may tend to receive the lowest ratings, ORIG a neutral rating, and HF0R the highest ratings, though the variations are the same all across the rating scale, and the means center around the neutral response (right panel in Figure 8.10A). For speakers from England and North America, medians suggest that ORIG and HF0R stimuli may be slightly preferred in terms of persuasiveness (Figure 8.10B). The trend of the LMM suggesting that HF0R stimuli were rated as more persuasive than ORIG stimuli is shown in the form of a smaller rating variation to the more negative ratings.

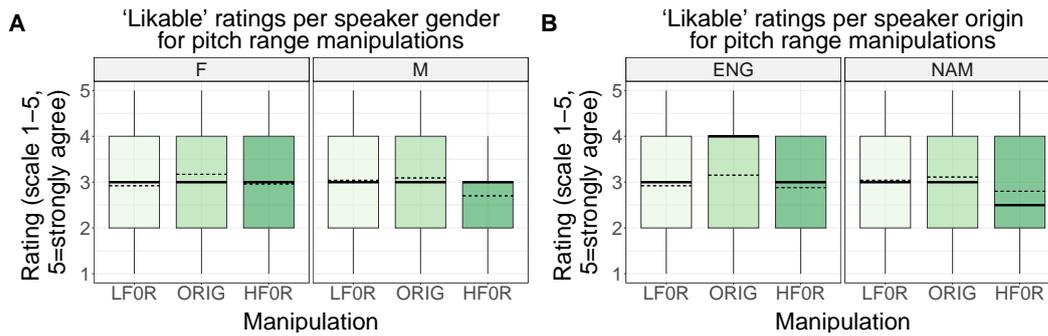


Figure 8.9: The results of the *likable* ratings of the stimuli with pitch range manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

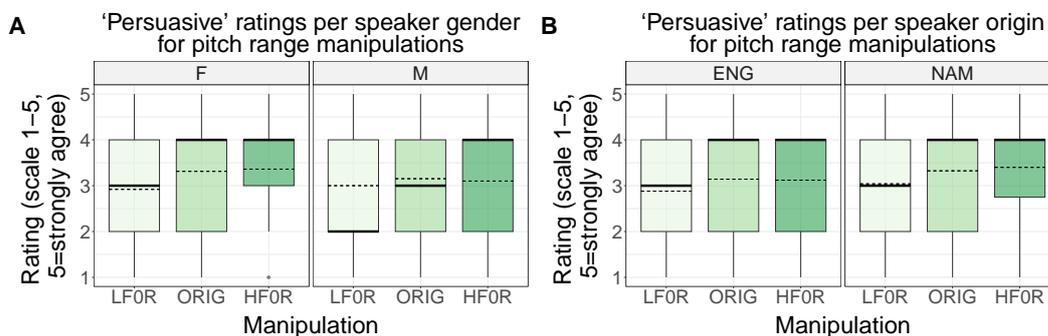


Figure 8.10: The results of the *persuasive* ratings of the stimuli with pitch range manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

8.4.3 Final contour direction

When investigating the effect of final contour direction on the *authentic* ratings, the LMM showed a significant interaction between *Manipulation* and *Origin* (see Table 8.8). It suggests that North American speakers were rated as significantly more authentic than English speakers when the stimulus was rising than when it was falling. For the English speakers, rising was the lowest rated stimulus and falling was rated the highest, meaning that opposite manipulations seem to be preferred for the two speaker groups. However, the interaction also suggests that the rating difference between the two stimulus conditions was significantly larger for speakers from England than speakers from North America. This is not shown in significant pairwise comparisons, but suggested by the estimates of the LMM (ENG: falling = 3.67, plateau = 3.43, rising = 3.14; NAM: falling = 3.13, plateau = 3.35, rising = 3.43). Similarly, there was a non-significant interaction suggesting that plateau stimuli were also higher rated than falling stimuli, but this only applied for North American speakers (see estimates above and Table 8.8).

Visually, it seems like stimuli with plateau or rising final contours were perceived as more authentic (median on 4, left panel in Figure 8.11A) with less variation in

Table 8.8: The output of the LMM analysis for final contour direction and the *authentic* ratings. The intercept is the falling stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: falling, plateau, rising), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|----------------|-------|------|---------|-----------|---|
| (Intercept) | 3.67 | 0.33 | 11.01 | <.001 | * |
| plateau | -0.24 | 0.30 | -0.79 | .431 | |
| rising | -0.53 | 0.40 | -1.33 | .186 | |
| male | -0.18 | 0.33 | -0.55 | .588 | |
| NAM | -0.54 | 0.35 | -1.55 | .138 | . |
| plateau × male | -0.21 | 0.33 | -0.66 | .513 | |
| rising × male | 0.33 | 0.40 | 0.83 | .406 | |
| plateau × NAM | 0.46 | 0.32 | 1.46 | .145 | . |
| rising × NAM | 0.83 | 0.42 | 1.97 | .050 | * |

Signif. codes: * .05, . .1

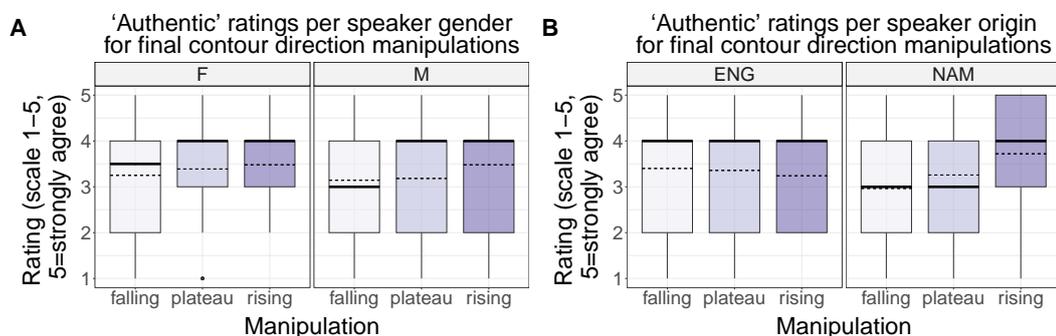


Figure 8.11: The results of the *authentic* ratings of the stimuli with final contour direction manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

rating than falling contours for female speakers. For male speakers (right panel in Figure 8.11A), the median rating mirrors that of female speakers, also suggesting a slight preference for plateau and rising contours in the context of authenticity, but the variation is much wider. For speakers from England, no pattern is discernible, but for speakers from North America it seems that rises were perceived as more authentic than falls or plateaus (Figure 8.11B), in line with the results from the inferential statistics.

In terms of the *enthusiastic* ratings, the LMM revealed no significant interactions, and no significant main effects of *Manipulation* or *Gender*. However, there was a non-significant main effect trend of *Origin* ($p = .09$; see Table F.6 in Appendix F). Post-hoc comparisons revealed that speakers from England tended to be rated as more enthusiastic than speakers from North America ($estimate = 0.59$, $SE = 0.28$, $t\text{-ratio} = 2.08$, $p = .07$). There was no visual pattern discernible concerning the manipulations for female speakers (left panel in Figure 8.12A). For male speakers, falling and rising stimuli tend to be rated as more enthusiastic than a) plateau stimuli, and b) any stimuli of female speakers (right panel in Figure 8.12A).

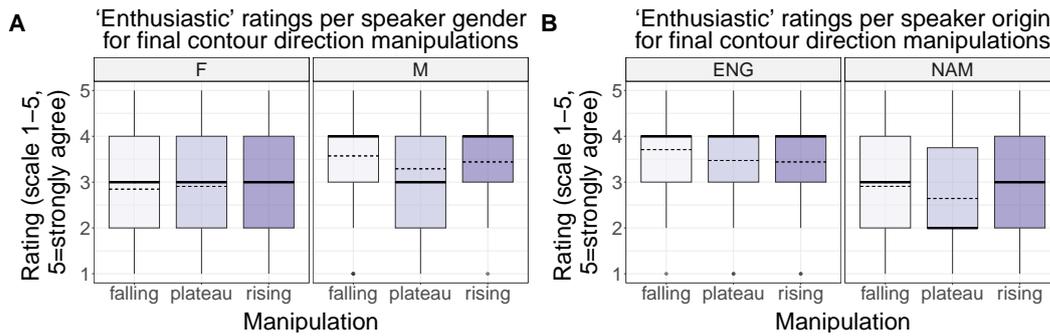


Figure 8.12: The results of the *enthusiastic* ratings of the stimuli with final contour direction manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

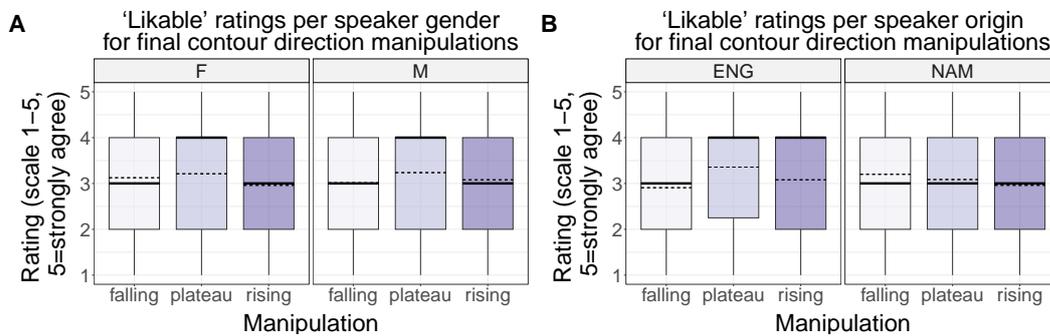


Figure 8.13: The results of the *likable* ratings of the stimuli with final contour direction manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

All three manipulations seem to receive fairly high ratings (median on 4, “agree”) with little variation for speakers from England (left panel in Figure 8.12B). There is more rating variation for speakers from North America, and plateau stimuli seem to receive the lowest enthusiastic ratings here (right panel in Figure 8.12B). This is in line with the trend from the LMM in that speakers from England received overall higher ratings.

The LMM for the *likable* ratings revealed no significant main effects or interactions (all p -values $\geq .5$; see Table F.7 in Appendix F). Visually, the variations are the same for male and female speakers, and speakers from England and North America (see Figure 8.13). Means and most medians are around the neutral response (3), boxes extend between “disagree” and “agree” (2 and 4, respectively), and whiskers reach to the extremes for all manipulations. Some of the medians differ, though—for plateau stimuli of male, female, and English speakers, and rising stimuli for English speakers. For these groups, the median is on “agree” suggesting slightly higher ratings for these stimuli.

Finally, the LMM for the *persuasive* ratings revealed a significant interaction be-

Table 8.9: The output of the LMM analysis for final contour direction and the *persuasive* ratings. The intercept is the falling stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: falling, plateau, rising), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|----------------|-------|------|---------|-----------|---|
| (Intercept) | 2.87 | 0.42 | 6.90 | <.001 | * |
| plateau | 0.49 | 0.30 | 1.62 | .106 | . |
| rising | 0.33 | 0.41 | 0.82 | .415 | |
| male | 0.08 | 0.41 | 0.19 | .855 | |
| NAM | 0.53 | 0.42 | 1.26 | .229 | |
| plateau × male | -0.22 | 0.32 | -0.69 | .491 | |
| rising × male | 0.11 | 0.39 | 0.30 | .768 | |
| plateau × NAM | -0.66 | 0.31 | -2.11 | .036 | * |
| rising × NAM | -0.56 | 0.41 | -1.37 | .172 | |

Signif. codes: * .05, . .1

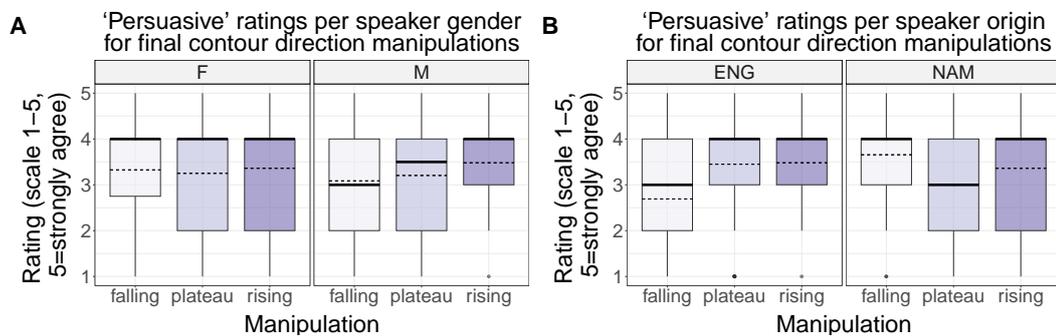


Figure 8.14: The results of the *persuasive* ratings of the stimuli with final contour direction manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

tween *Manipulation* and *Origin* for plateau and NAM (see Table 8.9). This interaction suggests that plateau stimuli received lower *persuasive* ratings than falling stimuli, but this was only the case for North American speakers. For speakers from England, plateau stimuli were higher rated than falling stimuli (estimates from LMM: ENG—falling = 2.87, plateau = 3.36; NAM—falling = 3.4, plateau = 3.23). Additionally, the interaction revealed that the rating difference between plateau and falling stimuli was significantly larger for speakers from England than North American speakers. This is in line with the EMMs from the post-hoc analysis, though there were no significant pairwise comparisons, but the EMMs differ slightly in value from the estimates of the LMM directly (ENG: falling = 2.91, plateau = 3.29, rising = 3.30; NAM: falling = 3.44, plateau = 3.16, rising = 3.27).

Visually, there are no differences between the manipulations for the female speakers. All medians are on “agree” suggesting that the speakers may be perceived as fairly persuasive overall, while the variation of the ratings is wide. This variation is the smallest for the falling stimuli, perhaps suggesting a slight preference of falling contours for the perception of persuasion (see left panel in Figure

8.14A). For male speakers, the medians suggest that falling stimuli get rated less persuasive than plateau stimuli, which get rated less persuasive than rising stimuli—for the rising stimuli, the variation is also much smaller than for the other two manipulations (see right panel in Figure 8.14A). Speakers from England seem to be rated more persuasive with plateau or rising contours, while North American speakers seem to be perceived as more persuasive with falling stimuli (Figure 8.14B), which corresponds to the findings from the LMM.

8.4.4 Speech rate

Regarding the effect of speech rate manipulations on the *authentic* ratings, the LMM showed a significant main effect of *Origin* (see Table 8.10) which suggests that in general, the North American speakers were perceived as less authentic than the speakers from England. While this was not confirmed by the post-hoc pairwise comparisons ($p = .23$), the EMMs suggest generally higher ratings for speakers from England as well (ENG = 3.37; NAM = 3.10). Additionally, though, there was a non-significant trend of an interaction between *Origin* and *Manipulation* (see Table 8.10) which may also have an influence on the main effect. The interaction suggests that MSR stimuli received a higher authenticity rating than the LSR stimuli, but only for North American speakers. For the English speakers, LSR was higher rated than MSR. Additionally, the rating difference tended to be larger for North American speakers. This was also suggested by the EMMs (ENG: LSR = 3.51, MSR = 3.29, HSR = 3.29; NAM: LSR = 2.93, MSR = 3.21, HSR = 3.16), though not the pairwise comparisons.

Visually, the medians of the *authentic* ratings are the same for all manipulations (4, “agree”) for the female speakers. While the variation is smallest for MSR stimuli, the variation here also does not extend to the extreme ratings, perhaps suggesting a slightly more neutral rating, while LSR and HSR received ratings across the board. HSR has a slightly smaller box suggesting perhaps a slight preference for HSR for female speakers (left panel in Figure 8.15A). For male speakers, no pattern is visible (right panel in Figure 8.15A). Speakers from England were rated more authentic for LSR stimuli, and the variation there was also smallest (left panel in Figure 8.15B). There was no pattern for speakers from North America (right panel in Figure 8.15B).

For the *enthusiastic* ratings, the LMM revealed a significant interaction between *Manipulation* and *Origin* as well as a significant main effect of *Origin* which is not interpreted separately as it is also involved in the interaction (see Table 8.11). The interaction suggests that HSR stimuli were rated as more enthusiastic than LSR stimuli for both speakers from England and North America, but that this was only a significant difference for the North American speakers (estimates from the LMM

Table 8.10: The output of the LMM analysis for speech rate and the *authentic* ratings. The intercept is the LSR stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: LSR, MSR, HSR), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|-------------|-------|------|---------|-----------|---|
| (Intercept) | 3.56 | 0.20 | 17.71 | <.001 | * |
| MSR | -0.16 | 0.33 | -0.49 | .628 | |
| HSR | -0.06 | 0.33 | -0.19 | .849 | |
| male | -0.09 | 0.23 | -0.38 | .708 | |
| NAM | -0.59 | 0.24 | -2.43 | .026 | * |
| MSR × male | -0.12 | 0.36 | -0.33 | .744 | |
| HSR × male | -0.31 | 0.36 | -0.87 | .383 | |
| MSR × NAM | 0.50 | 0.33 | 1.53 | .129 | . |
| HSR × NAM | 0.45 | 0.36 | 1.27 | .206 | |

Signif. codes: * .05, . .1

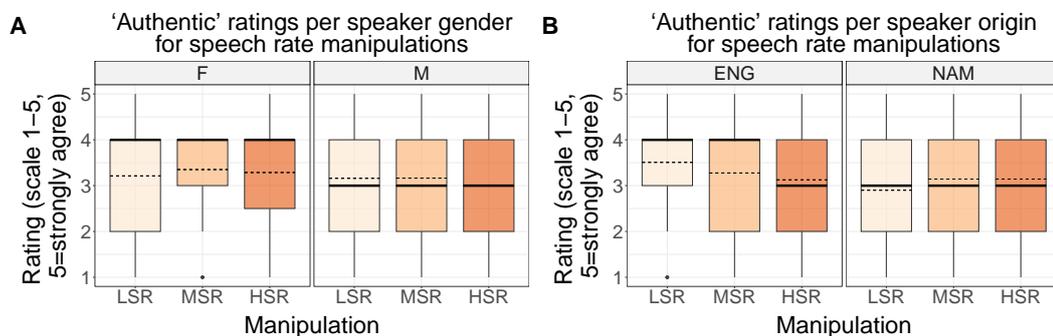


Figure 8.15: The results of the *authentic* ratings of the stimuli with speech rate manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

output: ENG—LSR = 3.33, HSR = 3.36; NAM—LSR = 2.37, HSR = 3.05), which is also implied in the visual inspection of the data (see Figure 8.16). This was also confirmed by the post-hoc pairwise comparisons which reveal that for North American speakers, LSR stimuli were rated as significantly less enthusiastic than HSR (estimate = -0.94, $SE = 0.24$, t -ratio = -4.00, $p = .001$), which was not the case for speakers from England (estimate = -0.29, $SE = 0.25$, t -ratio = -1.19, $p = .84$). Additionally, the pairwise comparisons suggest that LSR stimuli from North American speakers were rated as less enthusiastic than LSR stimuli (estimate = 0.96, $SE = 0.29$, t -ratio = 3.27, $p = .07$, a non-significant trend), MSR stimuli (estimate = 1.13, $SE = 0.32$, t -ratio = 3.49, $p = .03$, a significant difference), and HSR stimuli (estimate = 1.25, $SE = 0.33$, t -ratio = 3.86, $p = .02$, also a significant difference) from speakers from England.

The LMM also revealed a non-significant trend in the interaction between *Gender* and *Manipulation* (see Table 8.11) suggesting that HSR stimuli received higher *enthusiastic* ratings than LSR stimuli, but this difference was only relevant for male speakers (estimates from the LMM output: female—LSR = 3.33, HSR = 3.36; male—LSR = 3.56, HSR = 4.12). Pairwise comparisons revealed that on the one hand, the ratings for LSR stimuli for male speakers were significantly lower than

Table 8.11: The output of the LMM analysis for speech rate and the *enthusiastic* ratings. The intercept is the LSR stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: LSR, MSR, HSR), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) |
|-------------|-------|------|---------|-----------|
| (Intercept) | 3.33 | 0.28 | 11.96 | <.001 |
| MSR | -0.01 | 0.30 | -0.04 | .966 |
| HSR | 0.03 | 0.29 | 0.09 | .927 |
| male | 0.23 | 0.31 | 0.75 | .469 |
| NAM | -0.96 | 0.29 | -3.29 | .008 * |
| MSR × male | 0.36 | 0.34 | 1.05 | .295 |
| HSR × male | 0.53 | 0.32 | 1.64 | .103 · |
| MSR × NAM | 0.40 | 0.31 | 1.29 | .197 |
| HSR × NAM | 0.65 | 0.33 | 1.99 | .048 * |

Signif. codes: * .05, · .1

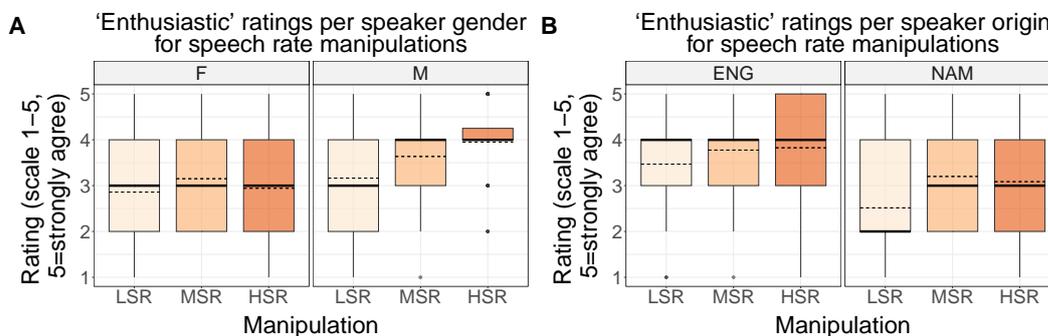


Figure 8.16: The results of the *enthusiastic* ratings of the stimuli with speech rate manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

ratings for HSR stimuli ($estimate = -0.88$, $SE = 0.25$, $t\text{-ratio} = -3.51$, $p = .007$), and also tended to be rated lower than MSR (non-significant trend; $estimate = -0.54$, $SE = 0.22$, $t\text{-ratio} = -2.5$, $p = .13$), but there was no difference between HSR and MSR ratings for male speakers ($estimate = -0.34$, $SE = 0.24$, $t\text{-ratio} = -1.43$, $p = .71$). Additionally, the pairwise comparisons revealed that HSR of male speakers was rated significantly more enthusiastic than LSR of female speakers ($estimate = -1.11$, $SE = 0.33$, $t\text{-ratio} = -3.42$, $p = .04$).

The visual speech rate results suggest no influence of the manipulation for female speakers (left panel in Figure 8.16A)—all manipulations have the same median response, similar means, and variation. For male speakers (right panel in Figure 8.16A), LSR stimuli tend to be perceived as least enthusiastic, but with wide variation. MSR stimuli are perceived as more enthusiastic with less variation, and HSR stimuli as even more enthusiastic (same median, higher mean, even less variation), in line with the results from the inferential statistics. Visually, it seems like speakers from England are generally perceived as more enthusiastic than North American speakers, and—in line with the LMM results—LSR stimuli received the lowest ratings for North American speakers (two right panels in Figure 8.16).

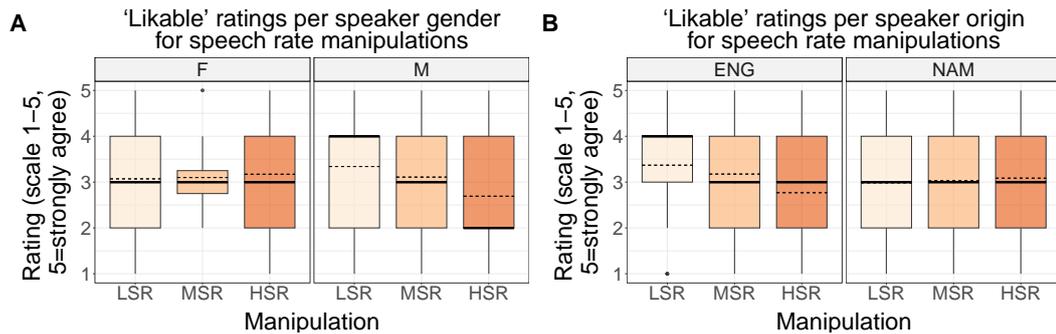


Figure 8.17: The results of the *likable* ratings of the stimuli with speech rate manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

In terms of the *likable* rating, the LMM did not reveal significant interactions or main effects (all p -values $\geq .2$; see Table F.8 in Appendix F). Visually, there is mainly a difference in variation for female speakers (left panel in Figure 8.17A). The medians (and means) for LSR, MSR, and HSR are all on the neutral response (3). However, the response variation is much smaller for MSR, suggesting that LSR and HSR invite more extreme ratings and perhaps polarize more. For male speakers, the variation is the same across manipulations (right panel in Figure 8.17A), but the median for LSR is higher than MSR, and the median of HSR is lower than MSR. That suggests that stimuli with a lower speech rate may be slightly preferred for males when it comes to likability. There is no discernible difference between the three speech rate categories for North American speakers (right panel in Figure 8.17B). MSR and HSR stimuli of speakers from England are similarly rated all over the scale, but LSR here has a) less variation than HSR and MSR, especially to the lower extreme, and b) the median lies higher (see right panel in Figure 8.17B), suggesting that LSR may be perceived as more likable in English speakers.

For the *persuasive* ratings, the LMM revealed no significant interactions or main effects (all p -values $\geq .4$; see Table F.9 in Appendix F). Visually, MSR stimuli may be slightly preferred for female speakers (left panel in Figure 8.18A), though this is only based on a smaller variation; the medians are all on “agree”. For male speakers (right panel in Figure 8.18A) and speakers from England, LSR stimuli seem to be preferred in terms of persuasiveness, while there is no clear pattern visible for North American speakers (Figure 8.18B).

8.5 Results II: Direct charisma ratings and familiarity

This section presents the results of the effect of the acoustic manipulations as well as the familiarity with the speakers on the perception of charisma. Again, both possible differences in speaker gender and speaker origin are included (with the

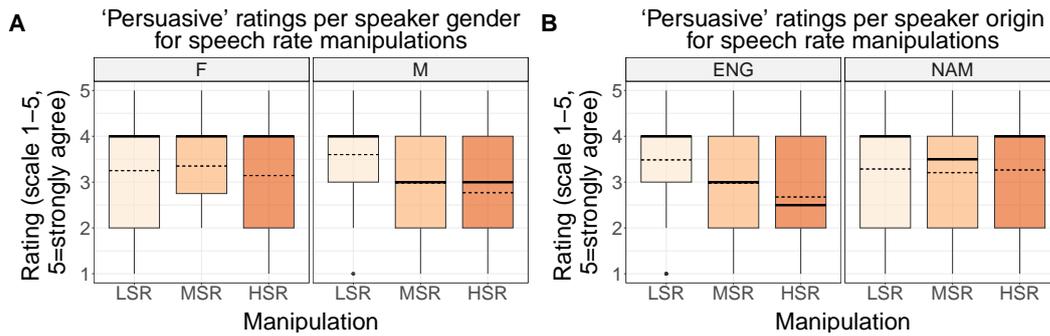


Figure 8.18: The results of the *persuasive* ratings of the stimuli with speech rate manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

exception of the pitch level analyses which only include the male speakers, see below). The different acoustic manipulations are looked at one after the other, both visually and via LMMs and Pearson correlations.

8.5.1 Pitch level

Since the majority of the pitch level manipulations had to be excluded for the female speakers, the data set again only consists of the male speakers' stimuli. The speakers from England and North America are therefore only the male speakers in the sample (three from England, two from North America).

The LMM showed significant main effects for *Origin* and *Familiarity* (see Table 8.12). However, these variables are also involved in significant interactions, so only the interactions are interpreted. There was a significant interaction between *Manipulation* and *Origin* (see Table 8.12) which suggests that HF0 stimuli received lower charisma ratings than ORIG stimuli, but that this was only the case for North American speakers. The rating difference between the two speaker groups was roughly the same (as seen in the estimates of the LMM in Table 8.12 and the EMMs below). The post-hoc pairwise comparisons did not reveal significant group differences, but the EMMs suggest that HF0 was highest rated for speakers from England, and this rating was therefore higher than that for ORIG stimuli, but for speakers from North America, the ORIG stimuli received the highest and the HF0 stimuli the lowest charisma ratings (ENG: LF0 = 3.48, ORIG = 3.60, HF0 = 3.83; NAM: LF0 = 3.98, ORIG = 4.10, HF0 = 3.88).

Additionally, there was a significant interaction between *Origin* and *Familiarity* (see Table 8.12) which suggests that speakers who were familiar to the listeners were rated as less charismatic than known speakers, but only for North American speakers. For the English speakers, those who were familiar to the listeners were rated as more charismatic than those who were known. Additionally, the rating difference between the two familiarity categories was significantly larger for speakers

Table 8.12: The output of the LMM analysis for pitch level and the *charismatic* ratings. The intercept is the ORIG stimulus for speakers from England who are known to the participants. The independent variables are *Manipulation* (three levels: ORIG, LF0, HF0), *Origin* (two levels: ENG, NAM), and *Familiarity* (four levels: I know the speaker, They seem familiar, Unsure, I do not know the speaker).

| | Est. | SE | t value | Pr(> t) | |
|---------------------------------|-------|------|---------|----------|---|
| (Intercept) | 3.48 | 0.34 | 10.35 | <.001 | * |
| LF0 | -0.30 | 0.31 | -0.95 | .341 | |
| HF0 | 0.22 | 0.31 | 0.70 | .486 | |
| NAM | 1.07 | 0.51 | 2.09 | .051 | * |
| They seem familiar | 0.79 | 0.36 | 2.20 | .038 | * |
| Unsure | 0.23 | 0.40 | 0.57 | .577 | |
| I do not know the speaker | -0.52 | 0.32 | -1.64 | .112 | . |
| LF0 × NAM | 0.01 | 0.22 | 0.03 | .976 | |
| HF0 × NAM | -0.45 | 0.22 | -2.07 | .040 | * |
| LF0 × They seem familiar | -0.09 | 0.40 | -0.23 | .819 | |
| HF0 × They seem familiar | -0.16 | 0.39 | -0.42 | .675 | |
| LF0 × Unsure | 0.33 | 0.43 | 0.77 | .444 | |
| HF0 × Unsure | 0.28 | 0.44 | 0.64 | .521 | |
| LF0 × I do not know the speaker | 0.46 | 0.34 | 1.34 | .182 | |
| HF0 × I do not know the speaker | -0.05 | 0.34 | -0.15 | .882 | |
| NAM × They seem familiar | -1.22 | 0.43 | -2.82 | .006 | * |
| NAM × Unsure | -0.61 | 0.48 | -1.28 | .205 | |
| NAM × I do not know the speaker | -0.46 | 0.41 | -1.13 | .265 | |

Signif. codes: * .05, . .1

from England. This was also suggested by the EMMs (ENG: I know the speaker = 3.45, They seem familiar = 4.15, Unsure = 3.88, I do not know the speaker = 3.07; NAM: I know the speaker = 4.37, They seem familiar = 3.86, Unsure = 4.19, I do not know the speaker = 3.52), but not by pairwise comparisons. For both speaker groups, the stimuli were rated least charismatic when the speaker was unknown to the listeners. This was a significant pairwise comparison for speakers from England in comparison to “They seem familiar” (estimate = 1.09, $SE = 0.29$, t -ratio = 3.93, $p = .02$), but there were no such pairwise comparisons for North American speakers (all p -values > .5). In general, this suggests that unknown speakers were perceived as less charismatic than known speakers, which was also confirmed by a Pearson correlation between the charisma ratings and the (numeric) familiarity ratings ($r = 0.3$, $t = 5.51$, $p < .001$).

Looking at the data visually, no clear pattern is visible for the male speakers (Figure 8.19A), though all stimuli manipulations receive fairly high charisma ratings. For the speakers from England, the HF0 stimuli seem to be preferred, though the rating medians for ORIG and LF0 are equally high on “agree” (4), but the variation is reaching into the “strongly disagree” response (1), which is not the case for HF0 (see left panel in Figure 8.19B). For the North American speakers, on the other hand, LF0 and ORIG stimuli seem to be rated as more charismatic with a higher mean and smaller variation than HF0, though HF0 is also rated highly, but with larger variation and lower mean (see right panel in Figure 8.19B), in line with the results from the inferential statistics.

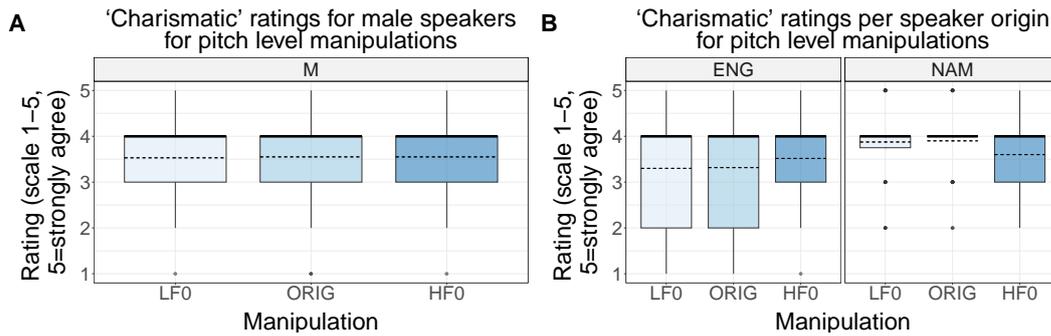


Figure 8.19: The results of the *charismatic* ratings of the stimuli with pitch level manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

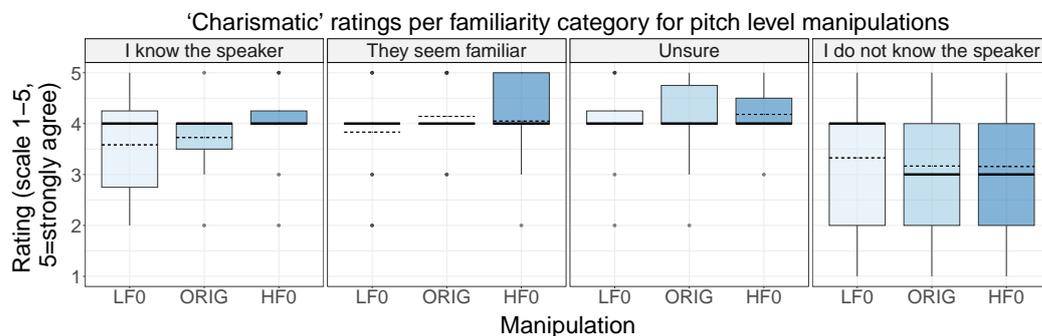


Figure 8.20: The results of the *charismatic* ratings of the stimuli with pitch level manipulations per familiarity category. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

In terms of familiarity, visually (see Figure 8.20), the median responses suggest that charisma is rated higher (on “agree”) if the participants indicated they knew the speaker, the speaker was familiar, or the participant was unsure, which is also in line with the LMM results. Especially with HF0 when the speaker was known and familiar, and LF0 and ORIG when the speaker was familiar, the variation is slim to non-existing, suggesting overall high charisma ratings. But in these three familiarity categories, there are no ratings with the “strongly disagree” response, again suggesting overall quite high charisma ratings. The medians for the charisma ratings when the speaker was not known by the participants are on the neutral “neither agree nor disagree” response (as are the means), and the variation ranges from the “strongly agree” extreme (5) to the “strongly disagree” extreme (1), suggesting overall lower charisma ratings and —more importantly—also more extreme ratings. The figures show that the manipulations polarize more when the speaker was unknown, and seem to mostly have an effect with known or familiar speakers.

8.5.2 Pitch range

The LMM revealed significant main effects of *Manipulation* and *Familiarity*, but both are involved in significant interactions (see Table 8.13), which is why only the interactions are interpreted. There was a significant interaction between *Manipulation* and *Gender* (see Table 8.13) which suggests that LF0R stimuli were rated as less charismatic than ORIG stimuli for both male and female speakers. The rating difference between the two manipulations was significantly smaller for male speakers, though, suggesting that the different stimuli may be less impactful or relevant for male than female speakers. This is also confirmed by the EMMs and pairwise comparisons, which revealed that the LF0R receive the lowest charisma ratings for both male and female speakers (female: LF0R = 2.87, ORIG = 3.38, HF0R = 3.27; male: LF0R = 3.62, ORIG = 3.77, HF0R = 3.91), but that the differences between the manipulations were only significant for the female speakers when ORIG is compared to LF0R (estimate = 0.51, $SE = 0.14$, t -ratio = 3.69, $p = .003$), and a non-significant trend for the comparison between LF0R and HF0R (estimate = -0.4, $SE = 0.14$, t -ratio = -2.79, $p = .06$). There were no significant pairwise comparisons for the male speakers (all p -values > .3).

There was a second significant interaction between *Manipulation* and *Familiarity* (see Table 8.13) which suggests that LF0R stimuli were rated significantly less charismatic than the ORIG stimuli when the speaker was not known to the listener, and also when the participants were unsure if they recognized a speaker (non-significant trend). However, the rating difference between the two stimuli was larger when the speaker was known than when they were unknown (a significant difference) or the participants were unsure (a trend). The EMMs suggest that if the speaker was known, the LF0R stimuli were rated less charismatic than ORIG and HF0R stimuli which was not the case when the speaker was not known (I know the speaker: LF0R = 2.96, ORIG = 3.71, HF0R = 3.60; They seem familiar: LF0R = 3.54, ORIG = 3.92, HF0R = 3.96; Unsure: LF0R = 3.53, ORIG = 3.64, HF0R = 3.49; I do not know the speaker: LF0R = 2.94, ORIG = 3.04, HF0R = 3.30). The pairwise comparisons also revealed that within the group of speakers that were not known, the LF0R stimuli received significantly lower charisma ratings than the HF0R stimuli (estimate = -0.37, $SE = 0.11$, t -ratio = -3.37, $p = .04$), while those two stimuli did not differ significantly from the ORIG stimuli (LF0R: estimate = 0.11, $SE = 0.11$, t -ratio = 1.02, $p = 1$; HF0R: estimate = -0.26, $SE = 0.11$, t -ratio = -2.37, $p = .43$). Additionally, speakers who were unknown to the participants (with the stimuli ORIG and LF0R) were rated as less charismatic than speakers who seemed familiar (with stimuli ORIG and HF0R) while there were no significant differences between known and unknown speakers (see Table G.1 in Appendix G for the relevant pairwise comparisons). A Pearson correlation analysis between charisma rating and

Table 8.13: The output of the LMM analysis for pitch range and the *charismatic* ratings. The intercept is the ORIG stimulus for female speakers from England who are known to the participants. The independent variables are *Manipulation* (three levels: ORIG, LF0R, HF0R), *Gender* (two levels: female, male), *Origin* (two levels: ENG, NAM), and *Familiarity* (four levels: I know the speaker, They seem familiar, Unsure, I do not know the speaker).

| | Est. | SE | t value | Pr(> t) | |
|----------------------------------|-------|------|---------|-----------|---|
| (Intercept) | 3.52 | 0.41 | 8.56 | <.001 | * |
| LF0R | -0.90 | 0.29 | -3.14 | .002 | * |
| HF0R | -0.24 | 0.29 | -0.82 | .410 | |
| male | 0.23 | 0.41 | 0.58 | .568 | |
| NAM | 0.15 | 0.49 | 0.30 | .766 | |
| They seem familiar | 0.48 | 0.41 | 1.17 | .254 | |
| Unsure | -0.23 | 0.41 | -0.56 | .581 | |
| I do not know the speaker | -0.81 | 0.36 | -2.23 | .039 | * |
| LF0R × male | 0.35 | 0.17 | 2.12 | .035 | * |
| HF0R × male | 0.24 | 0.18 | 1.33 | .185 | |
| LF0R × NAM | -0.05 | 0.17 | -0.32 | .747 | |
| HF0R × NAM | 0.01 | 0.18 | 0.05 | .958 | |
| LF0R × They seem familiar | 0.37 | 0.34 | 1.11 | .267 | |
| HF0R × They seem familiar | 0.16 | 0.34 | 0.47 | .637 | |
| LF0R × Unsure | 0.64 | 0.35 | 1.83 | .068 | . |
| HF0R × Unsure | -0.03 | 0.37 | -0.08 | .940 | |
| LF0R × I do not know the speaker | 0.64 | 0.29 | 2.20 | .028 | * |
| HF0R × I do not know the speaker | 0.38 | 0.30 | 1.28 | .202 | |
| male × They seem familiar | 0.06 | 0.35 | 0.17 | .868 | |
| male × Unsure | 0.48 | 0.36 | 1.32 | .189 | |
| male × I do not know the speaker | 0.11 | 0.31 | 0.36 | .717 | |
| NAM × They seem familiar | -0.62 | 0.44 | -1.41 | .169 | |
| NAM × Unsure | -0.17 | 0.46 | -0.36 | .719 | |
| NAM × I do not know the speaker | 0.17 | 0.42 | 0.40 | .690 | |

Signif. codes: * .05, . .1

familiarity showed a positive correlation ($r = 0.27$, $t = 6.83$, $p < .001$) which equally suggests that more familiar speakers were rated as more charismatic.

Visually, the median (and mean) of the LF0R stimuli is slightly lower than for the other manipulations for female speakers, while there are no obvious differences for male speakers (Figure 8.21A). This is in line with the results from the LMM. The visual inspection of the data additionally suggests that the stimuli of the male speakers were generally rated as more charismatic than those of the female speakers. In particular, the variation of ratings—except for two outliers—does not extend to the “strongly disagree” response for the male speakers. There are no obvious patterns in terms of manipulation for speakers from England and North America (Figure 8.21B), in line with the lack of significant results from the LMM. However, the median responses for ORIG and HF0R are slightly higher than LF0R for speakers from England, and there the variation of HF0R also has a smaller variation, perhaps suggesting a slight preference. For North American speakers, the ORIG stimuli have higher median ratings with smaller variation than the other two manipulations.

When looking at the familiarity categories and the charisma ratings, for participants who indicated they knew the speaker, ORIG and HF0R stimuli received especially high ratings with low variation. The LF0R stimuli have more variation in

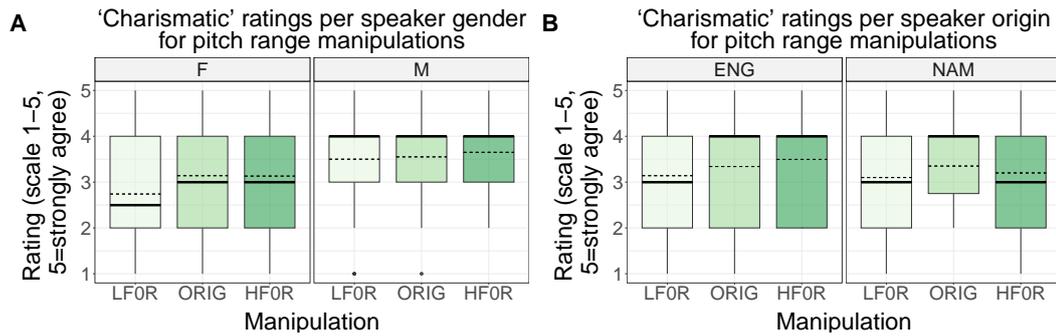


Figure 8.21: The results of the *charismatic* ratings of the stimuli with pitch range manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

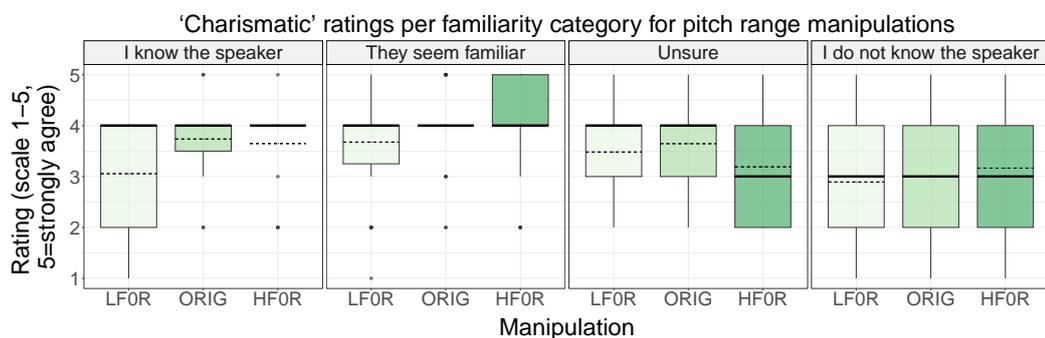


Figure 8.22: The results of the *charismatic* ratings of the stimuli with pitch range manipulations per familiarity category. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

charisma responses, especially to the negative end of the rating scale (see left panel in Figure 8.22). For participants who thought the speaker sounded familiar to them, all three manipulations were rated as charismatic, especially HF0R stimuli (second panel from the left in Figure 8.22). If participants were unsure if they knew the speakers, HF0R is seems to be slightly lower rated than the other two categories, but all three seem to be generally rated as fairly charismatic (second panel from the right in Figure 8.22). If the participants did not know the speaker (right panel in Figure 8.22), the ratings were all over the scale independent of the manipulation, and the means and medians were lower than the other categories, in line with the LMM results. Manipulation effects are again clearer with familiar speakers.

8.5.3 Final contour direction

The LMM for the final contour direction had significant main effects of *Manipulation* and *Familiarity*, though they are not reported separately as these variables were also involved in significant interactions. There was a significant interaction between *Manipulation* and *Gender* (see Table 8.14) which suggests that plateau stimuli were rated as less charismatic than falling stimuli, but only for male speakers. Female

speakers were rated as most charismatic with plateau stimuli and as less charismatic with falling stimuli. Additionally, the estimates of the LMM (see Table 8.14) suggest that the rating difference between the two manipulations is significantly smaller for male than female speakers. While there were no significant pairwise comparisons, the EMMs suggest that male speakers were rated most charismatic with rising stimuli, and female speakers as most charismatic with plateau stimuli (female: falling = 3.17, plateau = 3.43, rising = 3.34; male: falling = 3.81, plateau = 3.71, rising = 4.01).

Furthermore, there was a significant interaction between *Manipulation* and *Familiarity* which suggests that rising stimuli were generally rated as more charismatic than falling stimuli (with the exception of unknown speakers where rising stimuli were predicted to be rated less charismatic, but this difference was minuscule). However, the rating difference between the two stimuli was significantly larger when the speaker was known than the other familiarity categories. Pairwise comparisons additionally reveal that if the speaker was unknown, all manipulations were rated significantly less charismatic than all manipulations of familiar speakers. But again, there were no significant pairwise comparisons between known and unknown speakers, again suggesting that known speakers may also have a detrimental effect on charisma perception. Table 8.15 provides the significant pairwise comparisons. All other comparisons have p -values above .05 significance level.

Finally, there was also an interaction between *Origin* and *Familiarity* which suggests that familiar speakers (significant) and speakers the participants were unsure of (trend) were perceived as less charismatic than when the speakers were known, but only for the North American speakers (see Table 8.14 above for the statistical values). For English speakers, the known speakers received much lower ratings than the other two familiarity categories mentioned here, and the rating differences between known and familiar speakers are significantly larger for speakers from England as well. The same trend emerges for known speakers and speakers the participants were unsure about. This is also shown by the EMMs (ENG: I know the speaker = 3.45, They seem familiar = 4.03, Unsure = 3.67, I do not know the speaker = 2.91; NAM: I know the speaker = 3.99, They seem familiar = 3.71, Unsure = 3.61, I do not know the speaker = 3.25), but not pairwise comparisons. The pairwise comparisons did reveal that for speakers from England, the charisma ratings for unknown speakers were significantly lower than for familiar speakers (estimate = 1.12, $SE = 0.22$, t -ratio = 5.2, $p < .001$) and speakers the participants were unsure about (estimate = 0.76, $SE = 0.19$, t -ratio = 3.92, $p = .009$), but there was no significant rating difference between speakers that are known and unknown (estimate = 0.54, $SE = 0.39$, t -ratio = 1.39, $p = .84$). The higher charisma ratings of recognized speakers compared to unknown speakers was confirmed by a positive Pearson correlation between charisma rating and familiarity response ($r = 0.31$, $t = 7.95$, $p < .001$).

Table 8.14: The output of the LMM for final contour direction and the *charismatic* ratings. The intercept is the falling stimulus for female speakers from England who are known to the participants. The independent variables are *Manipulation* (three levels: falling, plateau, rising), *Gender* (two levels: female, male), *Origin* (two levels: ENG, NAM), and *Familiarity* (four levels: I know the speaker, They seem familiar, Unsure, I do not know the speaker).

| | Est. | SE | t value | Pr(> t) | |
|-------------------------------------|-------|------|---------|-----------|---|
| (Intercept) | 2.94 | 0.40 | 7.28 | <.001 | * |
| plateau | 0.30 | 0.28 | 1.07 | .286 | |
| rising | 0.60 | 0.28 | 2.17 | .031 | * |
| male | 0.54 | 0.39 | 1.37 | .178 | |
| NAM | 0.50 | 0.44 | 1.13 | .263 | |
| They seem familiar | 0.79 | 0.39 | 2.00 | .048 | * |
| Unsure | 0.33 | 0.39 | 0.84 | .403 | |
| I do not know the speaker | -0.38 | 0.36 | -1.05 | .298 | |
| plateau × male | -0.37 | 0.16 | -2.24 | .025 | * |
| rising × male | 0.02 | 0.16 | 0.12 | .901 | |
| plateau × NAM | 0.02 | 0.16 | 0.14 | .889 | |
| rising × NAM | 0.08 | 0.16 | 0.46 | .643 | |
| plateau × They seem familiar | -0.00 | 0.33 | -0.01 | .991 | |
| rising × They seem familiar | -0.72 | 0.33 | -2.20 | .028 | * |
| plateau × Unsure | -0.10 | 0.33 | -0.30 | .763 | |
| rising × Unsure | -0.49 | 0.34 | -1.47 | .142 | . |
| plateau × I do not know the speaker | -0.08 | 0.28 | -0.27 | .787 | |
| rising × I do not know the speaker | -0.64 | 0.28 | -2.30 | .022 | * |
| male × They seem familiar | 0.07 | 0.33 | 0.20 | .843 | |
| male × Unsure | 0.18 | 0.34 | 0.52 | .604 | |
| male × I do not know the speaker | 0.15 | 0.30 | 0.51 | .612 | |
| NAM × They seem familiar | -0.85 | 0.39 | -2.20 | .031 | * |
| NAM × Unsure | -0.60 | 0.41 | -1.48 | .144 | . |
| NAM × I do not know the speaker | -0.20 | 0.37 | -0.53 | .595 | |

Signif. codes: * .05, . .1

The charisma ratings for the final contour direction manipulations show very few visual patterns (Figure 8.23). Mostly, variation ranges from “agree” to “disagree” with the whiskers ranging to the extreme ratings. Means tend to be close to the neutral answer. Male speakers, however, overall received higher charisma ratings than female speakers, with higher medians, means, and smaller variation. There seems to be no rating difference between manipulations for speakers from England. For North American speakers, plateau stimuli seem to be perceived as most charismatic (high median, small variation range), and falling stimuli have a lower median than the other two manipulations.

For familiarity, charisma rating medians were higher if the participants knew the speakers, the speakers seemed familiar, or the participants were unsure, compared to the median ratings of the “I do not know the speaker” category (Figure 8.23), in line with the LMM results. Especially the rising stimuli were rated as charismatic when the speakers were known or seemed familiar. When the speakers seemed familiar, the falling stimuli were equally highly rated. For unknown speakers, the responses ranged across the scale.

Table 8.15: The EMMs for final contour direction and the *charismatic* ratings. Only the significant pairwise comparisons are included, between the speakers that were identified as “They seem familiar” (= familiar) and the speakers that were identified as “I do not know the speaker” (= unknown).

| | contrast | estimate | SE | t.ratio | p.value |
|--|--|----------|------|---------|---------|
| | falling × familiar - falling × unknown | 0.80 | 0.21 | 3.81 | .013 |
| | falling × familiar - plateau × unknown | 0.75 | 0.21 | 3.54 | .029 |
| | falling × familiar - rising × unknown | 0.79 | 0.21 | 3.76 | .016 |
| | plateau × familiar - falling × unknown | 0.92 | 0.19 | 4.96 | .001 |
| | plateau × familiar - plateau × unknown | 0.87 | 0.19 | 4.60 | .002 |
| | plateau × familiar - rising × unknown | 0.91 | 0.19 | 4.89 | .001 |
| | rising × familiar - falling × unknown | 0.72 | 0.19 | 3.74 | .022 |
| | rising × familiar - plateau × unknown | 0.68 | 0.20 | 3.43 | .047 |
| | rising × familiar - rising × unknown | 0.71 | 0.19 | 3.66 | .027 |

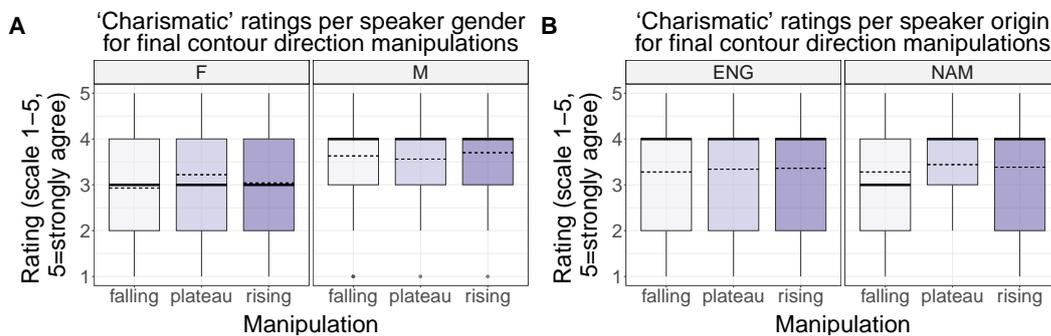


Figure 8.23: The results of the *charismatic* ratings of the stimuli with final contour manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

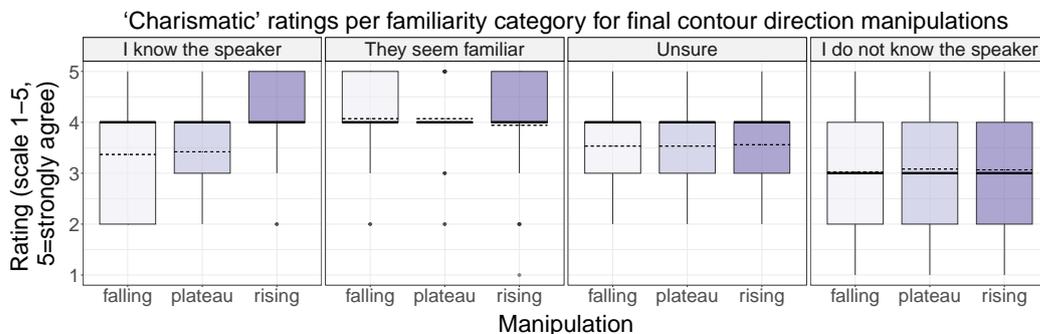


Figure 8.24: The results of the *charismatic* ratings of the stimuli with final contour direction manipulations per familiarity category. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

8.5.4 Speech rate

The LMM for the speech rate analysis revealed a significant main effect of *Familiarity* ($p = .053$; see Table F.10 in Appendix F) which suggests that unknown speakers were rated as less charismatic than known speakers. *Familiarity* was involved in some non-significant trends in interaction, though: with *Gender*, suggesting that familiar speakers were also rated as less charismatic than known speakers, but only

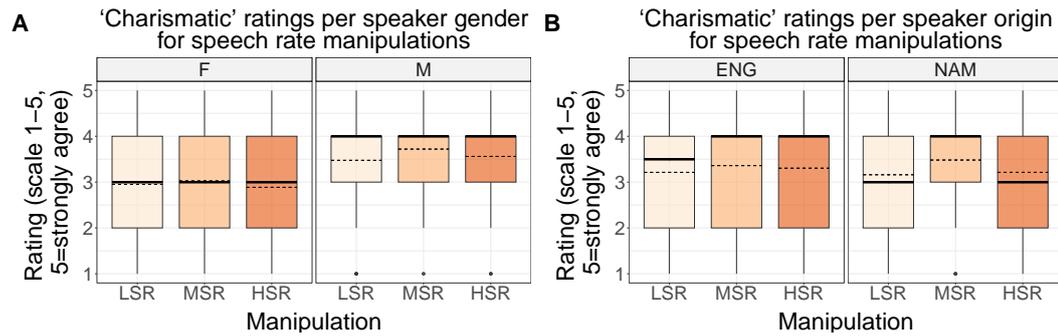


Figure 8.25: The results of the *charismatic* ratings of the stimuli with speech rate manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

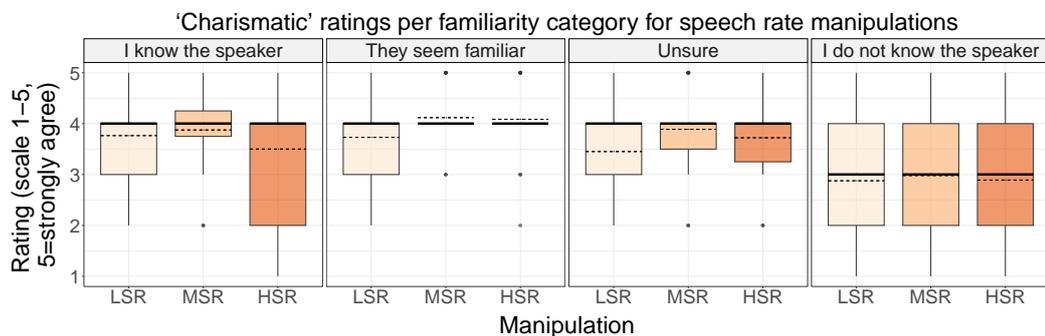


Figure 8.26: The results of the *charismatic* ratings of the stimuli with speech rate manipulations per familiarity category. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

for male speakers, and the rating difference between male and female speakers was much larger when they were known rather than familiar. There was another non-significant trend interaction with *Manipulation* which suggests that HSR stimuli tended to be rated more charismatic when speakers were familiar or the participants were unsure if they recognized them, but the opposite was the case when the speakers were known. Additionally, the rating difference between the stimuli was larger for known speakers. This was not corroborated by pairwise comparisons. A Pearson correlation analysis revealed a significant positive correlation between charisma rating and familiarity ($r = 0.36$, $t = 9.06$, $p < .001$) in line with the main effect of the LMM analysis.

Visually, the charisma ratings for the speech rate manipulations were very similar to those of the final contour direction manipulations (Section 8.5.3), and showed very few patterns (Figure 8.25). Mostly, variation ranges from “agree” to “disagree” with the whiskers ranging to the extreme ratings. Means tend to be close to the neutral answer. Male speakers overall received higher charisma ratings than female speakers, with higher medians, means, and smaller variation. For North American speakers, MSR stimuli seem to be perceived as more charismatic (high median, small variation range) than the other two manipulations.

In terms of familiarity (Figure 8.26), the charisma ratings were lower and more variable if the participants indicated they did not know the speakers. In the other three familiarity categories, MSR stimuli are always rated as charismatic with little rating variation, so are LSR stimuli. If the speaker seemed familiar or participants were unsure, the HSR stimuli were equally highly rated. For HSR stimuli of speakers that were recognized, the variation ranged between both extreme rating options, suggesting that this manipulation polarizes especially if the speaker is known.

8.6 General discussion

8.6.1 Summary of the experiment results

Table 8.16 offers an overview of the most likely candidates of acoustic characteristics evoking the different attributes. The most likely candidates are based on the visual differences in the figures in the two previous sections, sometimes only based on a difference in variation of responses compared between the manipulations. When there was a significant effect or a non-significant trend in the inferential statistics, this is explicitly marked in the table.

In terms of pitch level (note that this is only the case for the male speakers in the sample), the stimuli with lowered pitch level seem to be preferred, especially when it comes to authenticity, persuasiveness, and (to a degree) charisma. Mostly relevant is that the increased pitch level (HF0) was rated lowest in terms of authenticity and persuasiveness, which was even significant for authenticity. In terms of charisma ratings, this is dependent on the speaker origin: North American speakers were also perceived as more charismatic with unchanged or decreased pitch level, while speakers from England were (in contrast to the *authentic* and *persuasive* ratings) perceived as more charismatic with increased pitch level. Similarly, unchanged and increased stimuli received significantly higher enthusiasm ratings than the decreased pitch level, irrespective of *Gender* or *Origin*, and in contrast to the *authentic* and *persuasive* ratings. For likability, there seems to be a suggestion of a cultural difference, though this was not significant. Visually, LF0 seems to be preferred by the raters for speakers from England, but ORIG or HF0 for speakers from North America, which is opposite to the results of the *charismatic* ratings.

For pitch range, again authenticity and persuasiveness seem to be closely related in terms of preferred acoustic characteristics, but charisma also aligns. For the three attributes, the widened pitch range was preferred. ORIG stimuli seem to be preferred when rating enthusiasm, as there was a trend suggesting that LF0R was lower rated than ORIG, and the EMMs also suggest that LF0R and HF0R are rated similarly in this case. For likability, there was no obvious pattern, though for

Table 8.16: An overview of the results from the influence of the prosodic manipulations on the different charisma-adjacent attributes *authentic* (AU), *enthusiastic* (EN), *likable* (LI), and *persuasive* (PE), as well as the direct *charismatic* ratings (CH). Unless marked, the given feature characteristic represents the most likely candidates to evoke highest attribute ratings based on visual results from the descriptive statistics. A ⁺ marks a non-significant trend (where $p > .05$ but $< .15$), and a * marks a significant main effect or interaction from the inferential statistics.)

| | | Features | | | | |
|-----------|---------------|---|---|---|---|---|
| | | Pitch level | Pitch range | Final contour | Speech rate | |
| AU | female | — | HF0R | falling | no pattern | |
| | male | LF0 | HF0R | rising | no pattern | |
| | ENG | LF0 | HF0R | falling* | > NAM* | |
| | NAM | LF0 | HF0R | rising* | < ENG* | |
| | <i>Manip.</i> | LF0/ORIG* | — | — | — | |
| EN | female | — | < male ⁺ | no pattern | no pattern | |
| | male | ORIG/HF0 | > female ⁺ | falling/rising | HSR/MSR* | |
| | ENG | > NAM ⁺ | > NAM* | > NAM ⁺ | > NAM ⁺ | |
| | NAM | < ENG ⁺ | < ENG* | < ENG ⁺ | HSR*; < ENG ⁺ | |
| | <i>Manip.</i> | ORIG/HF0* | ORIG ⁺ | — | — | |
| LI | female | — | no pattern | no pattern | no pattern | |
| | male | no pattern | ORIG/LF0R | no pattern | LSR | |
| | ENG | LF0 | no pattern | no pattern | LSR | |
| | NAM | ORIG/HF0 | no pattern | no pattern | no pattern | |
| | <i>Manip.</i> | — | — | — | — | |
| PE | female | — | HF0R | falling | MSR | |
| | male | LF0 | HF0R | rising | LSR | |
| | ENG | ORIG/LF0 | ORIG/HF0R | plateau* | LSR | |
| | NAM | LF0 | HF0R | falling* | no pattern | |
| | <i>Manip.</i> | — | — | — | — | |
| CH | female | — | HF0R/ORIG*; < male | plateau*; < male | no pattern; < male | |
| | male | no pattern | HF0R*; > female | rising*; > female | no pattern; > female | |
| | ENG | HF0* | HF0R | no pattern | no pattern | |
| | NAM | LF0/ORIG* | ORIG | plateau | MSR | |
| | Familiarity | Known/ Familiar/ Unsure > Unknown* | Known/ Familiar/ Unsure > Unknown* | Known/ Familiar/ Unsure > Unknown* | Familiar/ Unsure > Known/ Unknown* | Known/ Familiar/ Unsure > Unknown* |

male speakers, LF0R and ORIG may be slightly preferred over HF0R, which would be the opposite of the *authentic*, *persuasive*, and *charismatic* ratings.

Again, there is a similarity in acoustic features between *authentic*, *persuasive*, and *charismatic* for the final contour direction ratings, at least in terms of *Gender*. Female speakers tended to be rated as more authentic, persuasive, and charismatic when the final contour in the stimuli was falling. The male speakers received higher ratings with rising stimuli. For charisma, the differences were significant. In particular, the rating differences were significantly smaller for male than female speakers (almost negligible, even). When looking at the speaker origin, the two attributes *authentic* and *persuasive* were oppositely rated (there were no clear patterns for the *charismatic* ratings): English speakers were rated more authentic with falling stimuli, and North American speakers with rising stimuli. At the same time, North

American speakers were rated as significantly more persuasive with falling contours, and English speakers with plateaus. The rating difference was significantly larger for English speakers than for North American speakers in both interactions. There were no obvious patterns regarding the most likely manipulations for enthusiasm and likability.

Finally, speech rate elicited the only *Origin* effect for authenticity in that these stimuli elicited significantly lower *authentic* ratings for North American speakers than speakers from England. This time, likability and persuasiveness seem to pattern together more in terms of acoustic characteristics, though only for male speakers and speakers from England. For these two attributes and the two speaker groups, a low speech rate seems to be preferred. Enthusiasm seems to be the opposite, for male speakers and North American speakers this time, where the medium or high speech rates were preferred. The charisma analyses showed no obvious patterns, except that a medium speech rate between 5.6 and 6.6 syll/s seems to be slightly preferred for North American speakers.

The enthusiasm ratings show one pattern across the different manipulated acoustic features, either as significant effects, or trends. Irregardless of acoustic feature or manipulation thereof, the North American speakers were rated as less enthusiastic than the speakers from England. Similarly, independent of acoustic feature, the descriptive statistics suggest that male speakers generally tend to be perceived as more charismatic than female speakers.

In terms of familiarity, there were significant positive correlations between the charisma ratings and the indication of familiarity with the speakers for all acoustic features. These correlations suggest that the less familiar a speaker is to listeners, the lower they are rated in terms of charisma. There were also significant main effects or interactions of *Familiarity* for all acoustic features. They all indicate that if a speaker was not known (or at least not recognized), the participants rated them as less charismatic than if the speaker was known, familiar, or the participants were unsure. In many cases, though, there was no significant difference between the responses “I know the speaker” and “I do not know the speaker”.

8.6.2 Manipulation effects on direct charisma ratings

The general hypothesis H1 predicted for the features investigated in this chapter that a larger pitch range, higher pitch level, medium or high speech rate, and non-rising pitch contours would be perceived as more charismatic. In general, this hypothesis could only be partially supported by the data. The results indeed suggest that stimuli with a larger pitch range are perceived as more charismatic. More specifically, a narrower pitch range, i.e. more monotonous speech, was perceived as less charismatic, irrespective of speaker gender or origin.

The expected higher pitch level was only perceived as more charismatic for the speakers from England, while a medium or high speech rate only seems to be relevant for speakers from North America. This might be a first indication that speakers from the two general origins are perceived differently in the YouTube context, or that different acoustic characteristics are connected to charisma perception in the two speaker cultures.

The expected preference for non-rising contours for charisma perception only occurred for the female speakers in the form of plateaus, while rising contours were preferred for the male speakers. It might be that for female speakers, it is important to be perceived as authoritative. The Frequency Code suggests that low pitch is associated with authority (Ohala, 1984; see, e.g., Winter et al., 2021 for an overview of potential issues of the Frequency Code). The association of low pitch and authority may be achieved by using non-rising contours. For the female speakers in the sample, plateaus were preferred which is not as definitive and authoritative as a fall, and at the same time also suggests that the speaker will say more, without raising the pitch too much. A rise could—following the Frequency Code—be interpreted as submissiveness and uncertainty (Ohala, 1984; Gussenhoven, 2002, 2016). At the same time, the male speakers perhaps are perceived as more charismatic when they are putting more effort into their speech as is assumed for rising contours by the Effort Code (Gussenhoven, 2002) but also sounding more lively, while this may also be interpreted as less authoritative. Additionally, rising final contours in English are often connected to so-called “uptalk” which is mostly seen as connected to young female speakers and frequently triggers negative associations, though there can be many different interpretations of listeners (Tyler, 2015). This phenomenon is often connected to American English, especially in California, but is actually present throughout and, importantly, also produced regularly in British English varieties (Tomlinson and Fox Tree, 2011; Arvaniti and Atkins, 2016). It might therefore be perceived as less charismatic especially for female speakers of English. Since the statistical models did not include a three-way interaction between *Manipulation*, *Gender* and *Origin*, it would be interesting for future studies with larger sample sizes to see if this negative impact of rising contours for female speakers holds, and if it may be more severe for North American speakers than for speakers from England. Likewise, it would be interesting to see if these results would differ with listeners from North America.

8.6.3 Manipulation effects on attribute ratings

The discussion in the previous section mainly offered insights into the main hypothesis regarding charisma directly. Other hypotheses were put forward in this

chapter to also compare the different attributes (including charisma) and possible influences of the manipulations. These hypotheses are discussed in the following.

H_{P1}1 predicted that acoustic features produced with more vocal effort like a larger pitch range or a higher pitch level would be connected to higher charisma, enthusiasm, and persuasiveness ratings, but lower authenticity and likability ratings. This was generally not the case for pitch level (only for speakers from England and their charisma ratings) and speech rate manipulations. Interestingly, for pitch level and authenticity, the North American speakers also received fairly high ratings for the HF0 stimuli (median almost on “agree”), but the rating median for LF0 stimuli was even higher (on “agree”). However, the LF0 stimuli ratings showed an extremely wide variation, ranging almost across the rating scale from one extreme to the other with their middle 50 percent of data. That means that this stimulus polarized the raters. It also suggests strong individual differences between the speakers, though. Since the pitch level analysis only included the male speakers in the sample, only two North American speakers (and three speakers from England) were rated. It could be that the ratings have a bimodal distribution with one speaker primarily receiving high authenticity ratings, and the other speaker receiving a larger amount of low authenticity ratings.

For pitch range, though, the likely preferred acoustic feature characteristics (descriptive mostly, some also statistically significant) for charisma and persuasiveness align, with a preference for the more effortful increased pitch range. This was also the preferred feature characteristic with the authenticity ratings, though, which was not expected. In a previous preliminary study with the same speaker sample focusing on the vowel space, the acoustic feature was positively correlated for charisma (larger vowel space, higher charisma rating, as is expected from the literature, see Niebuhr and Gonzalez, 2019; Niebuhr, 2020), but negatively correlated for authenticity (Berger et al., 2023). These results were interpreted as a first indication that vocal effort (important for the perception of charisma as it shows listener-oriented speech) may not be advantageous when appearing authentic on YouTube which tends to go in hand with being approachable, relaxed, and like being involved in a conversation with a friend (see Kyncl and Peyvan, 2017). These interpretations from the vowel space do not seem to apply to pitch range. Likewise, it was expected that—like for charisma and persuasiveness—a larger pitch range would lead to higher enthusiasm ratings, mainly due to an increased amount of liveliness and expressiveness. This was not the case for the current sample where, irregardless of speaker gender or origin, the unchanged stimuli were rated as most enthusiastic. This could be an indication that the widened pitch ranges were too wide to still be perceived as enthusiastic, but rather as too enthusiastic. For example, Fischer et al. (2022) found that robots with larger pitch ranges of 18 st were perceived as significantly more enthusiastic (and charismatic), and experiment par-

ticipants felt more energetic than robots with a smaller range of 12 st. For most speakers in the sample for the current experiment, the pitch ranges of the HF0R stimuli were higher than 18 st, and those of the ORIG stimuli were around 18 st (see Table B.1 in Appendix B). This may indicate a threshold for pitch range and enthusiasm as well as charisma in the context of YouTube.

For the final contour direction, a similar alignment can be observed in that there were no patterns for *enthusiastic* and *likable* ratings, but the most likely preferred acoustic feature characteristics were similar for the *charismatic*, *authentic*, and *persuasive* ratings—in particular for male and female speakers. As for the charisma ratings, female speakers received higher ratings with non-rising contours (in this case, plateaus), and male speakers with rising contours. This also occurs with the *authentic* and *persuasive* ratings, and may again suggest that female speakers need to show more authority in their speech to be perceived as charismatic and persuasive, and male speakers need to show more listener-oriented speech, but at the same time perhaps less dominating speech for charisma and persuasiveness. It might be that male and female speakers encode charisma and persuasiveness differently which warrants further research, but is also indicated by previous research (e.g., Niebuhr et al., 2019).

In general, it was not expected that authenticity ratings would behave similarly to charisma and persuasiveness, though, especially not in the context of YouTube, so also the connection between the three attributes needs further research since it came up with two of the investigated acoustic features. While YouTubers tend to aim to be perceived as genuine and authentic (see, e.g., Kyncl and Peyvan, 2017; Raun, 2018), they also need to persuade their viewers to interact with sponsored content and advertisements. This can potentially reduce the perceived authenticity of a YouTuber if there are not “authentic” sections included in a video (Bishop, 2018). Furthermore, and also unexpectedly, the *likable* ratings did not offer many visual and no statistically significant results in the current study. This may suggest two things. Either, the rating attribute is coded by different acoustic features that are not investigated here, or the likability depends on outside factors more heavily than acoustic features. Alternatively, it could be that the attribute is too difficult, too subjective, or too abstract for participants to rate. Future studies should also look into the understanding of the attributes from each participants’ perspective. Additionally, future studies should investigate if certain speakers pull the ratings to one direction.

As a whole, the expectations put forward in Table 8.1 regarding which feature characteristics would most likely elicit more positive ratings for the attributes were not fulfilled by the current data set. There are only a few cases where the predictions seem to apply. These were the higher pitch level connected to enthusiasm, the larger pitch range for persuasiveness and charisma, and the lack of effect for speech

rate and the *authentic* ratings. There were also a few predictions that applied to the current data, but only for one or two of the speaker groups (male or female, England or North America; see below for a detailed discussion of gender and origin related effects), mainly for the final contour shape manipulations. The predictions should be revisited with a larger data set and perhaps revised manipulations.

8.6.4 Speaker gender effects

H_{P1}2 predicted that male speakers would be perceived as more charismatic and persuasive than female speakers in the sample. This is only partially supported by the data. There was no gender effect observable for the persuasiveness ratings across acoustic features (neither descriptive, nor inferential). But there was a descriptive pattern that male speakers may be perceived as more charismatic than female speakers, irrespective of acoustic feature. While this was not a statistically significant effect, it is suggesting similar results as previous research (e.g., Jokisch et al., 2018; Niebuhr et al., 2018a; Niebuhr et al., 2019; Gutnyk et al., 2019). Meanwhile, H_{P1}3 predicted that there would be no gender differences between the *authentic*, *enthusiastic*, and *likable* ratings, which also was only partially, though in the majority of acoustic features, the case. For the enthusiasm ratings and the pitch range manipulations, though, male speakers were perceived as more enthusiastic than female speakers. This was a non-significant trend and *Gender* was not involved in interactions, which suggests that this perception may be the case irregardless of acoustic manipulation. At this point, it is still unclear though why this effect is only based on the data set with the pitch range manipulations, especially since it seems like the manipulation itself is not particularly relevant.

Furthermore, and still regarding speaker gender, H_{P1}4 assumed that male speakers would receive higher ratings in terms of charisma and the charisma-adjacent attributes with stimuli with an increased pitch level. This was only the case for the enthusiasm ratings in a way, where the decreased pitch level stimuli received significantly lower ratings than the unchanged or increased stimuli. For the other attributes, there were either no patterns apparent, or the stimuli with decreased were likely preferred (based on mostly descriptive results, though). That could suggest a few different things. It could be that the increased pitch level manipulations ended up sounding still too unnatural for the rest of the voice characteristics—which in that case simply did not come out in the pilot study—to be rated as *authentic*, *likable*, *persuasive*, and *charismatic*, but perhaps this was less relevant or even an advantage for increased expressivity and enthusiasm. On the other hand it could also be that the manipulated pitch level and the original pitch range (which was in this case not manipulated) did not fit together well, as both influence each other (see Mixdorff et al., 2018). Future studies should therefore look even closer into the manip-

ulations and combined changes. *Manipulation* was a significant main effect in the statistical models for authenticity (showing a negative influence of HF0) and enthusiasm (showing a negative influence of LF0), which might suggest that these two attributes are connected to opposite acoustic feature characteristics in this sample. However, previous research also suggests that pitch level might not be as telling a feature for charismatic speech in general, and that it is more a matter of other pitch features like pitch range or the combination with different features that influences pitch level (Berger et al., 2017; Mixdorff et al., 2018; Niebuhr et al., 2018a). On the other hand, these two main effects are the only ones related to *Manipulation* and not involved in an interaction, which means that for the other acoustic features, there are rating differences depending on manipulation and speaker group (origin and gender). Note that there are no results for female speakers and pitch level because of exclusions.

This leads directly to the next hypothesis, H_{p15} , which predicted that there would otherwise not be gender-related differences for the other three acoustic features pitch range, final contour direction, and speech rate. This was also not the case. For final contour shape, this was already mentioned above, where female speakers received higher ratings for the *authentic*, *persuasive*, and *charismatic* attributes with non-rising contours, and male speakers with rising contours. Again, this could be related also to authority and effort, though further research is needed. Additionally, future studies should also adjust and play with the manipulation method for final contour direction, as it is not clear yet if the targeted contours were actually met by the manipulations. In the future, this should be tested further with additional perception experiments categorizing the manipulations. In particular, it is possible that the manipulations may have inadvertently altered or changed the shape, type, or timing of the pitch accent which should be controlled more closely and investigated further in future studies.

For the persuasiveness ratings, there was a second (descriptive) gender difference, this time for the speech rate manipulations. Male speakers tended to be perceived as more persuasive with a low speech rate, while female speakers tended to receive higher ratings with a medium speech rate. Previous research says that faster speech rate is connected to charismatic speech, but also higher persuasiveness ratings (Apple et al., 1979; Rosenberg and Hirschberg, 2009; Banzina, 2021). Furthermore, research found that female speakers have to work harder acoustically than male speakers to be perceived as just as charismatic (e.g., Novák-Tót et al., 2017). This may explain why the higher (medium) speech rate was rated as more charismatic than the low speech rate for female speakers, but the other way around for male speakers. At the same time, the speech rates that are included in the manipulations are very high to begin with. The low speech rate group more or less corresponds to the reference sample from the literature as well as the values

of the investigated speaker Steve Jobs in Niebuhr et al. (2016b) and those of the female CEO speakers in Novák-Tót et al. (2017). The medium speech rate group in the current experiment already lies outside the references in that investigation, meaning the high speech rate group is even further away. It follows that a) it is likely that the high speech rate group is simply too high for being persuasive which could be an indication for a possible threshold after which the speech is no longer perceived as charismatic, b) female speakers tend to be more persuasive if they speak faster than male speakers (and therefore may “need to deliver a *better* performance” for similar ratings to male speakers; Novák-Tót et al., 2017, p. 2251, italics in original), but c) that the low speech rate of the male speakers in the sample is still quite fast and comparable to reference values, and therefore is somewhere in the range of “normal” speech rates. At the same time these speakers are—at least descriptively—similar to Steve Jobs (see Niebuhr et al., 2020a), a speaker who is often regarded as a very charismatic speaker from the business world. Of course, the similarity in this feature does not mean they are similar in other features that might be more relevant for persuasiveness (and charisma). Future studies should revisit the manipulation method of the current study, but adjust the speech rates. This could be done in different ways: The manipulation could take place in smaller steps to get even more fine-grained insights into relevant speech rates for different speaker groups and attributes. Alternatively, or rather in combination with that, the range of the speech rate manipulations should be adjusted more in reference to “normal” speech rates, but also reach clearly slower and faster rates than average. The current study took the natural speech rates of the stimuli as the basis, and therefore did not control for “average-ness” of the original stimuli and then going below and above that average speech rate.

Finally regarding this hypothesis, there was also a gender difference between pitch range manipulations for the *charismatic* ratings. The decreased pitch range stimuli were rated as less charismatic than the unchanged stimuli (and there was also a non-significant trend in the pairwise comparison between LF0R and HF0R). This was only the case for the female speakers, though. For the male speakers, there were no significant rating differences between the manipulations. That suggests that perhaps for the female speakers a more monotonous voice is less forgiving, and this could be related also to the overall pitch level of the speakers: since the pitch level is higher for the female speakers, monotony could be more striking than for male speakers with lower pitch level. However, general pitch range differences between male and female speakers are known. For example, female speakers have larger pitch ranges than male speakers, but this difference is larger in a story-telling task (Daly and Warren, 2001; see also Clopper and Smiljanic, 2011). Since YouTube vlogs can be considered a type of story-telling, this may explain why female speakers may be perceived as more charismatic with larger pitch ranges and

the manipulation was not relevant for male speakers. It could also be that the raters were more sensitive to pitch range difference in female speakers.

8.6.5 Speaker origin effects

Continuing on from the hypotheses regarding speaker gender and now focusing on speaker origin, hypothesis H_{P16} predicted no rating attribute differences between the two speaker origins (England and North America). This was not the case. For the speech rate manipulations, speakers from England were generally and independent of *Manipulation*, rated as more authentic than speakers from North America. At this point, it is still unclear, though, why this effect is only present in the speech rate data set, and not throughout. This could be investigated further with a larger speaker sample.

Additionally, and perhaps more crucially, irregardless of acoustic feature, the speakers from England were rated as more enthusiastic than the speakers from North America. This is an unexpected result to some degree, since cultural science research has previously suggested that people from England or Great Britain tend to be stereotypically less expressive than speakers from North America (see Lewandowska-Tomaszczyk and Wilson, 2021), and expressiveness and enthusiasm seem to be closely related. It might be that these results are skewed by the fact that only listeners originally from the British Isles rated the stimuli³. It could be that they perceive a more similar language variety to their own as more enthusiastic, but at the same time, and keeping in mind the stereotype mentioned just above, it could also be that for British ears, the North American speakers make use of acoustic features (perhaps also outside the scope of the current experiment) that overshoot what is perceived as enthusiastic in England or Great Britain. It could be that the North American speakers simply do too much and are perceived as too enthusiastic and trying too hard. Future studies could add questions about the degree of the rating—do participants disagree with a statement because it is not enthusiastic enough or too enthusiastic? Additionally, future studies should add listeners from North America to see if those raters perhaps rate the North American speakers more highly.

The second hypothesis related to speaker origin, H_{P17}, predicted that there were no origin differences depending on the acoustic manipulations. This was also not the case. For the *likable* and *charismatic* ratings, the pitch level manipulations revealed origin differences which were significant for charisma, and descriptive for likability. Additionally, they are the opposite between the two attributes. Speakers from North America seem to be rated as less likable with decreased pitch level,

³For example, previous research suggests that British English speakers may have a prejudice towards American English in general, but North American Network English still received quite high ratings, higher than some British varieties (Hiraga, 2005).

while this seems to be the preferred pitch level characteristic for the speakers from England. The opposite is the case for the charisma ratings: here, North American speakers are rated as significantly less charismatic with increased pitch level, but this again is the preferred pitch level characteristic for the speakers from England. This suggests two things: a) it seems to be that the speakers in this sample (note that this is only the male speakers and therefore only a small number per origin) are perceived differently based on their origin, but that this is likely tied to different acoustic manipulations as well; and b) that the *likable* and *charismatic* ratings may be coded differently, especially in terms of pitch level. Since there were no significant effects for the *likable* ratings on any of the acoustic features, and all results here are descriptive or even showed no pattern, future studies should consider if the attribute “likable” is even feasible to rate in this context. It could be that the attribute is either too difficult, abstract or subjective to rate properly, or that the acoustic features investigated are not directly connected to the attribute.

It is also likely that the choice of speakers and participants was too limited in number (very likely for the speakers) and too varied in terms of background, as this could also have an effect on the rating results (Weiss et al., 2021). In general, though, Weiss et al. (2021) suggest that “[while] people might not be inclined to immediately judge whether they truly like a person from a few seconds of interaction of recorded voice samples [(as was done in the current study)], listeners can express their gradual preference of the voice of a speaker, and thus his or her likability” (p. 246). For German, they found that lower pitch may be preferred for male speakers (which may also be suggested by the current data, at least for speakers from England), and higher pitch level for female speakers (which could not be tested in the current study); a stronger correlate seems to be articulation rate which did not show an effect in the current study. The variation of pitch within stimuli also seems to be an even stronger correlate of likability for German (see Weiss et al., 2021), which will be included in the analyses in Chapter 10. Weiss et al. (2021) also found that linear modeling seems to be problematic for likability, suggesting that different (non-linear) analysis methods may yield clearer or at least different results in future investigations with the data set of this project. Additionally, likability seems to be closely tied to familiarity, which has a major effect on the ratings in the current study (see below for a discussion), so combining that with likability directly in future studies would bring important insights.

Furthermore, there was an origin difference with the charisma ratings and the pitch range manipulations suggesting that for English speakers, an increased pitch range may be perceived as more charismatic (these results are only descriptive, not statistically significant) while an unchanged speech rate seemed to be perceived as more charismatic for North American speakers. It may be that the original pitch ranges for North American speakers are perceived as similarly wide as the

increased pitch ranges for speakers from England, even though there is quite a difference between the values (see Table B.1 in Appendix B). This could also again be related to the stereotype of British speakers apparently being less expressive (Lewandowska-Tomaszczyk and Wilson, 2021), and might suggest that the English speakers in the sample need to do more in terms of pitch range to be perceived as charismatic as North American speakers, but this still needs further research that cannot be accomplished with the current data.

Additionally, there is an origin difference between the final contour direction manipulations for the attribute *authentic*. North American speakers in the sample were perceived as more authentic with rising contours, while speakers from England were perceived as more authentic with falling contours. This is similar to the gender difference for the *authentic*, *persuasive*, and *charismatic* ratings that was discussed above, and this origin difference was a significant result. This could suggest that English speakers may need to be more sure of themselves and definitive when speaking to be perceived as authentic, while North American speakers may be perceived as more authentic with more flamboyancy and openness in their stimuli (which could again be related to the biological codes; Ohala, 1984; Gussenhoven, 2002). It may also be that this authenticity is applied because of uptalk, a “[rising] declarative pitch” (Tomlinson and Fox Tree, 2011, p. 58) which is a phenomenon that is often connected to American English, especially in California, but also occurs frequently in British English varieties (Tomlinson and Fox Tree, 2011; Arvaniti and Atkins, 2016). It could be that the listeners connect this phenomenon subconsciously to American English and therefore see rising intonation as more authentic for this speaker group. This cannot be investigated in further detail here, but future studies could take a closer look into this direction as well.

8.6.6 Familiarity effects

Finally, this section addresses the third research question (RQ3) regarding the relationship between charisma and familiarity. The corresponding hypothesis predicted that the charisma rating and the indicated familiarity with a speaker were positively correlated (independent of acoustic feature and manipulation; H3). This hypothesis holds true for the current data set, in line with findings from previous research (e.g., Rosenberg and Hirschberg, 2009; Lavan et al., 2016; Jokisch et al., 2018). However, this study has the opportunity to look a little bit deeper in that the participants were presented with different response options allowing them to also indicate a degree of familiarity to some extent. The response options here were “I know the speaker”, “They seem familiar”, “Unsure”, and “I do not know the speaker”. The two middle options are semantically quite close together, so there are few differences between these two responses. And while overall there were sig-

nificant correlations between the charisma ratings and the (numerical) familiarity responses (“I know the speaker” was given a higher numeric value for the analyses, 4, than “I do not know the speaker”, 1), the statistical models and pairwise comparisons also showed that mostly, there were significant rating differences between familiar speakers and unknown speakers, and not significant differences between known and unknown speakers. This was mostly the case for the speakers from England, but also occurred for North American speakers. This suggests that knowing a speaker can be equally detrimental to that speaker’s perceived charisma, as the information if the speaker is actually liked by the listener is not available in the current study. It is, however, reasonable to assume that speakers who are known (and remembered) mainly because they are, for example, annoying to the listener are then also rated lower in terms of charisma. This possible influence on the rating could be addressed in future studies by also asking the participants if they like a particular speaker or a speaker’s voice. Additionally, the question in the experiment did not include the task for the participants to actually name the speakers. Therefore, it cannot be confirmed that they correctly recognized the speakers. This could be included in future replications.

Future studies should also include individual investigations of charisma and familiarity with each speaker to get more fine-grained insights into the connection between charisma and familiarity. Another interesting approach might be to extrapolate the familiarity rating for each speaker and participant from the second experiment and relate them to the ratings of the charisma-adjacent attributes. That way the influence of familiarity on the other ratings could also be assessed. Finally, future studies could also examine if and how the different manipulations impact the familiarity ratings, and if a speaker can become unrecognizable or at least less frequently recognized. Furthermore, future studies could focus more on a balanced amount of speakers with many subscribers and a huge reach also in the traditional media, and speakers with very small channels, to see if the familiarity impact is actually tangible. This study only used successful YouTubers, though they were not all similarly widely known.

This chapter focused on four acoustic features that are fairly frequently included in charismatic speech research. The following chapter then deals with two more features—pause duration and the presence or absence of breathing—which are also relevant features of charismatic speech, but not as frequently researched. They are also at the same time directly connected to speech on YouTube (e.g., *The Film Theorists*, 2020), and to some degree—since cuts are included—do not even appear in speeches as in politics and business which are the most researched speech genres in terms of charismatic speech.

Chapter 9

Perception 2: Pauses, breathing & cuts

9.1 Introduction

Pauses are silence within connected speech of one speaker, but not necessarily without breathing noise (Trouvain and Werner, 2022). The majority of pauses in conversations are one second long or shorter, though they can be significantly longer (Goldman-Eisler, 1961; Lundholm Fors, 2015). Previous research has also shown that, in connected speech, pauses need to have an average duration of about 120 ms in order to be perceived as a pause (Heldner, 2011; Lundholm Fors, 2015), which is the average duration of the closure time of a voiceless stop in English (Lundholm Fors, 2015).

Pauses have several functions in speech. For example, silent pauses can add emphasis to an utterance. They can also signal non-assertiveness (Benus et al., 2006). Additionally, there is a higher frequency of pauses and a tendency for shorter pauses in truths than in lies (Benus et al., 2006). Different emotional states can also be expressed by pauses and the durations and frequency thereof. For example, more excitement is connected to a higher speech tempo—at least as far as perception is concerned (Trouvain and Werner, 2022). Trouvain & Barry (2000) found that—for example—horse race commentaries become faster the closer to the end the race is, and therefore the higher the needed degree of expressivity and excitement. They found out that this perceived increase in tempo was not caused by a faster articulation rate which stayed relatively consistent, but rather because of more and shorter pauses with more frequent audibly strong breathing noises towards the end of a race (Trouvain and Barry, 2000; see also Trouvain and Werner, 2022).

In terms of pauses and charisma, shorter pauses tend to sound more charismatic than longer pauses (D'Errico et al., 2013; Niebuhr et al., 2016a; Berger et al., 2020; Niebuhr et al., 2020a). On the other hand, the degree of humility and dominance seems to be also connected to the use of pauses: politicians who are perceived as humble (e.g., Barack Obama) tend to use more and longer pauses, while politicians who are perceived as more dominant (e.g., Donald Trump) tend to use longer

phrases with shorter pauses, perhaps to hold the floor for longer (D'Errico et al., 2019).

On YouTube, there is an interaction between pause duration and perceived awkwardness. In a video, Matthew Patrick—as the host of the YouTube channel The Film Theorists—looks at YouTube videos of American Late Night Shows broadcast during lockdown in 2020, and why this broadcasting format does not seem to be successful for most late night show hosts when they keep their presentation style with long pauses, waiting for laughter from an audience that is not in the same room. One of the shows he looked at—*The Daily Show with Trevor Noah*—actually incorporated YouTube methods like “editing tricks to inject energy—we’re talking jumpcuts pushing into tighter angles” (The Film Theorists, 2020, min. 10:35f.) which usually goes in hand with an audible jump as well. The following quote illustrates what is behind the frequent use of jumpcuts (also known as “smash cuts”) and the short pauses on YouTube:

Everyone is always eager to make fun of how frenetic YouTubers are with their smash cuts every few seconds. There’s a reason for that: The internet runs at a mile a minute, and pauses in this universe are used to show that something is awkward, not that something is funny. You let it sit and hang because it’s uncomfortable. (The Film Theorists, 2020, min. 7:03f., punctuation added)

This suggests that the perception of YouTube relies on short pauses, cuts, make it snappy to keep viewers engaged. This seems to be especially important for the first section of a video, as two other YouTubers mention: “The first 20 seconds or so are the most important, so be energetic and to the point so you draw your audience in” (Howell and Lester, 2015, p. 135).

This chapter (like the previous chapter) addresses the first and third research questions of the dissertation, which were introduced in Chapter 5, and which are repeated here again as RQ1 and RQ3:

RQ1: *How should acoustic parameters be configured to be perceived as charismatic (both in terms of charisma directly and charisma-adjacent attributes) in the context of YouTube vlogs?*

RQ3: *Is there a connection between charisma ratings and familiarity with the speakers in the sample?*

In this chapter, the focus lies on the second part of the experiment where the influence of pause-related manipulations is investigated. More specifically, this means that the influence of pause duration, and presence or absence of breathing noise on the perception of charisma-adjacent attributes—*authentic, enthusiastic, likable, and persuasive*—and charisma directly as well as the combination of charisma and familiarity are addressed.

While Chapter 5 introduced a general hypothesis for this and the prior chapter, Section 9.2 further introduces specific hypotheses and expectations for the current part of the experiment. Section 9.3 explains the methods behind the experiment, before the results are presented (Sections 9.4 and 9.5), and discussed (Section 9.6). For information on the speakers and videos included in the sample, refer back to Chapter 6 for an overview of data selection and its criteria.

9.2 Hypotheses

The main hypothesis of this part of the investigation is repeated again below as H1. It is the second half of the hypothesis first mentioned in Chapter 5, and refers directly to the influence of pause duration and breathing noise on the ratings of charisma, and the adjacent attributes *authentic*, *enthusiastic*, *likable*, and *persuasive*.

H1: *Stimuli with [...] audible breathing and shorter pauses [...] are perceived as more charismatic.*

This general hypothesis is broken down into several more specific hypotheses that are tested in this part of the investigation. It is generally predicted that for most attributes, short pauses receive higher ratings than long(er) pauses (H_{P21}; the p₂ refers to the connection of these hypotheses to the second perception experiment that is part of this chapter). Cuts instead of pauses, however, are expected to be perceived as more enthusiastic than the other pause durations, while cuts are expected to receive the lowest ratings on all other attributes (H_{P22}). This is assumed because cuts make a stretch of speech faster and perhaps livelier than it was before, but at the same time usually also sound artificial. This might not matter as much for enthusiasm, where the speed is likely priority, but could sound too unnatural for positive ratings on the other attributes.

H_{P21}: *Stimuli with short pauses are perceived as more charismatic, enthusiastic, authentic, and likable than medium pauses, long pauses, or cuts.*

H_{P22}: *Stimuli with cuts instead of pauses are perceived as more enthusiastic than the other pause durations, but get the lowest ratings for the other attributes.*

Stimuli with long pauses are predicted to be perceived as more persuasive than the other pause durations, but are predicted to get lower ratings for the other attributes (H_{P23}). This is predicted because longer pausing might a) be used for emphasis, or b) suggest that the speaker is thinking about what they want to say next, both of which may project competence and increase persuasiveness.

HP₂₃: *Stimuli with long pauses are perceived as more persuasive than medium or short pauses, or cuts, while these stimuli are expected to get lower ratings for the other attributes.*

Of all the different pause duration manipulations, it is expected, though, that the mix of pause durations (one long, one medium, and one short pause in each stimulus) is given the highest ratings by the participants, and this is expected for all attributes (HP₂₄). This condition shows variety, and variety in turn signals liveliness which is an integral part of charisma and persuasiveness, but also enthusiasm, and likely also authenticity and likability. Table 9.1 provides an overview of the expected ratings for each of the rated attributes, ordered from the expected highest-rated manipulation on the left to the expected lowest-rated manipulation on the right of the table.

HP₂₄: *A mix of pause durations gets the highest ratings for all attributes because of its variability.*

Furthermore, it is expected that stimuli that include an audible breathing noise in the pauses are rated higher on all attributes than stimuli with pauses without breathing noise (HP₂₅). However, it is expected that stimuli with a mix of pauses with and without breathing noise are given the highest ratings for all attributes (HP₂₆). This is predicted because it may be too unnatural to hear breathing in every pause, so the variation in the mix condition may be perceived more positively. Likewise, having variety in breathing noises also draws the attention to certain parts of a stretch of speech which may also increase the naturalness.

HP₂₅: *Stimuli with audible breathing noises are expected to be more highly rated on all attributes.*

HP₂₆: *Stimuli with a mix of breathing and non-breathing pauses (i.e., pauses with and without breathing noise) get the highest ratings for all attributes.*

Additionally, there are no effects of speaker gender (HP₂₇) or speaker origin (HP₂₈) expected for this part of the investigation. This is the case both for breathing presence/absence, as well as the pause duration. While previous research has shown that male speakers had shorter pause durations than a reference population from studies without ties to charisma or emotional prosody (D'Errico et al., 2013; Niebuhr et al., 2020a), female speakers have not been investigated so far regarding these features and charismatic speech. Likewise, speakers from England have not been investigated, so while the male speakers in Niebuhr et al. (2020a) were both from America, there is nothing known about speakers from England that would provide reason for expecting differences between the speaker groups. Therefore, as an exploration and a collection of data for female speakers and speakers from

Table 9.1: An overview of the expected ratings of the different attributes and the effect of the pause duration manipulations on the ratings.

| Attribute | Expected ratings of pause durations (highest to lowest) | | | | | | | | |
|---------------------|---|---|-------|---|--------|---|--------|---|------|
| Authentic | mix | > | short | > | medium | > | long | > | cut |
| Enthusiastic | mix | > | cut | > | short | > | medium | > | long |
| Likable | mix | > | short | > | medium | > | long | > | cut |
| Persuasive | mix | > | long | > | medium | > | short | > | cut |
| Charismatic | mix | > | short | > | medium | > | long | > | cut |

England, these two aspects are also investigated, especially because differences are likely, but the direction cannot be predicted based on previous research of charismatic speech.

H_{P27}: Rating differences between male and female speakers are likely, but directions cannot be predicted.

H_{P28}: Rating differences between speakers from England and North America are likely, but directions cannot be predicted.

Finally, this chapter also addresses the third research question of the dissertation regarding how charisma and familiarity are connected (see also Chapter 5, and the corresponding general hypothesis H3 (repeated below). Similarly to the expectations of the investigation in the previous chapter, it is also assumed that the charisma ratings and the claimed degree of familiarity with the speakers by the experiment participants correlate positively and significantly. That means that is expected—as suggested also by previous research (e.g., Rosenberg and Hirschberg, 2009; Lavan et al., 2016; Jokisch et al., 2018)—that known speakers are perceived as more charismatic, and equally that the more familiar listeners are with the speakers, the more charismatic they are perceived, but while this is expected for both pause duration and breathing noise manipulations, no difference between the specific manipulations is predicted.

H3: The more familiar a speaker is to the listeners, the more charismatic they are perceived.

9.3 Method

9.3.1 Stimulus creation: Pause manipulations

For this part of the investigation, long stimuli consisting of four naturally connected phrases (either bordered by pauses, or by a pause and a minor phrase boundary) and three pauses each (6 to 15 s) were chosen. Here, the pauses were changed with cutting and pasting silent sections to receive different pause durations (medium,

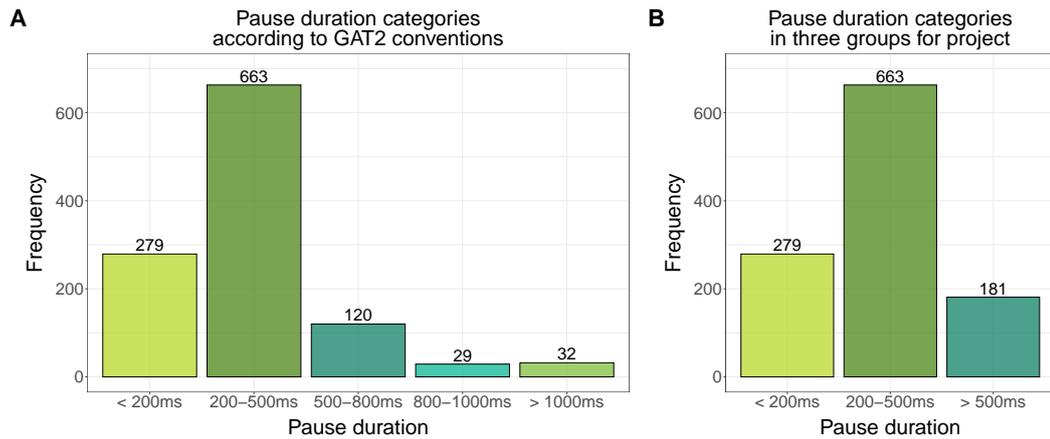


Figure 9.1: The frequency of different pause duration categories in the data. The pause durations were grouped A) following GAT2 conventions (Selting et al., 2009), and B) in three groups. The grouping in B) is used for the present investigation. Pauses below 200 ms were used as short pauses, between 200 and 500 ms as medium pauses, and over 500 ms as long pauses in the stimulus creation.

long, short, or an absence of pausing, i.e. cuts), presence or absence of audible breathing, as well as mixed conditions. The pauses had durations between 200 and 500 ms, as this was the duration of the majority of the pauses in the data.

To categorize the pauses, the pause durations in the full data set were measured and categorized, following GAT2 conventions (see Selting et al., 2009), in five categories—shorter than 200 ms, 200–500 ms, 500–800 ms, 800–1000 ms, and longer than 1000 ms—as well as an additional category of pauses longer than 1 second (see Figure 9.1A). Pauses longer than 500ms were in the minority, meaning the pauses in the data set were sorted into three groups that are used in the remainder of the study (see Figure 9.1B): short pauses (< 200 ms), medium pauses (200–500 ms), and long pauses (> 500 ms).

All pauses had naturally occurring inbreaths in each of the pauses with the exception of speaker ZS, where it was impossible to find a stimulus that met all criteria. Her “original/unchanged” stimulus contained one pause without a breath and two breathing pauses. It was categorized into another stimulus group with mixed breathing, and included with the others as a manipulation in the category with three medium breathing pauses. Additionally, there were no cuts in the original stimuli.

The length of the stimuli differs considerably between the speakers, as the choice of material was limited because of the many requirements. Stimulus length was considered a lower priority than other criteria. However, most stimuli fall into a duration range between 10 and 18 seconds, just two of the speakers are represented by less than 10 seconds of speech material.

Additionally, an (ideally) long segment of pause without breathing was extracted

from all speakers. This segment was chosen from another pause from the annotated sample of speech. This long pause segment was later used to create different pause durations in the manipulation which ensured consistent room acoustics and equipment within one speaker.

The different stimuli are listed below in Example 8. The positions of the pauses are indicated by vertical bars. One vertical bar indicates the weakest syntactic boundary in the stimulus, two vertical bars an intermediate syntactic boundary, and three vertical bars the strongest syntactic boundary. Syntactic boundary strength was determined relative to the other boundaries in each stimulus by comparing the constituents that were delimited by pauses in the original stimulus and checking which of the pauses ended a full syntactic sentence or a full syntactic phrase. For ease of reading, punctuation is not included. The starting time within the video is included below.

- (8) a. *As he stopped doing that as much he || assumed and kind of expected that a big chunk of his audience was gonna disappear because that's why they were subscribed ||| But he was happy for that chunk of his audience to disappear | to be able to just make content that he was enjoying and have a smaller audience (AD; Alfie Deyes Vlogs, 2018, min. 13:31)*
- b. *A lot of people think Netflix just came to me and said: "Hey, have a TV show" that's not how it worked ||| I've been writing the show for years with my brother | and then I got together with these other writers and we started developing the show even further || and then the four of us went around to every single network (CB; Colleen Ballinger, 2017, min. 3:10)*
- c. *And I really have to force myself | to get back on and that means taking basic self care || very seriously ||| which sounds really stupid and obvious (DH; Daniel Howell, 2017, min. 5:41)*
- d. *Most of you know I am a huge Mickey Mouse fan so | I was wearing my Mickey Mouse t-shirt and I was feeling pretty good It was like this like vivid lilac color and have bright pink lipstick and my hair was like: "Woohoo!" || I felt great I was really happy in it ||| And I was talking to the other moms (LP; Louise Pentland, 2017, min. 0:36)*
- e. *How YouTube has changed my life ||| When I started making videos I was depressed I felt lonely I did- had nowhere to turn to and || I found a family a community | and a purpose on YouTube it has given me a career that is unimaginable (LS; Lilly Singh Vlogs, 2017, min. 2:28)*
- f. *Everything that I am everything that I've ever done every single emotion and nuance of the person | that I not only want to be but have become || is laid out bare in the videos that I have made ||| that is me (MF; Markiplier, 2018, min. 2:10)*
- g. *And trying to like | give you guys at home as much information as we currently have ||| obviously we haven't read the whole thing and like what has officially gotten passed at this point but || according to the FAQ on the website (MP; GTLive, 2019, min. 26:31)*

- h. *Then I thought | | | if I'm a ghost | | am I gonna haunt this room | and is my ghost gonna have emoji pyjamas (PL; AmazingPhil, 2018, min. 3:59)*
- i. *Who might upload a video we don't know what might be in it we don't know if there's gonna be copyrighted material in it | | | so we are just going to ban all individual users like you said | from uploading but the people who wouldn't necessarily need to fall under that umbrella are | | the big companies (SP; GTLive, 2019, min. 22:39)*
- j. *Alongside videos alongside vlogs alongside editing alongside other massive projects that do take up | a lot of my time | | which I also enjoy doing so | | | it's not like I will ever stop those things to carry on those things (ZS; Zoe Sugg, 2018, min. 3:21)*

All stimuli were created in Praat by using the cutting, copying and pasting functions in the sound object. Sections that were selected could be cut out, and replaced at the same point in the signal by additional silences to create different lengths of pauses (long, medium, short), pauses with and without breathing (for both long and medium pauses), and cuts that auditorily mirror jumpcuts (cuts that create a distinctive acoustic jump between two phrases or words and thereby eliminating a pause or unwanted words) that are frequently used in YouTube videos (see *The Film Theorists*, 2020, for further information) and therefore represent practices from the genre the speech material is from, and that cannot be found as a factor for the charisma perception in other genres like political speeches or keynote presentations in business.

Including the original and unchanged starting stimulus that was chosen, eight different stimulus types were created for each speaker, amounting to 80 stimuli in total. All stimuli were re-synthesized by creating a manipulation object in Praat (see Section 8.3.1 for more information on Praat's manipulation functions) and then resynthesized using the "overlap-add" function. The resynthesis was performed in order to ensure that there was no substantial difference in audio quality between the first and the current second part of the experiment since all experimental parts were presented to the same participants in one session.

To ensure similar amplitude values between the different speakers on the one hand and the different stimuli from each speaker on the other hand, the normalization from the first data preparation (see Section 7.1) was repeated for all stimuli: They were all again normalized to -3dB in Audacity, because the re-synthesis process in Praat can affect the amplitude of the signal (E. Chodroff, p.c.).

Table 9.2 introduces the pause-related stimuli manipulations with their abbreviations, as well as a short explanation of the manipulation.

Table 9.2: An overview of the different pause-related manipulations of stimuli used in the perception experiments in the present study. Abbreviations of the manipulations and their explanations are provided.

| Abbreviation | Explanation |
|--------------|--|
| LONG_NBR | long pauses (> 500 ms), no breathing |
| LONG_BR | long pauses (> 500 ms), breathing |
| MED_NBR | medium pauses (200–500 ms), no breathing |
| MED_BR | medium pauses (200–500 ms), breathing |
| SHORT | short pauses (< 200 ms), no breathing |
| CUT | cut instead of pause |
| MIX_BR | one non-breathing, two breathing medium pauses |
| MIX_L | one pause of each length, no breathing |

9.3.2 Stimulus measurements

The pause durations always fell within the categories required for the stimuli. Pause durations were measured using ProsodyPro (Xu, 2013; version 5.7.8.1) in Praat (Boersma and Weenink, 2018, version 6.0.37). The breathing noises had a duration between 0.1 and 0.4 seconds and were measured with a script written by the author. Table H.1 in Appendix H provides a detailed overview of the phrase, pause, and breathing duration measurements. Since the rest of the stimuli is kept constant, rating differences are seen as caused by the manipulations. However, it is likely that other acoustic features also have an effect, which will be explored in Chapter 10.

9.3.3 Experiment design and participants

The experiment design of this experiment is the same as the experiment presented in the previous chapter (see Section 8.3.4 for the detailed experiment design and procedure). The two experiments only differ in terms of the stimuli. The participants were also the same. Those that rated the long stimuli (investigated in this chapter) in the first part of the experiment session in terms of charisma-adjacent attributes, rated the short stimuli (from the previous chapter) in the second experiment part in terms of charisma directly as well as familiarity. This was reversed for the other half of the participants. For the demographic information of the participants, see Table 8.4 in Section 8.3.5 of the previous chapter.

The participants rating the charisma-adjacent attributes *authentic*, *enthusiastic*, *likable*, and *persuasive* were split into four lists so that all 80 stimuli were rated by all listeners, but not on all attributes (see Table 9.3 for the lists). Only the unchanged stimuli were rated on all scales. There was no separation into different lists for the direct charisma ratings—all participants rated all 80 stimuli.

Table 9.3: Experimental lists for Part 1 of the experiment for the pause manipulations. The attributes that were rated for each stimulus type in each list are included.

| Stimulus | List 1 | List 2 | List 3 | List 4 |
|--------------------|--------------|--------------|--------------|--------------|
| MED_BR LONG_NBR | enthusiastic | persuasive | authentic | likable |
| MIX_L CUT | authentic | likable | persuasive | enthusiastic |
| MIX_BR SHORT | persuasive | enthusiastic | likable | authentic |
| MED_NBR LONG_BR | likable | authentic | enthusiastic | persuasive |

9.3.4 Statistical analyses

This experiment focuses on two acoustic features: pause duration and the presence or absence of audible breathing noises. In order to do that, the stimuli were grouped and (at least in part) included in the statistical analyses of both features as factor levels of the independent variable *Manipulation*. For the pause duration analyses, MED_BR and MED_NBR were collapsed into medium (MED), and the same was done for LONG_BR and LONG_NBR which were collapsed into LONG. This was done to have specific duration groups, but leave out the influence of breathing for the time being. The other groups in the pause duration analyses were SHORT, CUT, and MIX_L as a mixed feature. For the breathing analyses, the SHORT, CUT and MIX_L stimuli were excluded from the data set as they did not include breathing noises. The MIX_BR stimuli were part of this analysis as they showed a more varied use of breathing noises. The MED_BR and MED_NBR stimuli as well as the LONG_BR and LONG_NBR were included as comparisons between pauses with and without breathing, but also to see if there was a difference in breathing noise influence between different pause durations.

Before the statistical analyses, some of the stimuli ratings in this experiment had to be excluded. For each experiment part (the charisma-adjacent attributes and the direct charisma ratings), two ratings were excluded as the participants indicated they had not heard the audio. Additionally, 60 out of 3,200 ratings had to be excluded because the reaction time that is recorded by Praat was shorter than the duration of the stimulus. It was deemed important that the full stimulus was listened to before a judgment was made. This means that 3,136 ratings were analyzed in this part of the study.

The statistical methods overall did not differ from the methods used in the previous experiment chapter. The independent variables were *Manipulation* (five levels per acoustic feature, see above), *Gender* (levels: male and female), *Origin* (levels: ENG and NAM), and *Familiarity* (in connection with the direct charisma ratings; four levels: "I know the speaker", "They seem familiar", "Unsure", and "I do not

know the speaker”). For further information on the procedure as well as the general model structures, see Section 8.3.6 in the previous chapter.

The LMMs are knowledge-based models and were kept constant across experiments to allow model comparisons. This method was chosen despite three of the models in this analysis not converging. These three models—the *enthusiastic* and *persuasive* ratings for the pause duration analyses, and the *likable* ratings for the breathing analyses—cannot be interpreted. Three models not converging means that the majority of models (21 out of the 24 models related to the charisma-adjacent attributes) are still converging and interpretable. Models that returned a singular fit were interpreted normally (see Barr et al., 2013). For models that did not converge or that had a singular fit, different optimizers were tested to see if this would solve the convergence problems (see Clark, 2020). This information along with the actual model structures can be found in Appendix I.

For the investigation of the direct charisma ratings, the rating responses were also correlated with the numerical ratings of the familiarity with a speaker (from “I know the speaker” = 4 to “I do not know the speaker” = 1). The correlations were calculated using Pearson correlations (r). Pearson’s r was chosen because while the data were not normally distributed, the pause duration data set had around 1,300 data points and the breathing noise data set about 980 data points, and were therefore considered a large sample size which allows for using this method (see Levshina, 2015). The calculations were run with the `cor.test()` function.

The significance level for testing is set at $\alpha = .05$ also for the current study. Non-significant trends— p -values that can be rounded to 0.1 (i.e., $p > .05$ but $< .15$)—are also reported.

9.4 Results I: Charisma-adjacent attribute ratings

For the results of the charisma-adjacent attribute ratings, both the descriptive statistics from the figures are interpreted, as are the inferential statistics. Unlike in the study in the previous chapter, no stimuli were excluded in the current investigation. That means that there are always five male and five female speakers in the sample, as well as five speakers each from North America and England.

9.4.1 Pause duration

The LMM for the *authentic* ratings and the pause duration manipulations did not reveal any significant main effects or interactions (all p -values $\geq .2$). The LMM output can be found in Table J.1 in Appendix J. The descriptive statistics for the *authentic* ratings suggest that both male and female speakers tend to be fairly highly rated, though the rating variation is generally smaller for the female speakers. For

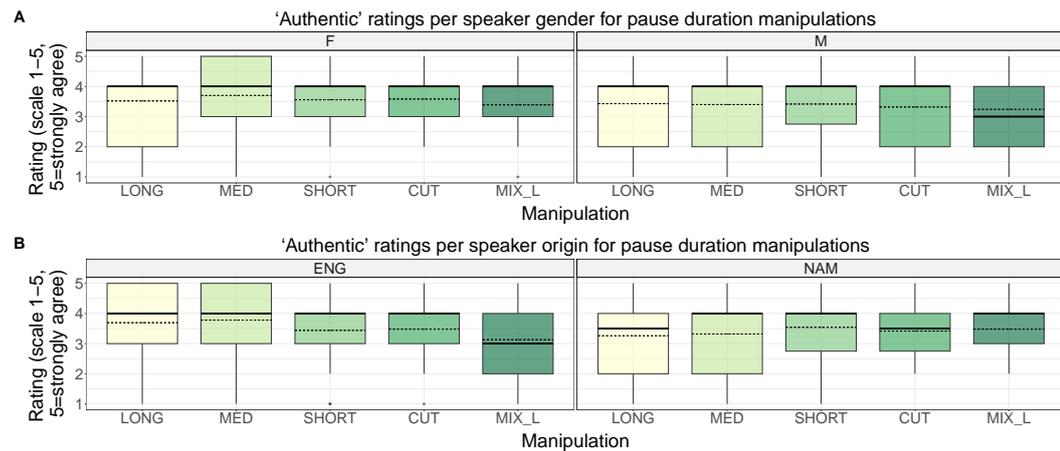


Figure 9.2: The results of the *authentic* ratings of the stimuli with pause duration manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

the female speakers, it seems like the SHORT, CUT and MIX_L stimuli are all highly rated with a high median and mean, and relatively small variation (see left panel in Figure 9.2A). The MED stimuli seem to get the most “strongly agree” responses which may suggest this stimulus is slightly preferred, but there is also variation to the negative extreme. The LONG stimuli seem to be rated as authentic, but with rating variation across the scale. For the male speakers, the variation of rating responses is much wider than for female speakers (from extreme to extreme, see right panel in Figure 9.2A). The variation is smallest for the SHORT stimuli as well, which may suggest it is less frequently dispreferred. The median for MIX_L stimuli is lower than for the other stimuli, on the neutral rating, suggesting more negative or neutral ratings for this manipulation.

For the comparison of the speakers from England and North America, all ratings are again quite high with most means on the “agree” (2) response. For speakers from England, there seem to be a number of “strongly agree” ratings for the LONG and MED stimuli which may suggest a slight preference for these manipulations, though the median rating of the SHORT stimuli is equally high with less variation to the negative extreme (see left panel in Figure 9.2B). The median response for the MIX_L stimuli is lower on the neutral answer which suggests that for speakers from England, having varied pause durations in a stimulus may be perceived as less authentic. For the North American speakers (right panel in Figure 9.2B), on the other hand, the MIX_L stimuli may be preferred as they have a high median as well as smaller rating variation than the other stimuli. LONG and MED stimuli seem to polarize more in terms of authenticity, as the ratings for these stimuli range across the board.

The LMM of the *enthusiastic* ratings did not converge, but was not adjusted to

keep the model structures consistent (see Section 9.3.4 for more information). The model is therefore not interpreted and also not reported.

Visually, male and female speakers seem to be fairly similarly rated in terms of the median enthusiasm response (see Figure 9.3A). Variation of the responses is smaller for the female speakers, though. Especially for MIX_L stimuli, there basically is no rating variation for the female speakers, suggesting that these stimuli may be preferred when it comes to enthusiasm, as most of the raters seem to agree. This is similar for the CUT stimuli, while the variations are wider for the other three stimuli. For male speakers, the CUT and MIX_L stimuli also seem to be preferred for enthusiasm, with the boxes reaching the positive extreme response, and the variation not reaching the “strongly disagree” (1) response. For both male and female speakers, the MED stimuli also seem to be perceived as similarly enthusiastic as the CUT and MIX_L stimuli.

When looking at the speakers from England and North America, the main difference between the two groups is that the ratings for the speakers from England are generally higher than those for the speakers from North America (Figure 9.3B), which would be in line with the results from Chapter 8. While the median responses are the same for both speaker groups (with the exception of the SHORT stimuli, which have a neutral and therefore lower median response than the rest of the stimuli for the North American speakers), the variation ranges up to the extreme “strongly agree” response for the speakers from England, and down to the neutral or even the “disagree” response for the speakers from North America. Otherwise, the rating variations for the different stimuli are the same between the two speaker groups, suggesting that for both speakers from England and North America, MED, CUT, or MIX_L stimuli are perceived as more enthusiastic, while LONG and SHORT pauses seem to polarize more.

While the LMM of the *likable* ratings did not reveal significant main effects or interactions, there were two non-significant trends for *Manipulation* and *Origin*, respectively. The non-significant trend of *Manipulation* suggests that in general, the SHORT stimuli with only pauses shorter than 200 ms tended to be perceived as less likable than the LONG pauses. This is also what the estimates from the LMM suggest: SHORT stimuli receive the lowest *likable* ratings (LONG = 3.81; MED = 3.71; SHORT = 3.11; CUT = 4.13; MIX_L = 4.35). The second non-significant trend of *Origin* suggests that speakers from England tend to be rated as more likable than speakers from North America, which may come back to the more polarizing responses for the North American speakers that was observed in the descriptive statistics (estimates from LMM output: ENG = 3.81; NAM = 3.26; see Table 9.4 below).

Visually, when looking at the results for the *likable* ratings, it seems like the listeners are more divided when it comes to the male speakers than with the female

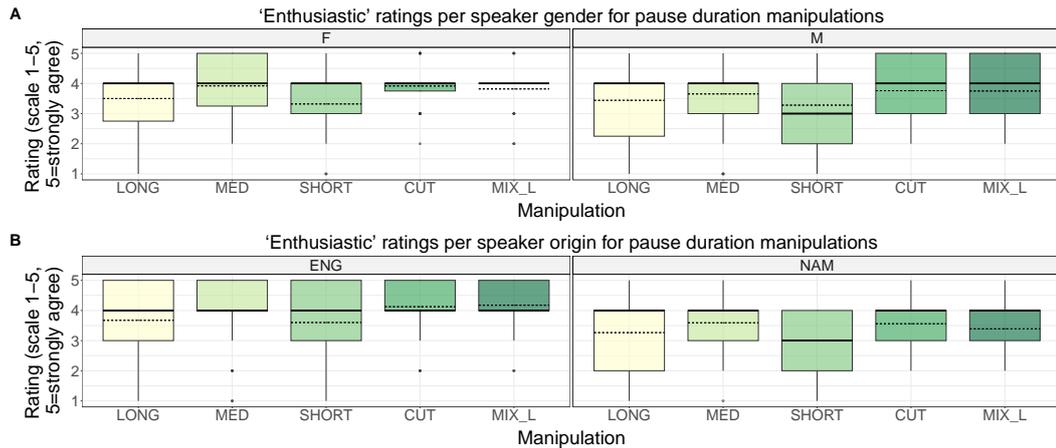


Figure 9.3: The results of the *enthusiastic* ratings of the stimuli with pause duration manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

speakers, which is seen in the larger rating variation for all stimuli with the male speakers (see Figure 9.4A). For both male and female speakers, the SHORT stimuli seem to be perceived as slightly less likable or at least neutral, since the median response here is lower. Similarly, the median response is lower for CUT stimuli with the female speakers, which may be the slightly preferred stimulus for the male speakers. For the male speakers, the median response for the MIX_L stimuli is also lower, which again seems to be one of the preferred stimuli for the female speakers. LONG and MED stimuli seem to be perceived as more likable for both male and female speakers, and MIX_L stimuli are in this group for female speakers, while CUT stimuli are also preferred for male speakers.

Speakers from North America in general seem to polarize the listeners more when it comes to likability, as the rating variation ranges across the board which is only the case for the SHORT stimuli for the speakers from England (see Figure 9.4B). For speakers from England, the CUT and MIX_L stimuli may be slightly preferred in terms of likability, while there may be a slight preference for MED stimuli for the North American speakers, as suggested by the higher median response.

The LMM for the *persuasive* ratings did not converge. It is therefore neither reported nor interpreted.

The descriptive results suggest that the stimuli with short pauses seem to be slightly preferred for persuasiveness and female speakers (left panel in Figure 9.5A). The LONG stimuli received the lowest rating median which suggests that long pauses may be perceived as less persuasive for female speakers. The other medians are on the “agree” response, but the ratings for the SHORT stimuli spread less into the “strongly disagree” response. In contrast, the MED stimuli may be perceived as most persuasive for the male speakers, which is indicated by a high

Table 9.4: The output of the LMM analysis for pause duration and the *likable* ratings. The intercept is the LONG stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG, MED, SHORT, CUT, MIX_L), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|--------------|-------|------|---------|-----------|---|
| (Intercept) | 3.81 | 0.27 | 14.20 | <.001 | * |
| MED | -0.10 | 0.26 | -0.38 | .702 | |
| SHORT | -0.70 | 0.44 | -1.60 | .117 | . |
| CUT | 0.32 | 0.43 | 0.74 | .466 | |
| MIX_L | 0.54 | 0.43 | 1.26 | .216 | |
| male | -0.21 | 0.27 | -0.78 | .442 | |
| NAM | -0.55 | 0.33 | -1.66 | .108 | . |
| MED × male | 0.11 | 0.29 | 0.39 | .698 | |
| SHORT × male | 0.22 | 0.43 | 0.52 | .608 | |
| CUT × male | -0.01 | 0.43 | -0.01 | .989 | |
| MIX_L × male | -0.47 | 0.43 | -1.09 | .284 | |
| MED × NAM | 0.06 | 0.29 | 0.22 | .828 | |
| SHORT × NAM | 0.62 | 0.54 | 1.16 | .256 | |
| CUT × NAM | -0.74 | 0.53 | -1.38 | .180 | |
| MIX_L × NAM | -0.58 | 0.53 | -1.08 | .290 | |

Signif. codes: * .05, . 1

median (shared by most other manipulations) and a smaller rating variation to the negative extreme (right panel in Figure 9.5A). The SHORT stimuli here get a slightly more neutral median, suggesting a slightly lower persuasiveness rating.

When looking at differences between speakers from England and North America (Figure 9.5B), a few observations are striking. For the speakers from England, the SHORT stimuli are perceived as least persuasive which is shown by a small variation that mostly extends below the neutral answer and does not reach the positive extreme. The other stimuli are rated similarly across the board (median on 4 = “agree”, variation ranging to both extreme responses), so no particular preference is visible. In contrast, the SHORT stimuli seem to get the highest persuasiveness ratings for the speakers from North America with small variation that does not reach below the neutral answer (except for an outlier). The other stimuli are rated across the board. The ratings for the CUT stimuli jump out, though. There is an extremely wide box suggesting the ratings vary widely and 75 percent of the data fall into a range between “strongly agree” and “disagree”. It therefore seems that for North American speakers, the CUT stimuli polarize especially, and that many listeners find them persuasive, but many do not.

9.4.2 Presence of breathing noises

The LMM of the *authentic* ratings showed a significant main effect of *Manipulation*, but there was also a significant interaction between *Manipulation* and *Gender* (see Table 9.5) which means that the main effect is not reported separately. There was a significant interaction effect for MED_NBR stimuli and male speakers. The interaction and the estimates suggest that the MED_NBR stimuli were generally rated

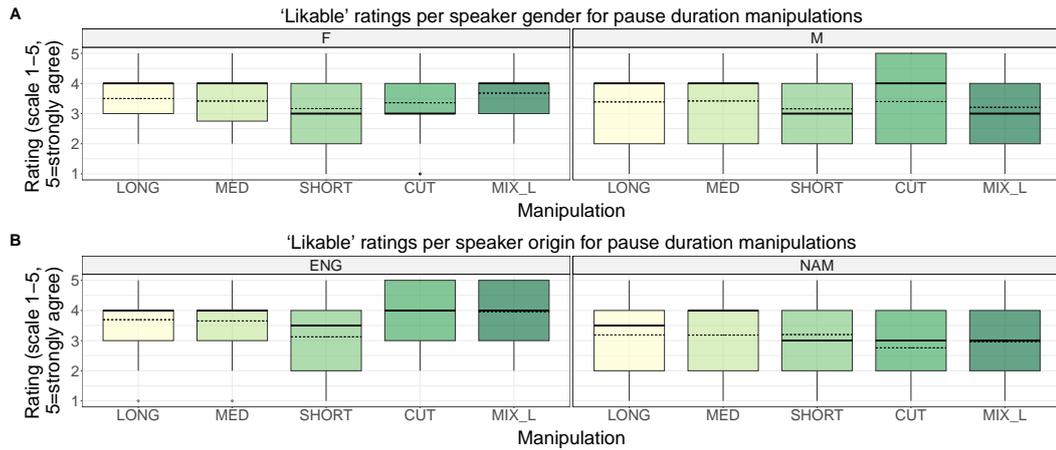


Figure 9.4: The results of the *likable* ratings of the stimuli with pause duration manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

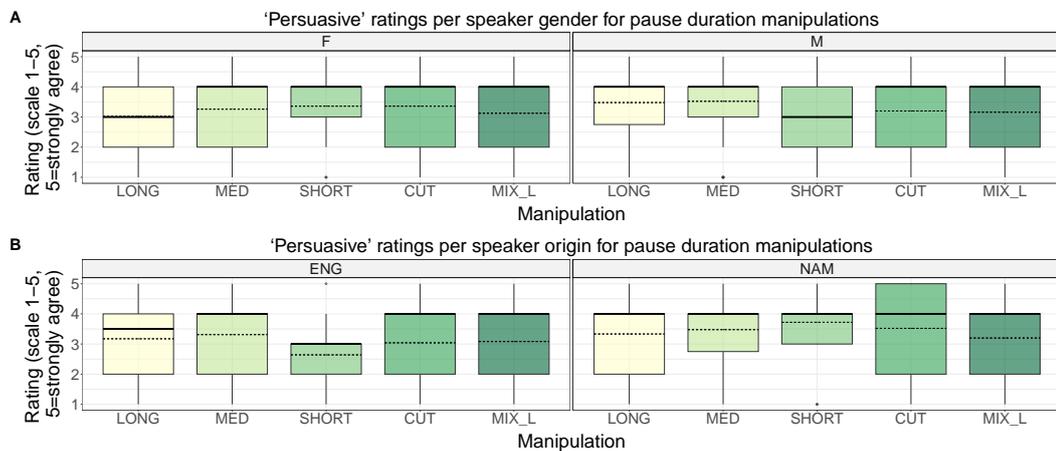


Figure 9.5: The results of the *persuasive* ratings of the stimuli with pause duration manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

as more authentic than the LONG_NBR stimuli, but that the rating difference between the two was significantly smaller for the male speakers than for the female speakers (female: LONG_NBR = 3.41, MED_NBR = 4.62; male LONG_NBR = 3.66, MED_NBR = 3.72; estimates from the LMM). There were no significant pairwise comparisons, though (all p -values $\geq .6$).

There was also a non-significant trend for the same interaction, this time for LONG_BR stimuli and male speakers. The interaction suggests that stimuli with long pauses and breathing noises were—compared to stimuli with long pauses and without breathing noises—predicted to be perceived as more authentic for female speakers, and as less authentic for male speakers. But the rating difference between the two stimuli was smaller for the male speakers than for the female speakers (fe-

Table 9.5: The output of the LMM analysis for the manipulation of breathing noises and the *authentic* ratings. The intercept is the LONG_NBR stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG_NBR, LONG_BR, MED_NBR, MED_BR, MIX_BR), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|----------------|-------|------|---------|-----------|---|
| (Intercept) | 3.41 | 0.46 | 7.39 | <.001 | * |
| LONG_BR | 0.78 | 0.58 | 1.35 | .188 | |
| MED_NBR | 1.21 | 0.58 | 2.10 | .044 | * |
| MED_BR | 0.02 | 0.39 | 0.05 | .959 | |
| MIX_BR | 0.13 | 0.58 | 0.23 | .821 | |
| male | 0.25 | 0.43 | 0.58 | .571 | |
| NAM | -0.15 | 0.54 | -0.28 | .782 | |
| LONG_BR × male | -0.87 | 0.52 | -1.69 | .108 | . |
| MED_NBR × male | -1.15 | 0.52 | -2.23 | .038 | * |
| MED_BR × male | -0.17 | 0.42 | -0.39 | .694 | |
| MIX_BR × male | -0.24 | 0.52 | -0.46 | .652 | |
| LONG_BR × NAM | -0.63 | 0.68 | -0.92 | .368 | |
| MED_NBR × NAM | -0.95 | 0.68 | -1.40 | .180 | |
| MED_BR × NAM | 0.17 | 0.42 | 0.39 | .694 | |
| MIX_BR × NAM | 0.16 | 0.68 | 0.24 | .815 | |

Signif. codes: * .05, . .1

male: LONG_NBR = 3.41, LONG_BR = 4.19; male: LONG_NBR = 3.66, LONG_BR = 3.57; estimates from the LMM).

The descriptive results for the *authentic* ratings suggest that the LONG_BR and the MED_NBR seem to be preferred for the female speakers (see left panel in Figure 9.6A), in line with the results from the LMM, though the other stimuli have the same high median rating on the “agree” response, but wider variations and fewer ratings on the “strongly agree” response. For the female speakers it therefore seems like the combination of presence or absence of breathing noises with the pause duration is important. In contrast, it appears that the MIX_BR and the LONG_NBR stimuli tend to be perceived as more authentic for the male speakers. For the MIX_L stimuli this is suggested by the box that reaches the most positive answer “strongly agree” and represents that there were a number of such responses. For the LONG_NBR stimuli, the variation (in this case, the whiskers) does not reach the most negative answer “strongly disagree” and also has no outliers there.

In terms of speaker origin, it seems like the raters were more decisive or in agreement with each other with the speakers from England than the speakers from North America (Figure 9.6B). The variations were wide for the North American speakers, and the only very slight suggestion of a preference is that the box of the MIX_BR stimuli ratings reaches to “strongly agree” answer, suggesting more positive ratings than for the other stimuli. The rating variations for the speakers from England is much smaller and only reaches the most negative answer in the case of the LONG_BR stimuli which also received the most positive responses, suggesting that this stimulus polarizes with speakers from England. The stimulus perceived

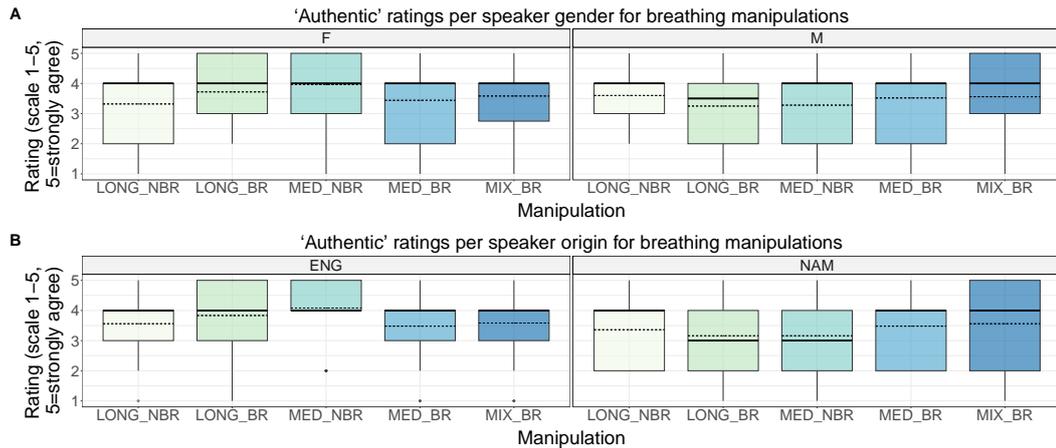


Figure 9.6: The results of the *authentic* ratings of the stimuli with breathing manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

as most authentic for English speakers was MED_NBR, with its box only ranging between “agree” and “strongly agree” without variation to lower ratings.

The LMM for the *enthusiastic* ratings only revealed a non-significant trend for the main effect of *Origin* ($p = .13$; see Table J.2 in Appendix J). The estimates suggest that the speakers from North America tended to be rated as less enthusiastic than the speakers from England (estimates from LMM: ENG = 3.58; NAM = 2.93). This is in line with the findings of the prosodic manipulations in the previous Chapter 8.

The descriptive results for the *enthusiastic* ratings suggest that the LONG_BR stimuli were perceived as most enthusiastic for the female speakers, with rating variation ranging between the “agree” and “strongly agree” responses (with the exception of a few outliers; see left panel in Figure 9.7A). This is closely followed by the MED_BR stimuli, and then the MIX_BR stimuli. The LONG_NBR stimuli seem to be rated most neutrally, though there is also variation. In general, with the exception of a few outliers for the LONG_BR stimuli, there are no “strongly disagree” ratings. On the other hand, the ratings for the LONG_NBR and MIX_BR stimuli range across the rating scale for the male speakers—the latter stimuli have the lowest median rating for the male speakers at the neutral answer (right panel in Figure 9.7A). Like with the female speakers, the preferred stimulus for the perception of enthusiasm seems to be MED_NBR.

When looking at the speaker origin, it seems like the North American speakers are overall rated as less enthusiastic than the English speakers, in line with the non-significant trend from the LMM. While all rating medians of the speakers from England are on the “agree” response, three of the five medians of the North American speakers are on the neutral response and the variation reaches further into not agreeing responses (see Figure 9.7B). But for both speaker groups, the most pre-

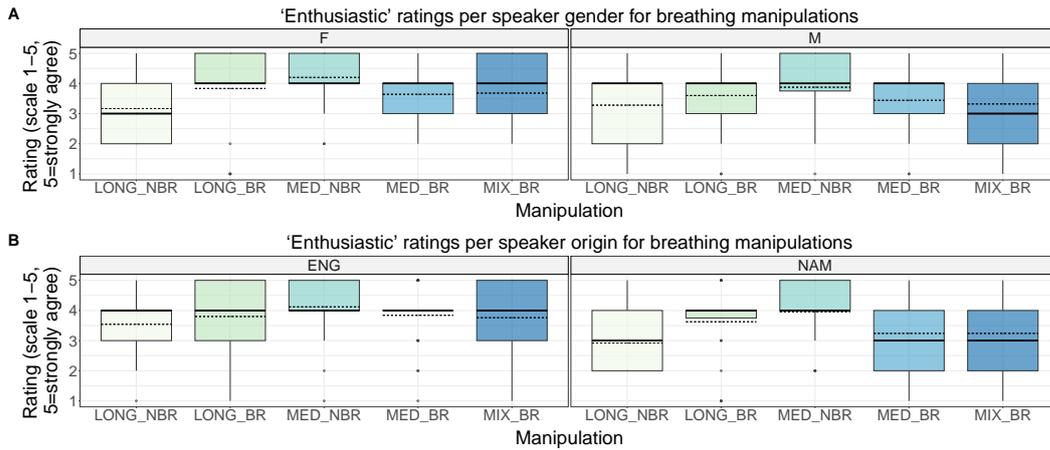


Figure 9.7: The results of the *enthusiastic* ratings of the stimuli with breathing manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

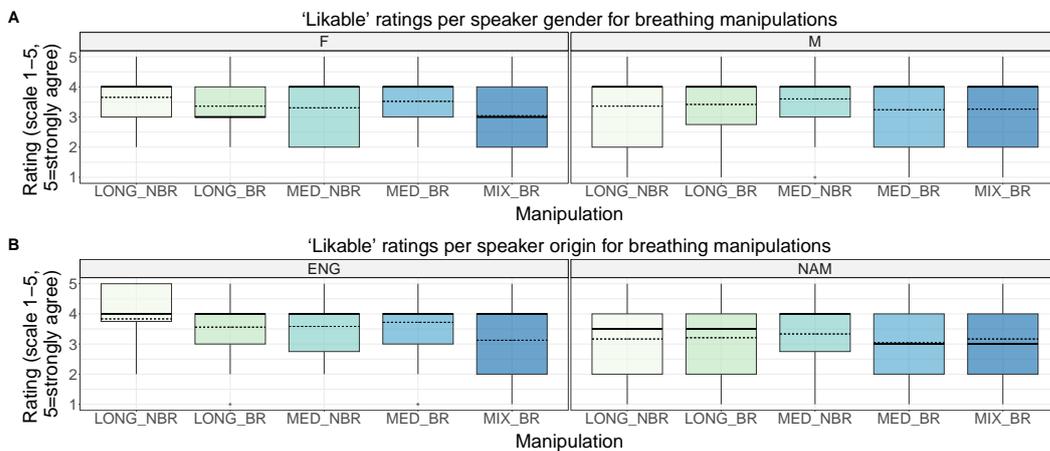


Figure 9.8: The results of the *likable* ratings of the stimuli with breathing manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

ferred stimulus seems to be MED_NBR again, as with the male and female speakers, suggesting that a lack of breathing noise in combination with pauses between 200 and 500 ms length are perceived as especially enthusiastic—at least in so far as the fairly short stimuli of the experiment are concerned. Otherwise, all other stimuli are similarly highly rated for speakers from England. The MED_NBR stimulus ratings simply include more “strongly agree” ratings than the other stimuli, though the MED_BR stimuli here are very consistently rated with the “agree” response and just a few outliers. For the speakers from North America, this is also the case, but for the LONG_BR stimuli.

The LMM for the *likable* ratings did not converge (see Section 9.3.4 for further information). It is therefore not reported or included.

The visual inspection of the results from the *likable* ratings does not show obvious

Table 9.6: The output of the LMM analysis for the manipulation of breathing noises and the *persuasive* ratings. The intercept is the LONG_NBR stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG_NBR, LONG_BR, MED_NBR, MED_BR, MIX_BR), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|----------------|-------|------|---------|-----------|---|
| (Intercept) | 3.01 | 0.49 | 6.19 | <.001 | * |
| LONG_BR | -0.23 | 0.55 | -0.43 | .673 | |
| MED_NBR | -0.12 | 0.55 | -0.22 | .830 | |
| MED_BR | 0.32 | 0.38 | 0.82 | .411 | |
| MIX_BR | 0.06 | 0.54 | 0.10 | .918 | |
| male | 0.42 | 0.56 | 0.74 | .465 | |
| NAM | 0.07 | 0.48 | 0.14 | .889 | |
| LONG_BR × male | 0.15 | 0.65 | 0.24 | .814 | |
| MED_NBR × male | 0.15 | 0.65 | 0.24 | .816 | |
| MED_BR × male | -0.29 | 0.42 | -0.68 | .495 | |
| MIX_BR × male | -0.74 | 0.64 | -1.16 | .263 | |
| LONG_BR × NAM | 0.31 | 0.50 | 0.61 | .546 | |
| MED_NBR × NAM | 0.21 | 0.51 | 0.41 | .683 | |
| MED_BR × NAM | 0.13 | 0.42 | 0.31 | .759 | |
| MIX_BR × NAM | 0.76 | 0.50 | 1.53 | .140 | . |

Signif. codes: * .05, . .1

patterns for the speaker gender (Figure 9.8A). LONG_NBR and MED_BR stimuli may be slightly preferred for female speakers, seen from the smaller rating variation and high median on the “agree” response. For the male speakers, MED_NBR stimuli seem to be perceived as more likable than the other stimuli, though there are no other obvious patterns.

In terms of the speaker origin, the results for the speakers from North America mirrors those of the male speakers in that the MED_NBR stimuli seem to be slightly preferred (smaller variation and higher median, see right panel in Figure 9.8B). Overall, the ratings of the speakers from England seem to be slightly higher than those for North American speakers, also with less variation. For English speakers, the LONG_NBR stimuli seem to be perceived as more likable than the other stimuli, which is suggested by the box reaching the “strongly agree” response representing more of such responses than the other stimuli (left panel in Figure 9.8B).

The last attribute to look at are the *persuasive* ratings. The LMM did not reveal significant main effects or interactions, though there was a non-significant trend for the interaction between *Manipulation* and *Origin* (see Table 9.6). This interaction suggests that MIX_BR stimuli tended to be rated as more persuasive than LONG_NBR stimuli, but that this difference was significantly larger for the North American speakers than for the speakers from England (ENG: LONG_NBR = 3.01, MIX_BR = 3.07; NAM: LONG_NBR = 3.08, MIX_BR = 3.90; estimates according to the LMM). This is not corroborated by significant pairwise comparisons.

Visually, the MIX_BR stimuli received a high rating with small variation for the female speakers (see left panel of Figure 9.9A). The MED_BR stimuli received a similarly high median rating on “agree”, but the variation is larger. The LONG_BR

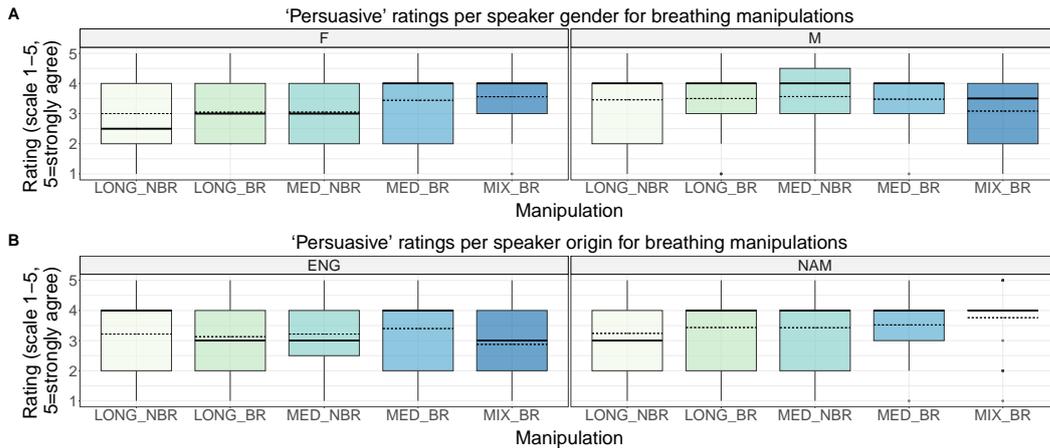


Figure 9.9: The results of the *persuasive* ratings of the stimuli with breathing manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

and MED_NBR are fairly neutrally rated with larger variation, though the variation of the responses to the LONG_BR stimuli does not extend to the “strongly disagree” response. The LONG_NBR stimuli receive the lowest median rating with wide variation. For the male speakers (right panel in Figure 9.9A), all medians are higher around the “agree” response. LONG_BR and MED_BR stimuli seem to be preferred as they have smaller rating variation. The responses for the MED_NBR stimuli are similarly high, and the box even extends slightly higher than “agree” suggesting more positive answers than the other two mentioned stimuli, but the lower whisker also extends to the low end of the scale which (in comparison to the outliers of the other two stimuli) suggests that there are more “strongly disagree” ratings and therefore polarization than with LONG_BR and MED_BR stimuli. The LONG_NBR and MIX_BR stimuli are rated across the board, though both still have high median ratings. Both male and female speakers seem to be perceived as more persuasive when there is some breathing noise audible in the stimuli.

In terms of speaker origin, there is no obvious pattern for speakers from England (left panel in Figure 9.9B). The medians of the LONG_NBR and MED_BR stimuli are higher (on “agree”; 4) than those of the other stimuli which are on the neutral response. But the rating variation ranges across the scale for all stimuli. With the exception of the LONG_NBR stimuli, all other stimuli are rated as persuasive (“agree”; 4) for the North American speakers (right panel in Figure 9.9B), which suggests that these speakers may be perceived as more persuasive overall. For these speakers, the MIX_BR stimuli seem to be preferred as there is hardly any variation around the median rating of “agree” with the exception of some outliers in both directions. Likewise, the MED_BR stimuli are also highly rated with smaller variation, though there is more variation in rating than with the mixed condition.

9.5 Results II: Direct charisma ratings and familiarity

This part of the results is focused on the direct charisma ratings and at the same time the indication of familiarity of the raters with the speakers from prior to the experiment. Both the results of the pause duration manipulations and the manipulations of breathing presence and absence are presented.

9.5.1 Pause duration

The results for the *charismatic* ratings of the pause duration-related stimuli suggest that there are very few influences of the manipulations on the charisma ratings. The LMM revealed no significant main effects of *Manipulation*. There were significant main effects of *Gender* and of *Familiarity*, though they will not be reported separately as they are also involved in an interaction with each other (see Table 9.7 for the main effects and significant interactions of the LMM; the full output can be found in Table J.3 in Appendix J). Speakers tended to be rated as less charismatic when they were familiar than when they were known, but this was only the case for male speakers and the rating difference between the two familiarity categories was only relevant for the male speakers (female: known = 3.64, familiar = 3.72; male: known = 4.15, familiar = 3.91; estimates from the LMM output; this was a non-significant trend). Similarly, speakers the participants were unsure about were also rated as less charismatic than known speakers, but again this was only the case for the male speakers and the difference between the categories was significantly larger for the male speakers as well (female: known = 3.64, unsure = 3.66; male: known = 4.15, unsure = 3.5; estimates from the LMM output). Unknown speakers were perceived as less charismatic than known speakers for both male and female speakers, but again the rating difference was significantly larger for the male speakers (female: known = 3.64, unknown = 3.06; male: known = 4.15, unknown = 2.7; estimates from the LMM output). That shows that for this sample and these experiment participants, familiarity with the speaker was much more of an advantage in terms of charisma for male speakers than it was with the female speakers.

These results are confirmed by significant pairwise comparisons: they show (see Table 9.8) that for the male speakers, the group means between the familiarity categories differed significantly with the exception of the comparison of the known and familiar speakers. There was also a non-significant difference between the ratings for speakers the raters were unsure about and those that were unknown, which suggests that uncertainty about a speaker may have lead to more charismatic ratings for male speakers than the speaker being unknown, but this was only a trend in the current sample. The only significant difference within the ratings of the fe-

Table 9.7: The output of the LMM analysis for the manipulation of pause duration and the *charismatic* ratings. The intercept is the LONG stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG, MED, SHORT, CUT, MIX_L), *Gender* (two levels: female, male), *Origin* (two levels: ENG, NAM), and *Familiarity* (four levels: I know the speaker, They seem familiar, Unsure, I do not know the speaker). Only the main effects and the significant interactions are included here. The full output can be found in Table J.3 in Appendix J.

| | Est. | SE | t value | Pr(> t) | |
|----------------------------------|-------|------|---------|-----------|---|
| (Intercept) | 3.64 | 0.30 | 12.27 | <.001 | * |
| MED | 0.10 | 0.17 | 0.59 | .557 | |
| SHORT | 0.25 | 0.21 | 1.18 | .238 | |
| CUT | 0.26 | 0.21 | 1.27 | .206 | |
| MIX_L | 0.23 | 0.21 | 1.10 | .272 | |
| male | 0.51 | 0.27 | 1.91 | .067 | . |
| NAM | 0.14 | 0.32 | 0.44 | .659 | |
| They seem familiar | 0.08 | 0.23 | 0.36 | .719 | |
| Unsure | 0.02 | 0.25 | 0.08 | .935 | |
| I do not know the speaker | -0.58 | 0.27 | -2.18 | .035 | * |
| male × They seem familiar | -0.32 | 0.21 | -1.54 | .123 | . |
| male × Unsure | -0.67 | 0.20 | -3.29 | .001 | * |
| male × I do not know the speaker | -0.87 | 0.18 | -4.77 | <.001 | * |

Signif. codes: * .05, . 1

Table 9.8: The output of the EMM analysis for familiarity and the *charismatic* ratings for the pause duration data. Only the significant pairwise comparisons are included, as well as some non-significant comparisons that are relevant for the context of the significant results. (“I do not know the speaker” = known; “They seem familiar” = familiar; “I do not know the speaker” = unknown).

| | Est. | SE | t.ratio | p.value | |
|--------------------------------------|------|------|---------|---------|---|
| male × known - male × familiar | 0.16 | 0.22 | 0.727 | .994 | |
| male × known - male × unsure | 0.96 | 0.23 | 4.22 | .009 | * |
| male × known - male × unknown | 1.41 | 0.24 | 5.89 | <.001 | * |
| male × familiar - male × unsure | 0.80 | 0.16 | 5.08 | .001 | * |
| male × familiar - male × unknown | 1.25 | 0.19 | 6.53 | <.001 | * |
| male × unsure - male × unknown | 0.45 | 0.15 | 2.902 | .110 | . |
| female × known - female × unknown | 0.55 | 0.25 | 2.178 | .406 | |
| female × familiar - female × unknown | 0.70 | 0.21 | 3.37 | .037 | * |

Signif. codes: * .05, . 1

male speakers was that familiar speakers were perceived as more charismatic than unknown speakers.

In general, there was a significant positive correlation between the charisma ratings and the (numeric version) of the familiarity ratings (1 = unknown to 4 = known). This was tested and confirmed by a Pearson correlation between the two rating scores for each stimulus across participants ($r = 0.31, t = 11.92, p < .001$).

The descriptive statistics suggest that for female speakers (see left panel in Figure 9.10A), all stimuli are rated similarly, with medians on the “agree” response (= 4), the boxes ranging between “agree” and the neutral answer, and the whiskers extending to the highest response option and the “disagree” response (= 2) with a few outliers on the “strongly disagree” option. The mean for the SHORT stimuli seems to be slightly higher, but the difference is only small. Additionally, the dot

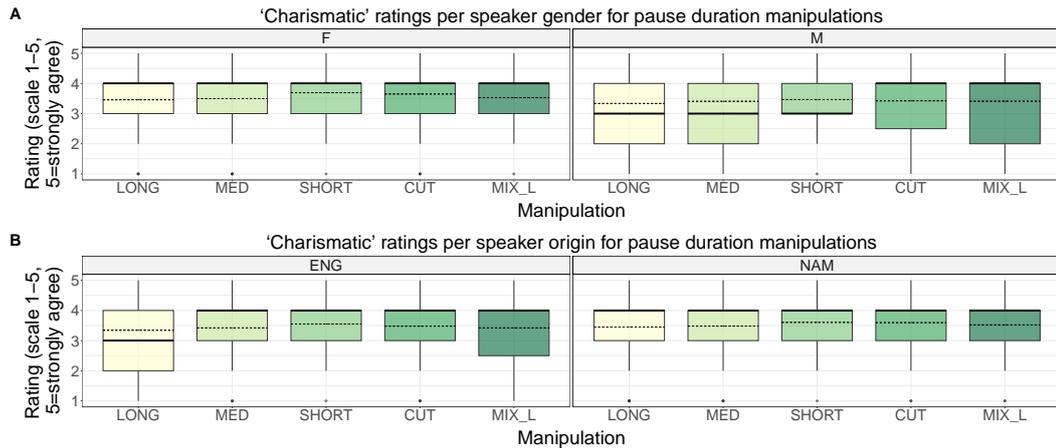


Figure 9.10: The results of the *charismatic* ratings of the stimuli with pause duration manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

signifying the outliers is lighter suggesting fewer data points and perhaps hinting at an extremely slight preference. In general, though, the ratings for the female speakers are overall higher with less variation than those of the male speakers. The rating variation spreads across the rating scale for all stimuli except the SHORT stimuli, where the variation resembles that of the female speakers, but the median is lower (right panel in Figure 9.10A). Therefore, there are no obvious preferences in terms of charisma and pause duration manipulation available for the male and female speakers, but perhaps a slight and tentative advantage of SHORT stimuli.

The rating results are similarly inconclusive visually for the speaker origin (Figure 9.10B). The boxes are exactly the same as for the female speakers with small variation and a high median for the North American speakers, as well as the MED, SHORT and CUT stimuli for the speakers from England. For those speakers, the LONG and MIX_L stimuli have wider variations perhaps suggesting that these stimuli are more polarizing for the listeners. The point signifying the outliers is again lighter for the SHORT stimuli than the other stimuli for both speakers from England and North America. This again—as with the speaker gender results—may suggest a tentative and very slight preference for these stimuli and charisma. The lack of obvious visual patterns is in line with the lack of statistical results regarding *Manipulation* from the LMM.

In terms of familiarity, the results mirror those from Chapter 8: the charisma ratings are highest and less varied for the “I know the speaker” and “They seem familiar” categories, and more neutral with much wider variation for the categories “Unsure” and “I do not know the speaker” (Figure 9.11). The LONG stimuli seem to be the most polarizing when the speaker is known or familiar. When the raters are unsure if they recognized the speakers, the SHORT stimuli may overall be per-

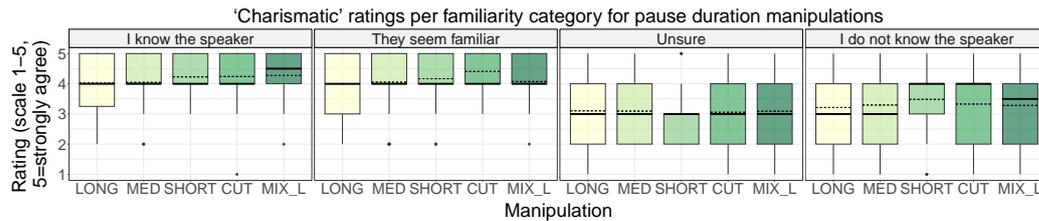


Figure 9.11: The results of the *charismatic* ratings of the stimuli with pause duration manipulations per familiarity category. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

ceived as slightly less charismatic, and but slightly more charismatic than the other stimuli (smaller rating variation) when the speaker was unknown to the listeners.

9.5.2 Presence of breathing noises

The results for the manipulation of breathing noises and the charisma ratings are similar to those with the pause duration manipulations. The LMM again did not reveal a significant main effect for *Manipulation*. There was a significant main effect of *Familiarity*, but this is also involved in a significant interaction with *Gender*, so only the interaction is reported (for the statistical output of the main effects and significant interactions of the LMM, see Table 9.9; the full output can be found in Table J.4 in Appendix J). First, there was a non-significant trend that suggests that known speakers were generally rated as more charismatic than speakers the participants were unsure about, but that this difference tended to be larger for the male speakers (female: known = 3.73, unsure = 3.48; male: known = 4.05, unsure = 3.44; estimates from the LMM output). Second, known speakers were rated as more charismatic than unknown speakers for both male and female speakers, but again this difference was significantly larger for the male speakers (female: known = 3.73, unknown = 3.01; male: known = 4.05, unknown = 2.82; estimates from the LMM output). There seem to be no significant rating differences between known and familiar speakers for either speaker gender (female: known = 3.73, familiar = 3.65; male: known = 4.05, familiar = 3.87; estimates from the LMM output).

Pairwise comparisons revealed that the means differed significantly between known and unknown (estimate = 1.14, $SE = 0.26$, t -ratio = 4.36, $p = .008$), familiar and unsure (estimate = 0.68, $SE = 0.19$, t -ratio = 3.66, $p = .03$), and familiar and unknown (estimate = 1.12, $SE = 0.21$, t -ratio = 5.33, $p < .001$) for male speakers. For the female speakers, there was only one non-significant trend in the pairwise comparison between familiar and unknown speakers (estimate = 0.71, $SE = 0.23$, t -ratio = 3.05, $p = .08$).

A Pearson correlation confirms that there is a significant positive correlation between the charisma and familiarity ratings ($r = 0.3$, $t = 9.71$, $p < .001$). However,

Table 9.9: The output of the LMM analysis for the manipulation of breathing noises and the *charismatic* ratings. The intercept is the LONG_NBR stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG_NBR, LONG_BR, MED_NBR, MED_BR, MIX_BR), *Gender* (two levels: female, male), *Origin* (two levels: ENG, NAM), and *Familiarity* (four levels: I know the speaker, They seem familiar, Unsure, I do not know the speaker). Only the main effects and the significant interactions are included here. The full output can be found in Table J.4 in Appendix J.

| | Est. | SE | t value | Pr(> t) | |
|----------------------------------|-------|------|---------|-----------|---|
| (Intercept) | 3.73 | 0.34 | 10.99 | <.001 | * |
| LONG_BR | 0.15 | 0.25 | 0.61 | .544 | |
| MED_NBR | 0.18 | 0.24 | 0.76 | .450 | |
| MED_BR | 0.16 | 0.25 | 0.65 | .516 | |
| MIX_BR | 0.31 | 0.25 | 1.24 | .217 | |
| male | 0.32 | 0.31 | 1.01 | .320 | |
| NAM | -0.02 | 0.37 | -0.07 | .948 | |
| They seem familiar | -0.08 | 0.31 | -0.27 | .787 | |
| Unsure | -0.25 | 0.33 | -0.76 | .451 | |
| I do not know the speaker | -0.72 | 0.30 | -2.36 | .024 | * |
| male × Unsure | -0.36 | 0.24 | -1.50 | .135 | . |
| male × I do not know the speaker | -0.51 | 0.21 | -2.42 | .016 | * |

Signif. codes: * .05, . 1

as with the pause duration results, the correlation seems to be more nuanced than can be depicted by a linear correlation method and should be revisited in future studies.

The visual inspection of the data does not reveal striking patterns for female speakers, as all medians are on “agree” with small variation and a few outliers on the “strongly disagree” response (left panel in Figure 9.12A). For the male speakers, there seems to be a slight preference for the MED_BR and MIX_BR stimuli, as these received higher medians and smaller rating variations, while the other three stimuli were rated more neutrally and the responses ranged to both extremes (see right panel in Figure 9.12A).

When looking at the speaker origin, there again are no clear preferences in terms of charisma. Only the LONG_NBR stimuli receive more polarizing ratings, and therefore also lower medians, for both speakers from North America and England (see Figure 9.12B). It may be that the MED_BR and MIX_BR stimuli are slightly preferred for speakers from England since they have slightly higher median ratings, but this seems to be only a tentative difference.

In terms of familiarity with the speakers, the results suggest (as with the pause duration manipulations) that known and familiar speakers received higher charisma ratings with smaller rating variations than speakers the participants were unsure about or who they did not know (see Figure 9.13). For known and familiar speakers, LONG_NBR and MED_NBR stimuli seem to be rated as slightly less charismatic than the other stimuli, but there seem to be no obvious patterns for the other two categories.

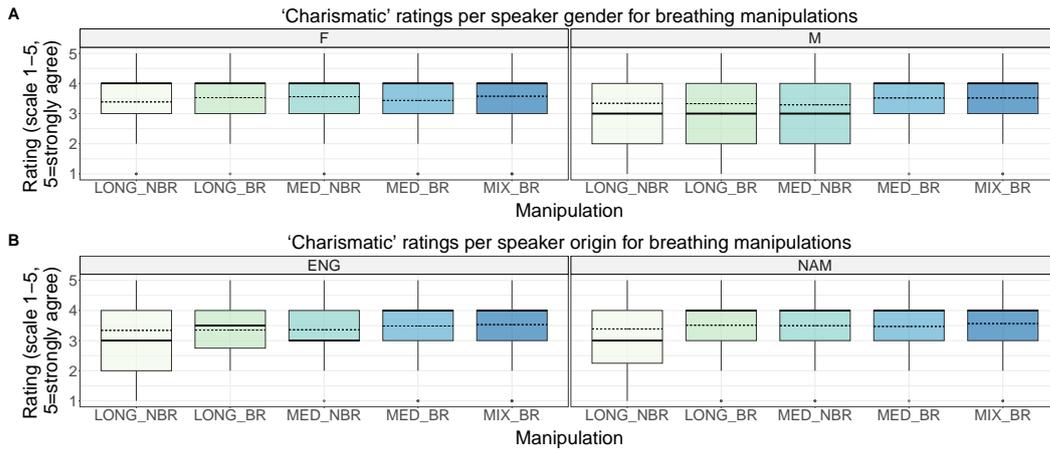


Figure 9.12: The results of the *charismatic* ratings of the stimuli with breathing manipulations, split by A) *Gender* and B) *Origin*. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

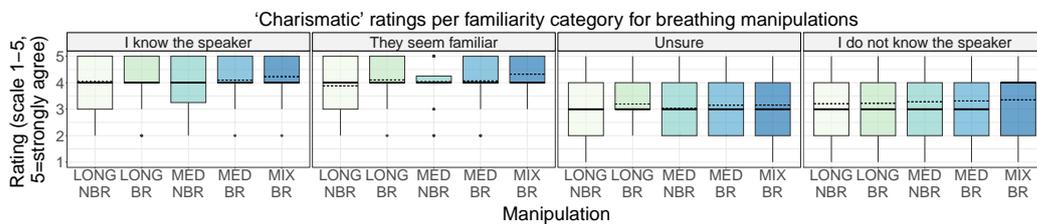


Figure 9.13: The results of the *charismatic* ratings of the stimuli with breathing manipulations per familiarity category. A solid horizontal line marks the median, a dashed horizontal line the mean. If only the median is visible, both fall on the same rating value.

9.6 Discussion

9.6.1 Summary of the experiment results

The results of the analyses of the attribute-related ratings are summarized in Table 9.10. The most likely candidates of acoustic characteristics for each speaker group and rating attribute that are included in the table are mostly based on the descriptive statistics presented above, unless specifically marked. Below, those results are also summarized in more detail.

In general it seems that for all attributes, with the exception of enthusiasm, different stimuli are preferred for male versus female speakers, as well as speakers from North America and England. And while there also are many commonalities between the different speaker groups within each rating attribute, different stimuli seem to be preferred.

Enthusiasm seems to be an exception in this case, so this is where the summary starts. When looking at the pause duration results, the descriptive statistics suggest that for all speaker groups (male, female, North American, and England) either the mixed condition (MIX_L) with one long, one medium, and one short pause, or the

Table 9.10: An overview of the results from the influence of the pause-related (duration and breathing noise) manipulations on the charisma-adjacent attributes *authentic* (AU), *enthusiastic* (EN), *likable* (LI), and *persuasive* (PE) as well as the direct charisma ratings (CH), including the effects of *Familiarity* (Fam.). Unless marked, the given feature characteristic represents the most likely preference based on descriptive statistics. A * marks a significant effect, and a ⁺ marks a non-significant trend from the inferential statistics.

| | | Features | |
|-----------|---------------|--|--|
| | | Pause duration | Breathing noise |
| AU | female | MED | MED_NBR or LONG_BR; LONG_BR > LONG_NBR ⁺ |
| | male | SHORT | MIX_BR or LONG_NBR; LONG_BR < LONG_NBR ⁺ |
| | ENG | not MIX_L | MED_NBR |
| | NAM | MIX_L | MIX_BR |
| | <i>Manip.</i> | – | MED_NBR > LONG_NBR* |
| EN | female | MIX_L or CUT | MED_NBR or LONG_BR |
| | male | MIX_L or CUT | MED_NBR |
| | ENG | MIX_L or CUT | MED_NBR; > NAM ⁺ |
| | NAM | MIX_L or CUT | MED_NBR; < ENG ⁺ |
| | <i>Manip.</i> | – | – |
| LI | female | LONG, MED, or MIX_L | LONG_NBR or MED_BR |
| | male | LONG, MED, or CUT | MED_NBR |
| | ENG | CUT or MIX_L; > NAM ⁺ | LONG_NBR |
| | NAM | MED; < ENG ⁺ | MED_NBR |
| | <i>Manip.</i> | SHORT < rest ⁺ | – |
| PE | female | SHORT | MIX_BR |
| | male | MED | LONG_BR or MED_BR |
| | ENG | not SHORT | no pattern |
| | NAM | SHORT | MIX_BR ⁺ |
| | <i>Manip.</i> | – | – |
| CH | female | SHORT | no pattern |
| | male | SHORT | MED_BR or MIX_BR |
| | ENG | SHORT | MED_BR or MIX_BR |
| | NAM | SHORT | no pattern |
| | <i>Manip.</i> | – | – |
| | Fam. | known/familiar > unknown*, mostly for male speakers | known/familiar > unknown*, mostly for male speakers |

CUT stimuli were perceived as more enthusiastic than the other stimuli. This could not be tested statistically in the current investigation as the models did not converge, but it would be worthwhile to take another closer look in the future. Likewise, the breathing noise stimuli also seem to have a consistently preferred stimulus with the enthusiasm ratings: MED_NBR (plus LONG_BR for female speakers, which seems to be a tie).

Similarly, the pause duration stimulus that is the most likely candidate to elicit *charismatic* ratings is consistently the one with SHORT pauses with a duration below 200 ms. Equally, there seems to be a preference for audible breathing noises for a charismatic impression.

The other attributes, as mentioned before, do not have such consistent patterns. Looking at the *authentic* ratings suggests a few things. For pause duration, male and female speakers seem to be perceived as more authentic with different stimuli:

females with the MED stimuli, and males with the SHORT stimuli. Both speaker groups seem to be preferred with shorter pauses rather than long pauses, as the MED stimulus condition with 200 to 500 ms is still quite short. Speakers from England and North America differ as well. North American speakers seem to be perceived as more (consistently) authentic with the MIX_L stimuli, meaning variation in pause duration may be most relevant here. On the other hand, speakers from England get the lowest rating for the MIX_L stimuli, and there is no obvious preference pattern.

When looking at the manipulations of the breathing noises, again male and female, as well as North American and English speakers were rated more authentic with different stimuli. For female speakers (based on visual observations), MED_NBR or LONG_BR stimuli were rated as authentic most consistently, while this was the case for MIX_BR or LONG_NBR stimuli for the male speakers. The LMM showed that LONG_BR was rated higher than LONG_NBR for female speakers, but lower for male speakers, and that the rating difference between the two stimuli was larger for the female speakers than for the male speakers. Additionally, MED_NBR stimuli were generally rated as more authentic than LONG_NBR stimuli, though the rating difference between the two was significantly larger for the female speakers than the male speakers. However, this seems to mostly be a duration difference, as both stimuli involved do not contain breathing noises. As for origin differences, North American speakers seem to be rated as more authentic with MIX_BR stimuli, so having variety in pauses with and without breathing noises, while the speakers from England seem to be perceived as more authentic with MED_NBR stimuli.

In terms of pause duration preferences for the *likable* ratings, the higher rated (or less variably rated) stimuli range more or less across the board. The one stimulus condition that is perceived as least likable, though, are the SHORT stimuli which tended to be lower rated than the rest of the stimuli conditions (a non-significant trend). This is directly opposite to the most likely stimulus candidate for *charismatic* ratings. Additionally, North American speakers tended to be perceived as less likable than the English speakers in the pause duration data set.

As far as the breathing noise manipulations are concerned, the speaker groups have one thing in common: the most likely stimulus candidate to elicit *likable* ratings did not contain breathing (with the exception of the female speakers where there seems to be a tie with MED_BR stimuli). So it seems like a lack of audible breathing noises is perceived as somewhat more likable, but the pause duration in this data set differs depending on the speaker group: both female and English speakers are perceived as more likable with long pauses without breathing noises, while medium length pauses without breathing seem to be preferred for both male and North American speakers.

Then, the results for the *persuasive* ratings show that there are group differences for gender and origin. As far as the pause duration manipulations are concerned, the SHORT stimuli seem to be preferred for the female speakers, and the MED stimuli for the male speakers, which is exactly the opposite than the stimuli most likely related to authenticity in this study. When looking at the origin differences, the pattern is similar to the *authentic* ratings: there seems to be a preference for the North American speakers, and a stimulus condition that is perceived as less persuasive than the other stimuli for the English speakers. The stimulus condition in question was MIX_L for authenticity, but it is the SHORT stimulus condition for the *persuasive* ratings. For two of the speaker groups (female; NAM), the most likely candidates for higher persuasiveness ratings seem to align with those for *charismatic* ratings.

The results of the breathing manipulations have something in common: while the pause duration differs depending on the speaker group, all most likely stimulus candidates to elicit *persuasive* ratings include audible breathing noises. For the English speakers, no pattern was visible. There was a non-significant trend, though, that suggests that MIX_BR stimuli were rated as more persuasive than LONG_NBR stimuli, but that this was especially the case for the North American speakers. This is also what the visual inspection of the data suggested. It is furthermore in line with the *authentic* rating results which also suggest that variety in the stimuli (i.e., the mixed stimulus conditions) were more highly rated for the North American speakers.

Another observation regards the breathing noise manipulations. In cases where there (visually) seems to be a tie between two most likely preferred stimulus conditions, there is one stimulus with breathing and one without breathing. This is the case mostly for the female speakers (for the *authentic*, *enthusiastic*, and *likable* ratings) and to some degree also the male speakers with the *authentic* ratings. For the female speakers, the results for authenticity and enthusiasm align—the two most likely stimulus candidates are MED_NBR or LONG_BR—while the preferred stimuli for the *likable* ratings are the opposite (MED_BR or LONG_NBR).

The biggest influence on the charisma ratings was the familiarity of the experiment participants with the speakers. Overall, the more known a speaker was to the raters, the higher was the charisma rating. But the statistical analyses and the interactions therein suggest that the reality is more nuanced than that, and there seems to be a gender effect. For the male speakers—while there was no significant difference between the ratings for known and familiar speakers—all other comparisons were significant (or at least trends) for the pause duration manipulations, and two groups (known/familiar vs. unsure/unknown) were found for the breathing noise manipulations.

9.6.2 Manipulation effects on direct charisma ratings

The general hypothesis for this part of the study stated that stimuli with audible breathing noises and shorter pauses are perceived as more charismatic (H1). When looking at the results from the LMMs, this was not the case: there were no main effects or interactions involving *Manipulation* for either pause duration or presence of breathing noises. That suggests that overall, the chosen pause duration and breathing noise manipulations do not have a significant effect on how charismatic the speakers are perceived, at least not when the group means are compared, as is the case with LMMs. The tentative observations in the descriptive statistics (in particular, medians and variation), on the other hand, seem to confirm the hypothesis. But while there are patterns that go in the direction of the predictions, this needs to be understood as a tendency to point to what might be found in future, perhaps more targeted studies.

The stimulus that elicited the most consistent and at the same time high ratings for charisma was the SHORT stimulus condition for all four speaker groups (male/female, ENG/NAM). That suggests that—as also found by previous research (e.g., D’Errico et al., 2013; Niebuhr et al., 2020a)—short pauses are also perceived as more charismatic in the context of YouTube. When looking at the average pause duration values of Steve Jobs and Mark Zuckerberg that are presented by Niebuhr et al. (2020a), the shortest average pause duration is the one by Mark Zuckerberg in his investor-oriented speech and its standard deviation is valued between around 300 and 600 ms. The other standard deviations start in similar places, but reach much higher, up to about 1,000 ms. The 300 to 600 ms pause duration range sits somewhere in the MED category at the border to the LONG pause duration category in the current investigation. That means that it is likely that on YouTube, pauses should be even shorter than what is generally considered short (the stimuli in the SHORT category in the experiment have pauses with durations below 200 ms) to be perceived as charismatic. This could be a genre effect that would be in line with what tends to be expected for pauses on YouTube: since YouTube as a whole is considered to be fast-paced, long pauses (or at least extremely long pauses) tend to be used to show awkwardness, while short pauses enhance the fast pace (The Film Theorists, 2020).

Similarly, and also tentatively, the stimuli that most likely and most consistently evoke positive charisma ratings include breathing noises, both with medium length pauses (MED_BR and MIX_BR). This was only the case for male speakers and speakers from England; no pattern was obvious for female speakers and North American speakers. The finding for male and English speakers (which may be connected, as there are more male than female speakers in the group of speakers from England) could be in line with previous research as well which suggests that

audible breathing noises, especially when they are short and intense, tend to be perceived as more charismatic—in particular, the audible breathing noises could be a result of faster chest breathing which has been shown to be advantageous for acoustic prosody in general (Michalsky and Niebuhr, 2019). Additionally, the mix of two pauses with breathing noise and one without could be understood as more variable and therefore more lively and expressive, which is also connected to charisma. This finding at the same time also enhances the previous pause duration finding. In the data set that was analysed for the breathing noise manipulations, only the LONG and MED stimuli were included, as was the mixed condition which also had medium-length pauses (200-500 ms). The stimuli with short pauses and cuts were excluded because they did not contain breathing noises. The result here therefore again suggests that shorter pauses may be perceived as more charismatic than longer pauses, and this time the MED stimuli are also within the range of pause durations of Mark Zuckerberg, though still significantly shorter than those of Steve Jobs, who is considered the more charismatic speaker (Niebuhr et al., 2020a).

9.6.3 Pause duration effects on attribute ratings

Moving into the hypotheses specifically put forward for the experiment dedicated to the pause durations and breathing noises, the first one predicted that stimuli with short pauses would be rated as more charismatic, enthusiastic, authentic, and likable than medium or long pauses, or cuts (H_{p21}). This was generally not the case in the current sample with a few exceptions. The SHORT stimuli seem to be preferred for the *charismatic* ratings (as was already discussed above). Additionally, male speakers tended to be rated as more authentic with SHORT stimuli as well, compared to the other pause durations. The ratings were not higher with short pauses than the other pause durations for the *enthusiastic* and *likable* ratings. In particular for the *likable* ratings, this was the only (non-significant) trend that included *Manipulation*, and it suggests that overall, SHORT stimuli were rated as less likable than the LONG stimuli, but also as less likable than all of the stimuli. The *likable* and *charismatic* findings (discussed above) therefore seem to be in direct opposition which suggests—at least in the context of YouTube—that speakers may have to find a balance between being fast and charismatic on the one hand, and likable on the other hand. That may be a possible explanation for the prevalence of pauses between 200 and 500 ms in the corpus (see Figure 9.1 in Section 9.3.1). These pauses were classified as medium in the current project, but are considered short elsewhere, whereas the short pauses in the current project are considered micropauses elsewhere (e.g., Selting et al., 2009). That means that the medium pauses in this experiment could still be considered as short (but average or medium in the context of YouTube) and therefore be long enough to be among the preferred

stimuli for likability, but also short enough to be perceived as charismatic, especially because the preference of the SHORT stimuli for charisma in the sample is only based on subtle visual patterns. With this in mind, the hypothesis seems to be tentatively supported for all attributes as there are always stimuli among the most likely candidates (see Table 9.10) that include medium length pauses.

Additionally, H_{P22} predicted that the CUT stimulus condition would be perceived as more enthusiastic than the other pause durations, but that it would receive the lowest ratings for all other attributes. While the second part was not the case in the current study (the CUT stimuli were also one of the stimuli perceived as more likable, and they are tied with others in many other cases), the first part seems to be accurate. Again, the results are based only on the descriptive statistics, but the CUT stimuli seem to be rated as most enthusiastic for all speaker groups, together with the MIX_L stimuli. Both of these stimuli could make a stimulus or in general a stretch of speech more fast-paced and therefore lively and expressive. The cuts do that by excluding time to breath or think, and creating the impression the speaker is extremely excited about what they want to say and intend to let you know quickly. At the same time, this is also a popular strategy on YouTube and therefore directly connected to the speech genre (The Film Theorists, 2020), though it can also be perceived negatively by the audience (see Hacker News, 2015). The mixture of pause durations in a stimulus could achieve the same thing—liveliness, expressiveness, excitement—by having variety and therefore being less monotonous as if the pauses were all of similar length. Equally, this condition draws more attention to the content, as the longest pause was placed at the strongest syntactic boundary of the content. It could therefore also show enthusiasm for what is being said. This needs to be investigated further and perhaps controlled even more in future studies.

The next hypothesis related to pause durations states that long pauses would be perceived as more persuasive than the other pause durations, and be lower rated for all other attributes (H_{P23}). This was not the case. Short or medium pauses seemed to be preferred for persuasiveness for the different speaker groups, though for English speakers the SHORT stimuli were explicitly not perceived as persuasive. Still, the LONG stimuli were either on par with other stimuli conditions, or they were rated more neutrally. Only for male speakers were the LONG stimuli rated as similarly persuasive than the preferred MED stimuli, though with more rating variation. However, the LONG stimuli are also not consistently rated the lowest on the other attributes which suggests that they are perceived as more neutrally, but that persuasiveness as well as the other attributes are less connected to long pauses, but rather adjusted to the fast pace of YouTube. It could also be that these results change if the visual channel is included. A previous case study of Barack Obama's speech suggests that shorter pauses tend to have fewer facial ges-

tures (though this is also linked to the syntactic position of the pause; Banzina and Niebuhr, 2023), and increased visual communication but with longer pauses could then be more persuasive also on YouTube.

The final pause-duration related hypothesis predicted that having a mixture of long, medium, and short pauses in stimuli (that is, the MIX_L condition) would receive the highest ratings on all attributes because of its variability (H_{P24}). This is also not the case. The MIX_L stimuli are never rated exclusively negatively, but rather neutrally, and they are included in the most likely preferred stimuli conditions for at least one speaker group with the *authentic*, *enthusiastic*, and *likable* ratings, but not with the *persuasive* and *charismatic* ratings. For example, the MIX_L stimuli tend to be rated as more authentic for North American speakers, but explicitly less authentic for speakers from England compared to any of the other stimuli. For the *enthusiastic* ratings, it is very consistently one of the two most likely preferred stimuli, together with the CUT stimulus condition (as discussed above), across all four speaker groups. For the speakers from England as well as female speakers, MIX_L stimuli are also included in the most likely preferred stimuli for the *likable* ratings. This condition is more neutrally rated or on par for the *persuasive* and *charismatic* ratings. This together suggests that the variation in pause duration may be helpful for some speaker groups and attributes, but not necessarily for charisma directly. But it also does not seem like it is a disadvantage to speak with varied pause durations.

Table 9.1 in Section 9.2 above presented the expected ratings for the different attributes and pause durations in the form of hierarchies. However, this does not fit the results of the data, as there are few if any obvious patterns or statistical results that allow for creating a hierarchy. The only point where this is slightly possible—and where the results seem to be in line with the expectations—are the stimuli resulting in higher *enthusiastic* ratings (MIX_L and CUT) which were consistently the higher rated stimuli, though this could not be tested by means of a LMM. In general, as mentioned, there are in general very no statistically significant rating differences, one trend (as far as *Manipulation* is concerned), and mostly subtle visual differences in the results of the pause duration manipulations. It is possible that the pause duration manipulations were too subtle to be picked up and evaluated by the participants. Previous research suggests that a gap in conversation can be detected with a duration of 120 ms (Heldner, 2011), but other research places the detection threshold for a pause between 100 and 300 ms (see Heldner and Włodarczak, 2016 for an overview), which would cover a substantial amount of the medium pauses and all of the short pauses in the experiment. Unlike the prosodic manipulation experiment, there was no pilot study run evaluating the stimuli used in this experiment. This should be remedied in a future study and the results should then be revisited and put into more context. Similarly, it could be that the manipulation

method of cutting sections of longer pauses to make them shorter or pasting other sections in to make them longer resulted in less natural sounding pauses. Testing a different method like extending or shortening the existing pauses using the TD-PSOLA algorithm in Praat (Charpentier and Stella, 1986; Moulines and Charpentier, 1990; Boersma and Weenink, 2018) may also be of interest. It could also be that English L1 speakers are not sensitive enough to pause durations in their own language, which is what was investigated here. It might be that L2 speakers react differently. This could also be a topic for future research.

9.6.4 Breathing noise effects on attribute ratings

Turning the discussion over to the breathing noise manipulations, the first related hypothesis expected that stimuli with audible breathing noises were predicted to receive higher ratings on all attributes (H_{P25}). This seems to be partly the case with the current sample. For the *authentic*, *persuasive*, and *charismatic* ratings, the majority of likely preferred stimuli include breathing noises. This is in line with research on charismatic speech (Michalsky and Niebuhr, 2019), which is then closely related to persuasiveness. In terms of authenticity, it might be that having breathing noises present makes the audio (and video, for that matter) seem more authentic—and therefore also the speaker—because it shows that there was no cutting involved in these instances or at least the naturally occurring breaths were left intact. However, for authenticity, male and female speakers also each had one non-breathing, most likely preferred stimulus, though this differed in pause duration from the breathing stimulus. This is discussed further below. On the other hand, the most likely stimuli to be perceived as enthusiastic or likable had pauses without breathing noises (with two exceptions for female speakers who also had similar ratings for a stimulus with pauses without breathing, though again the durations differed). It is unclear at this point why speakers tended to be perceived as more likable without audible breathing noises. For enthusiasm, it seems reasonable to assume that pauses without breathing are more immediate and the speaker aims to get their information out as fast and efficiently as possible, for which frequent and intense breathing could be detrimental. It has to be mentioned at this point as well that it could be that the pauses in the original stimuli do contain breathing, but that it is simply not audible and also was not visible in the video. It is striking that the results for authenticity, persuasiveness and charisma seem to align here, which was the same in the investigation of the prosodic manipulations for pitch range and final contour shape (Chapter 8). This may suggest that those three attributes are more closely related than enthusiasm or likability, at least in the context of YouTube.

The second breathing-related hypothesis predicted that the mixed condition with

two pauses with breathing noises and one without would get the highest ratings for all attributes (H_{P26}). This was generally not the case. The MIX_BR stimuli seem to be perceived as more authentic and persuasive than the other stimuli for North American speakers. This could potentially be connected to the stereotype of North American speakers apparently being more expressive than British speakers (Lewandowska-Tomaszczyk and Wilson, 2021), since the variety in breathing versus non-breathing could be seen as an effort to be less monotonous and more lively, though this cannot be seen as a definitive result. Tentatively, the same is the case for the charisma ratings, where the MIX_BR stimuli seem to be one of the preferred stimuli for the male speakers and for the speakers from England. This could again be related to expressiveness, and since the results for North American speakers simply showed no pattern, this also neither confirms nor denies the interpretation with stereotype in mind of the persuasive ratings.

In general, there again were only one significant result, one trend, and otherwise very few obvious visual patterns regarding the specific manipulations. However, it seems as if the *authentic*, *persuasive*, and *charismatic* ratings are encoded similarly in terms of breathing noises, as are the *enthusiastic* and *likable* ratings. Perhaps future studies with larger samples or revised manipulation methods can collect clearer results.

One further observation that can be made for the breathing noise manipulations that was not part of the hypotheses is that—as far as the attributes *authentic*, *enthusiastic*, and *likable* are concerned—when there are two most likely stimulus candidates for higher ratings, one of them had breathing noises, the other one did not, but they also differed in their duration. When there were two options, one with breathing noises, the other without, one of the stimuli had long pauses, the other medium pauses. The combination of the two features differed depending on attribute and gender (which leads to the discussion of the next hypothesis below). But it suggests that either there are certain combinations that elicit certain attributes, or that there are participants who may prefer one combination over the other.

9.6.5 Speaker gender effects

Directions of possible gender effects were initially not predicted for the pause duration and breathing noise manipulations (H_{P27}). The only attribute where there were consistently no gender effects is enthusiasm: both male and female speakers received higher ratings for the MIX_L and CUT conditions (pause duration), and the MED_NBR condition (breathing noise). For female speakers, the LONG_BR stimuli were equally highly rated for enthusiasm, but the two speaker groups have the MED_NBR stimuli in common. Similarly, both male and female speakers seem

to be rated as more charismatic with SHORT stimuli, though this is only a very tentative observation.

For the rest of the attributes, the ratings for stimuli for male and female speakers seem to differ. For authenticity and pause duration, female speakers tended to receive higher and more consistent ratings with medium length pauses and male speakers with short pauses, while this pattern is reversed for the *persuasive* ratings. This may suggest that since the *authentic* and *persuasive* attributes seem to align to some degree when it comes to acoustic features—in terms of breathing noises, in this case—maybe pause duration is the feature that sets the two attributes apart. It could be that women are perceived as more persuasive when they get the information out fast (since this may also mean that they are well-prepared and do not have to spend much time planning what they will say next), but more authentic when they take a little bit more time. For the male speakers, this could be the opposite. However, compared to average pause durations suggested in previous research (e.g., Selting et al., 2009), both SHORT and MED stimuli can be considered as short. If the results are interpreted with this in mind, the gender difference for the *authentic* and *persuasive* ratings disappears and they again align with the *charismatic* results.

For the breathing noise manipulations, there are also gender differences. In terms of the perception of persuasiveness, male and female speakers seem to have one thing in common: the most likely preferred stimulus includes audible breathing noises. While for the male speakers, both long and medium pauses elicit similar results, the mixture and variety of the MIX_BR stimuli receives the highest ratings for the female speakers. It seems like having shorter pauses with variable breathing is advantageous for female speakers, since it can portray liveliness, expressiveness, and previous planning which might increase the amount of persuasion. It could be that this is an example of something where female speakers may have to do more (variety, planning, etc.) in order to be perceived as persuasive, while male speakers do not have to put as much effort in (as is the case with other prosodic features like emphatic accent frequency; see Novák-Tóth et al., 2017). When looking at the *likable* ratings, these differences could be either duration-related (female: LONG_NBR; male: MED_NBR) or breathing-related (female: MED_BR; male: MED_NBR), as the female speakers seem to have two contenders for the most likely preferred stimulus. It is unclear at this point how these results can be interpreted and what they might mean, especially since there are no significant differences from the LMMs.

In terms of authenticity, there were also gender differences. Both male and female speakers show two most likely stimulus candidates to elicit higher *authentic* ratings. For the female speakers, they are the MED_NBR and LONG_BR stimuli, and the MIX_BR (with all pauses of medium length) or LONG_NBR stimuli for the male speakers. That shows that both speaker groups have two possible combinations of

features, but the combinations are opposite for them. Where the female speakers seem to be perceived as more authentic with medium pauses without breathing noises, the male speakers are rated as more authentic with medium pauses and breathing noises. The opposite is the case for the stimuli with long pauses. This could hint towards actual perception differences between male and female speakers when it comes to authenticity and breathing noises (and pause durations). It can also not be ruled out that there may have been unforeseen issues with the manipulations that may have led to biases in the ratings, though this seems less likely as the ratings were very high (see Figure 9.6 in this chapter), both for the male and the female speakers. In line with the descriptive findings on the stimuli with the long pauses detailed above, there was also a non-significant trend for an interaction between *Gender* and *Manipulation* which suggests that LONG_NBR stimuli are rated as more authentic than LONG_BR stimuli for the male speakers, but as less authentic for the female speakers. This may suggest that especially for female speakers, breathing noises are important to have when the pauses are longer than 500 ms. Otherwise they are likely no longer seen as authentic and real.

This together shows that—contrary to initial expectations—there are gender differences in how male or female speakers are perceived based on stimuli related to pause duration and presence or absence of breathing noises. While these results are still tentative, future studies should take a closer look at gender differences with a larger sample and also expanding the gender identities investigated. Likewise, taking the gender identity of the experiment participants who rated the speakers into account is beyond the scope of this investigation. It might, however, provide further insights into the matter of gender and charisma (and related attributes) perception, since previous research did not only reveal speaker gender differences (Jokisch et al., 2018; Niebuhr et al., 2018a; Niebuhr and Wrzeszcz, 2019; Niebuhr et al., 2019; Gutnyk et al., 2019; Niebuhr, 2020), but also initial evidence for audience gender differences (Gutnyk et al., 2020).

9.6.6 Speaker origin effects

Similar to the expectations for speaker gender, the next hypothesis predicted speaker origin differences, but not specific directions of differences (H_{P28}). For most attributes, origin differences were found which will be discussed further below. Exceptions were the pause durations and the *enthusiastic* and *charismatic* ratings, and the breathing noise manipulations and the *enthusiastic* ratings, where the highest rated stimuli matched between speakers from England and North America. For the *charismatic* ratings and the breathing noise manipulations, it was a comparison of a tentative possible preference for MED_BR or MIX_BR stimuli for the English speakers, and no pattern at all for the North American speakers. The results

here are therefore too inconclusive to interpret. Similarly, for breathing noise manipulations and the *persuasive* ratings (the second exception), a comparison is also difficult. There were no obvious patterns for speakers from England, and North American speakers were rated as more persuasive with MIX_BR stimuli which differed from LONG_NBR stimuli more than the same comparison for English speakers. In general, though, this is still too little information to be able to interpret the results accurately.

When looking further into the pause duration manipulations and origin, the results for *authentic* and *persuasive* seem to be similar for English and North American speakers. For both attributes, there seems to be a preferred stimulus for the North American speakers—MIX_L with authenticity, SHORT with persuasiveness. On the other hand, the same stimuli are rated least authentic/persuasive for the speakers from England. In these cases there is no pattern at all in terms of what might be preferred, but the MIX_L/SHORT stimuli stand out negatively. This may suggest that the two speaker groups are expected to encode authenticity and persuasiveness differently and the raters agree more on how that should be done for the North American speakers, but for the English speakers (rated by British listeners) there is less consensus. This suggests that authenticity may be connected to variation for North American speakers, and to consistency for speakers from England, at least to listeners from England. As with many other results, this opens more questions.

In terms of breathing noises, there seem to be some similarities in the most likely stimulus candidates to elicit *authentic* ratings: both candidates for English and North American speakers include audible breathing noises, and both have medium pause durations. The stimulus for speakers from England has less variation in breathing, though, since all three pauses include breathing noises (MED_BR), while the stimulus for speakers from North America in the sample has one pause without breathing noise (MIX_BR). This suggests that both speaker groups are similar to each other in terms of authenticity, just that North American speakers seem to be rated as more authentic with more variation. This is somehow similar to the *likable* ratings: both speaker groups have most likely preferred stimuli including breathing noises, here they differ in pause duration, though. Speakers from England were perceived as more likable with long pauses, and speakers from North America with medium length pauses. This may suggest that North American speakers are perceived as more likable when they are a little bit faster and perhaps expressive, while English speakers are perceived as more likable when they take their time, consider, and slow down.

Additionally, there were non-significant trends of main effects that suggest origin differences. In the data set with the pause duration manipulations, speakers from England were rated as generally more likable than speakers from North America. It

is unclear at this point why this effect came up only for the pause duration manipulations. It may be also present in the analyses of the breathing noise manipulations, but the statistical model did not converge and could therefore not be reported and interpreted. Additionally, this effect did not occur for the likable ratings with the prosodic manipulations in Chapter 8, but the data were collected from different participants, so a participant effect cannot be ruled out here, either. Another possible explanation could be that the intonation contours (falling, rising, plateau, as well as pitch accents) differ between speakers from England and North America, which was not controlled in the stimulus selection. However, a feature like final contour direction did not have a significant influence on the *likable* ratings in the analyses of the previous chapter where it was controlled, but an effect in connection with pauses and breathing cannot be ruled out.

Similarly, in the data set with the breathing noise manipulations, speakers from England tended to be rated as more enthusiastic than speakers from North America. This was visually also the case for the pause duration manipulations, but that model also did not converge, so information from the inferential statistics is not available. This result is in line with those from Chapter 8 where independent of the manipulated acoustic feature, the North American speakers were also rated as less enthusiastic than the speakers from England. Keeping in mind the stereotype put forward by Lewandowska-Tomaszczyk and Wilson (2021), this seems counter-intuitive, as the stereotype states that North American speakers tend to be known as more expressive. It could be a case of regional variety (in the broad sense of British versus North American English), since all listeners were from England, and the results may change with North American listeners in the sample.

In general, similar to the gender results, there are differences especially in some of the charisma-adjacent attributes between speakers from England and North America. The results regarding charisma directly are more inconclusive in the current sample. However, all these elements together suggest it is necessary to look at charismatic speech as well as enthusiastic, persuasive, etc. speech from different angles. So far, all published research on charismatic speech in English has focused on North American English. Especially the results from the other attributes like *authentic* or *persuasive* suggest that there may be more fruitful information in a more detailed comparison.

9.6.7 Familiarity effects

This chapter also addresses the third research question dealing with the relationship between charisma and familiarity (see also Chapter 5). The corresponding hypothesis (H3) predicted that there would be positive correlations between the charisma ratings and the familiarity, and that this was the case for both data sets

(pause duration manipulations and breathing noise manipulations) independent of the specific manipulations. This was generally the case. There were significant positive correlations for both data sets suggesting that indeed, the more familiar a speaker is to the listeners, the more charismatic they are rated. There were also significant interactions and pairwise comparisons between *Familiarity* and *Gender*, though, that suggest that there is much more nuance to these correlations. They show that there are differences between how charisma and familiarity play together for male and female speakers, at least in the current sample. When looking at the results from the male speakers, it seems like there are two groups of familiarity: known/familiar speakers on the one hand, and speakers the listeners are unsure about and unknown speakers on the other hand. This comes out clearly in the pairwise comparisons of the breathing noise data, and less clearly for the pause duration data. There, the pairwise comparisons suggest the same thing, but there was a non-significant trend that also suggests there might be a difference between unsure and unknown. This would then suggest that—at least in the pause duration data—there may be actual degrees of familiarity that could have an effect on charisma ratings, and it would mean that it is important for male speakers to be more known.

Crucially, there was a significant pairwise comparison between known and unknown male speakers, but the difference in charisma rating between the known and unknown female speakers was not significant. For the female speakers, only the comparison between the ratings for familiar and unknown speakers was significant. That suggests that for male speakers, knowing the speaker is generally beneficial for charisma, but there is no difference between speakers the participants really knew and those they just seemed familiar with. On the other hand the results suggest that knowing a speaker can be just as detrimental than not knowing the speaker when they are female. It is here that it seems to come into play if the speaker is actually liked by the listeners. This was not information that was elicited in the experiment but should be included in future experiments. It is likely that this effect came out for the female speakers because there are a few speakers in the sample that have been criticized in the media for business decisions they made, and these criticisms were picked up by traditional media and may therefore have increased the number of people who know the speakers, but likewise also reduced the amount a speaker is liked and supported (Deller and Murphy, 2020).

Additionally, this experiment (and the experiment in Chapter 8, which had similar findings) suggests that the concept of familiarity is much more nuanced than often included in research. Previous research that included charismatic speech and familiarity (e.g., Rosenberg and Hirschberg, 2009) only asked the participants if they recognized a speaker or not. Future studies should—like the present study—consider breaking down the degree of familiarity into smaller, more specific

categories in order to get more fine-grained insights into familiarity and charisma. This could also be combined with questions aiming to find out if the listeners know or recognize a speaker because they like them or specifically because they actually dislike them. Additionally, especially in the context of YouTube, it might also be interesting to find out if the listeners would consider watching videos or even subscribing after hearing a speaker, as this would give further, though more indirect, information.

Chapter 10

Perception 3: Acoustics and perception

10.1 Introduction

The last two chapters have investigated the influence of specific manipulations on the perception of charisma and the four related attributes *authentic*, *enthusiastic*, *likable*, and *persuasive*. This chapter investigates other acoustic features known to be relevant for charisma perception of the experiment stimuli. The manipulations were restricted to pitch level, pitch range, final contour shape and speech rate as well as the duration of pauses and presence or absence of breathing as a starting point for analyses. These manipulations are based on more global features that could be manipulated sufficiently in Praat (Boersma and Weenink, 2018) and at the same time have been shown to matter for charismatic speech by previous research (e.g., Rosenberg and Hirschberg, 2009; D’Errico et al., 2013; Niebuhr et al., 2020a).

The results of the previous chapters suggest, though, that there was generally little influence of the manipulations on the attribute ratings, at least in terms of mean ratings. The origin and gender of the speaker had more of an influence, but this was also limited. Especially the familiarity with a speaker was relevant for the direct charisma ratings, rather than the manipulations. The lack of significant results from the manipulations suggests that looking at these global manipulations may not be enough. It is therefore relevant to investigate what else is going on in the stimuli. In particular, this chapter looks at a selection of relevant acoustic features in the unchanged experiment stimuli and correlates the acoustic measurements and frequency counts with the mean rating a stimulus received. This method was tested and already used for the analysis of the vowel spaces of the same speakers in a previous study (Berger et al., 2023). The investigation is based on both the short stimuli that were used for the pitch-related and speech rate manipulations, and the long stimuli used for pause duration and breathing manipulations (Chapters 8 and 9), though they are analyzed separately since the experiment participants differed. All five attributes (*charismatic*, *authentic*, *enthusiastic*, *likable*, and *persuasive*) are included, though only the significant and descriptively relevant results are reported.

Using these methods, which are explained further in Section 10.3, the second

research question put forward in Chapter 5 is addressed. It is repeated as RQ2 below.

RQ2: *Do speakers in the sample who receive higher ratings for charisma and/or charisma-adjacent attributes in the perception studies employ acoustic features in ways that have been reported in other charisma literature?*

That means that this chapter works towards combining the acoustics and the perception side of the project. Only a selection of acoustic features relevant for charismatic speech are addressed here, namely a selection of more local pitch features (minimum, maximum, and mean F0, median pitch, excursion size, and the standard deviation of pitch), intonation features (pitch peak timing, prominence ratio, and frequency of different prominence levels), and tempo- and duration-related features (phrase duration, stimulus duration, speech rate, and speech rate variation). These features were chosen because they represent a combination of frequently researched and un-researched acoustic features in the context of charismatic speech and can therefore both replicate results, but also steer research in other directions. Additionally, many of the other features connected to charismatic speech are also highly dependent on the room acoustics, recording environment and equipment, post-processing and data compression (see, e.g., Zhang et al., 2021; Siegert and Niebuhr, 2021b; Berger and Neitsch, 2023), for example all voice quality and intensity-related features. Since equipment, environment, and processing information is not available for YouTube videos, and data compression always plays a role on YouTube and is more detrimental for intensity and voice quality, these features were excluded from the analyses for the time being. The hypotheses and assumptions for this study are collected in Section 10.2. Section 10.3 introduces the methods used in more detail, before the results are presented in Section 10.4.

10.2 Hypotheses

The general hypothesis for this study suggests that the ranking of the speakers based on the mean acoustic measurement of the stimuli is correlated with the mean rating response that speakers' stimuli received. The direction of the correlation (positive or negative) depends on the specific feature. The general hypothesis is repeated as H2 below.

H2: *The ratings from the experiments correlate with the acoustic feature values known to be used in charismatic speech and related attributes.*

Six pitch measurements are included in this study which all represent specific local aspects of the utterances: mean pitch level, median pitch, pitch variability, pitch

range, minimum and maximum F0. Pitch range and pitch level were also investigated in terms of the effect of manipulation in Chapter 8, but are addressed here explicitly regarding their actual measurements. Previous research suggests that larger variability and range can be expected (for both male and female speakers; see, for example, Niebuhr et al., 2018a; Niebuhr and Skarnitzl, 2019) for charismatic speech, as well as higher minimum and maximum F0 (at least for male speakers; see, e.g., Signorello et al., 2012a, 2012bb; D’Errico et al., 2013; Mixdorff et al., 2018). Since there is no previous research that suggests a gender difference for these features and its relation to charismatic speech, none is expected for this study. Since these features are connected to more vocal effort, it might be that they are perceived as less authentic and likable in the context of YouTube (see Section 4.3.4). The hypothesis is shown in H_{P3}1 (the subscript _{P3} indicates that these hypotheses are specifically connected to this third perception study). Additionally, previous research showed that male speakers are perceived as more charismatic with higher mean and median pitch, and female speakers when speaking with a lower mean and median pitch level (see Novák-Tót, 2016; Niebuhr et al., 2018a; Niebuhr et al., 2020a), which is also what is expected for the YouTubers in the current sample for charisma, persuasiveness, and enthusiasm, but not for authenticity and likability (H_{P3}2 and H_{P3}3).

H_{P3}1: *Speakers with larger pitch variability and range, and higher minimum and maximum F0 are perceived as more charismatic, persuasive, and enthusiastic, but less authentic and likable.*

H_{P3}2: *Male speakers with higher mean and median pitch receive higher charisma, persuasiveness, and enthusiasm ratings, but lower authenticity and likability ratings.*

H_{P3}3: *Female speakers with lower mean and median pitch receive higher charisma, persuasiveness, and enthusiasm ratings, but lower authenticity and likability ratings.*

In terms of intonation (i.e., features that are related to more global melody patterns), three features were investigated. First, there is the pitch peak timing, meaning the time difference between the accented vowel onset and the pitch peak of the accent in percent—the larger the positive value, the later the peak; the larger the negative value the earlier the peak; values close to zero percent coincide with the accented vowel onset. Previous research suggests that charismatic speech uses more later pitch accents (Biadys et al., 2007; Berger et al., 2020). Additionally, a previous study of the current speaker sample showed that the YouTubers used early peaks (i.e., peaks occurring early in or even before the prominent syllable) significantly less frequently than medial or late peaks (Berger and Zellers, 2021), so combining these previous findings would suggest that using later peaks is perceived as

more charismatic, persuasive, and enthusiastic. Since later peaks also go in hand with increased vocal effort (Gussenhoven, 2002), it might again lead to lower ratings in authenticity and likability in the context of YouTube, since YouTubers are generally expected to be laid back, relaxed, and genuine, as if they were engaged in a friendly conversation (Kyncl and Peyvan, 2017), which was also suggested by results of the vowel space in Berger et al. (2023). This is summarized in H_{P3}4. The second feature is the prominence ratio—this is the number of prominent divided by the total number of syllables in a phrase for the ratio, which is then normalized by phrase duration¹. As far as the prominence ratio in a stimulus is concerned, there has not been any research on this feature and its connection to charisma perception, as far as the author is aware. This is therefore included as an exploratory feature. Similar to the pitch peak timing, though, it is predicted that a higher ratio of prominent syllables in a stimulus is connected to higher vocal effort and therefore assumed to be perceived as more charismatic, persuasive, and enthusiastic, but perhaps again less authentic and likable (H_{P3}5). The third intonation-related feature is the frequency of prominences of different levels (weak, strong, emphatic) which is normalized to count per minute (cpm). The predictions here are in general the same as for the previous two features, though this is the case in particular for the emphatic accents: more emphatic accents are predicted to be perceived as more charismatic, persuasive, and enthusiastic, but less authentic and likable (H_{P3}6). There is no research investigating frequency effects of the other two prominence levels, so this is also included in the analyses in an exploratory manner.

H_{P3}4: *Speakers using later peaks are perceived as more charismatic, persuasive, and enthusiastic, but less authentic and likable.*

H_{P3}5: *Speakers using a higher ratio of prominent syllables are perceived as more charismatic, persuasive, and enthusiastic, but less authentic and likable.*

H_{P3}6: *Speakers using more emphatic accents are perceived as more charismatic, persuasive, and enthusiastic, but as less authentic and likable.*

Additionally, four acoustic features relating to tempo and duration were analyzed: the phrase duration, stimulus duration, speech rate, and the variation of speech rate between phrases (this last analysis and the stimulus duration analysis are only run for the long stimuli). Shorter phrases are expected to be perceived as more charismatic and related attributes (H_{P3}7; see also Niebuhr et al., 2020a), but at the same time the longer a stimulus is, the the higher is the expected rating (H_{P3}8; see Jokisch et al., 2018). Overall, a medium speech rate is expected to be perceived as more *charismatic, persuasive, enthusiastic, authentic, and likable* than a lower

¹This is not the same as the prominence ratio in psycho-acoustics, which is “intended for detection and evaluation of prominent tones in noise emissions” (Lennström et al., 2013, p. 2).

or higher speech rate (H_{P3}9; see also Niebuhr et al., 2020a). For YouTube in particular, a speech rate between five and six syllables per second is expected to receive higher ratings. This is higher than both the range of business speakers investigated for charismatic speech (Ginni Rometty, Mark Zuckerberg and Steve Jobs) and average reference values (see Novák-Tót et al., 2017; Niebuhr et al., 2020a). This is only addressed in an exploratory manner, but could point towards a genre difference. Additionally, larger speech rate variations are also expected to be perceived as more charismatic (and its related attributes), at least as far as female speakers are concerned (H_{P3}10; see Novák-Tót, 2016). This second feature is only analyzed for the long stimuli since these stimuli contain four phrases in which speech rate can be measured and where the standard deviation between phrases therefore can also be measured.

H_{P3}7: *Speakers using shorter phrases are perceived as more charismatic, persuasive, enthusiastic, authentic, and likable.*

H_{P3}8: *The longer the stimulus is as a whole (including pauses), the higher is the rating expected to be.*

H_{P3}9: *Speakers using medium speech rates are perceived as more charismatic, persuasive, enthusiastic, authentic, and likable.*

H_{P3}10: *Speakers using larger speech rate variation between phrases are perceived as more charismatic, persuasive, enthusiastic, authentic, and likable.*

Generally, no differences between male and female speakers, as well as speakers from England and North America are expected. One exception to this hypothesis is the pitch level analysis, where previous research suggests specific gender differences (see H_{P3}2 and H_{P3}3). No effect of gender and origin is mainly expected because there is no research about charismatic speech available yet for speakers from England, and very little research regarding charismatic speech and female speakers. Another reason is that for some features (e.g., pitch variability, pitch range, speech rate), the same tendencies were found for both speaker groups (male and female, see Novák-Tót, 2016; Niebuhr et al., 2020a). Therefore, while no specific difference is expected, possible speaker gender and origin differences are included in the analysis in an exploratory manner. Differences are not unlikely, but the direction cannot be predicted. Therefore, no formal hypothesis is attached to this exploration, especially because the data set is too small for inferential statistics. All observations are based on descriptive statistics as a first step until a larger data set can be investigated.

10.3 Methods

10.3.1 Measurements

For the current part of the investigation, only the unmodified stimuli are analyzed in terms of acoustics, and only the ratings directly connected to these stimuli are included. The short and long stimuli are considered separately as they were rated by different groups of experiment participants. Additionally, the measurements for the long stimuli were averaged across the four stimulus phrases for the statistical analyses. The short stimuli were used in the analyses in Chapter 8 and consist of only one phrase (either an IPU or part of one at a minor phrase boundary). The long stimuli were used in Chapter 9 and consist of four phrases (again, IPUs or the beginning and/or end was selected at a minor phrase boundary) and three pauses with a length between 200 and 500 ms. For the content of the stimuli, see Section 8.3.1 in Chapter 8 and Section 9.3.1 in Chapter 9.

In most cases, the mean value was used for correlations with the experiment ratings, but in the case of the pitch accent timing, the median was deemed more relevant in light of vast variability between the phrases. All measurements were made for the full annotated corpus of speech data with the prepared audio files that were checked for outliers and adjusted (see Section 7.1). The measurements for the full corpus were made on IPU level. Some of the phrases in the stimuli correspond to minor phrase boundaries and not IPUs, though. That means that for these phrases, the measurements of the full corpus would not be representative. Therefore, the measurements were run a second time with the full corpus, but the analysis intervals were now set at the exact beginnings and ends of the stimuli phrases. The data set for analysis then consisted of the full corpus measurements, and additionally the measurements for the phrases of the short and long stimuli (five phrases total per speaker) which were marked as such and only used as subsets. The two data sets were randomly checked, and the measurements were either identical or only differed slightly. This means that the stimulus measurements are based on the same audio files as the full corpus data and therefore more comparable. The data from the full corpus are used to situate the stimuli within the larger data set to see if it is reasonable to generalize from the stimuli to the full data set. If the stimuli happen to be outliers when the rest of the data are considered, that would mean that all results can only be seen as representative for the stimuli sample, not the sample as a whole. The measurements of the stimuli are displayed below, for the short stimuli in Table 10.1, and for the long stimuli in Table 10.2.

The measurements for minimum, maximum, and mean F0, as well as excursion size and phrase duration were made with ProsodyPro (Xu, 2013, version 5.7.8.1)²

²Settings: target tier = "Phrases-PP" (3; see Section 7.2); F0 range: 60-600 Hz (male speakers), 75-600

Table 10.1: Acoustic measurements of the short stimuli. These measurements are used for correlations with the mean experiments ratings, and for displaying the data in ascending order for visual correlations.

| | Speakers (ENG) | | | | |
|--|----------------|--------|--------|--------|--------|
| | LP | ZS | AD | DH | PL |
| Maximum F0 (st, normalized) | 3.26 | 3.93 | 3.63 | 4.09 | 9.06 |
| Minimum F0 (st, normalized) | -2.81 | -5.1 | -5.46 | -11.96 | -10.69 |
| Mean F0 (st) | 12.11 | 14.52 | 10.16 | 11.34 | 6.25 |
| Median pitch (st) | 11.33 | 13.65 | 9.23 | 10.23 | 3.63 |
| Excursion size (st) | 6.07 | 9.03 | 9.09 | 16.04 | 19.75 |
| Pitch variability (standard deviation, st) | -32.44 | -21.14 | -23.65 | 13.81 | -17.39 |
| Accent timing (% , normalized) | 13.28 | 47.40 | 48.51 | 96.3 | 60.16 |
| Prominence ratio (normalized) | 0.21 | 0.15 | 0.12 | 0.2 | 0.13 |
| Phrase duration (s) | 2.40 | 2.52 | 2.39 | 2.02 | 2.73 |
| Speech rate (syll/s) | 4.99 | 5.16 | 7.12 | 4.94 | 6.23 |
| | Speakers (NAM) | | | | |
| | CB | LS | SP | MF | MP |
| Maximum F0 (st, normalized) | 5.64 | 6.97 | 4.3 | 11.03 | 5.05 |
| Minimum F0 (st, normalized) | -11.31 | -4.12 | -11.62 | -6.59 | -4.65 |
| Mean F0 (st) | 11.78 | 13.08 | 10.93 | 2.57 | 7.36 |
| Median pitch (st) | 9.74 | 11.54 | 9.98 | 2.00 | 5.12 |
| Excursion size (st) | 16.95 | 11.09 | 15.92 | 17.62 | 9.71 |
| Pitch variability (standard deviation, st) | -16.64 | -17.35 | -21.03 | -18.2 | -26.67 |
| Accent timing (% , normalized) | 137.92 | 33.58 | 41.33 | 5.97 | 40.52 |
| Prominence ratio (normalized) | 0.13 | 0.24 | 0.24 | 0.2 | 0.15 |
| Phrase duration (s) | 2.24 | 2.07 | 2.08 | 1.99 | 2.18 |
| Speech rate (syll/s) | 7.58 | 4.83 | 4.81 | 5.03 | 5.52 |

in Praat (Boersma and Weenink, 2018, version 6.0.37). The measurement for the standard deviation of pitch was added to the ProsodyPro script by the author. The other measurements (pitch accent timing, prominence ratio, frequency of different prominence levels, speech rate, speech rate variability, and stimulus duration) were made with Praat scripts specifically written for the project.

Minimum, maximum, mean F0 and median pitch were measured in Hz and converted to st³ in R for more comparability between different speaker groups (Gradol, 1986; Henton, 1995). Additionally, maximum and minimum F0 were normalized to the mean F0 of each phrase by subtracting the mean from the maximum or minimum value in order to ensure comparability. Excursion size was measured in st, and phrase duration as well as stimulus duration in ms.

Pitch peak timing was measured as the time difference of the position of the F0 peak maximum to the accented vowel onset in percent relative to the vowel duration. The pitch peak is connected to a pitch accent and prominent syllable. If the peak appears much later than the vowel onset—for example, late in the following syllable—this can lead to percentages much higher than 100 percent. For the analy-

Hz (female speakers); maximum formant: 5000 Hz (male speakers), 5500 Hz (female speakers); rest of settings: default.

³Formula for semitones with a reference tone of 100 Hz: $12 * \log(\text{feature value} / 100) / \log(2)$, following http://phonetics.linguistics.ucla.edu/facilities/acoustic/pitch_unit_conversion.txt

Table 10.2: Acoustic measurements of the long stimuli. These measurements are used for correlations with the mean experiments ratings, and for displaying the data in ascending order for visual correlations. The mean or median—if mentioned in parentheses—refers to the mean or median across the four phrases in each stimulus.

| | Speakers (ENG) | | | | |
|---------------------------------------|----------------|--------|--------|--------|--------|
| | LP | ZS | AD | DH | PL |
| Maximum F0 (mean; st, normalized) | 6.89 | 5.6 | 13.78 | 5.51 | 4.74 |
| Minimum F0 (mean; st, normalized) | -16.53 | -5.13 | -10.73 | -6.78 | -9.32 |
| Mean F0 (mean; st) | 16.7 | 13.7 | 7.2 | 9.4 | 12.63 |
| Median pitch (mean; st) | 13.98 | 12.63 | 4.3 | 8.75 | 10.59 |
| Excursion size (mean; st) | 23.43 | 10.74 | 24.51 | 12.29 | 14.06 |
| Pitch variability (mean; st) | -7.95 | -21.78 | -15.99 | -21.46 | -14.77 |
| Accent timing (median; %, normalized) | 54.39 | 31.83 | 43.22 | 57.34 | 54.2 |
| Prominence ratio (mean; normalized) | 0.14 | 0.23 | 0.14 | 0.26 | 0.66 |
| Weak prominence (cpm) | 23.70 | 45.25 | 45.69 | 32.19 | 29.16 |
| Strong prominences (cpm) | 71.09 | 36.20 | 49.84 | 51.51 | 87.48 |
| Emphatic prominences (cpm) | 15.80 | 40.73 | 16.61 | 45.07 | 19.44 |
| Phrase duration (mean; s) | 3.54 | 2.87 | 3.31 | 2.16 | 1.27 |
| Speech rate (mean, syll/s) | 4.63 | 4.98 | 5.95 | 5.31 | 6.02 |
| Speech rate variation (syll/s) | 0.46 | 0.58 | 0.60 | 0.62 | 1.55 |
| Stimulus duration (s) | 15.19 | 13.26 | 14.45 | 9.32 | 6.17 |
| | Speakers (NAM) | | | | |
| | CB | LS | SP | MF | MP |
| Maximum F0 (mean; st, normalized) | 6.22 | 2.59 | 7.45 | 5.66 | 5.11 |
| Minimum F0 (mean; st, normalized) | -4.63 | -3.2 | -5.93 | -6.00 | -4.49 |
| Mean F0 (mean; st) | 13.06 | 9.57 | 13.05 | 1.97 | 6.00 |
| Median pitch (mean; st) | 11.96 | 9.4 | 11.7 | 1.37 | 5.21 |
| Excursion size (mean; st) | 10.84 | 5.79 | 13.39 | 11.66 | 9.59 |
| Pitch variability (mean; st) | -19.42 | -40.2 | -18.08 | -28.34 | -31.85 |
| Accent timing (median; %, normalized) | 63.64 | 46.17 | 48.14 | 23.76 | 39.64 |
| Prominence ratio (mean; normalized) | 0.13 | 0.15 | 0.22 | 0.31 | 0.23 |
| Weak prominence (cpm) | 43.98 | 32.92 | 55.83 | 14.17 | 29.76 |
| Strong prominences (cpm) | 48.38 | 32.92 | 64.41 | 42.50 | 69.44 |
| Emphatic prominences (cpm) | 35.18 | 23.51 | 21.47 | 51.95 | 54.56 |
| Phrase duration (mean; s) | 3.18 | 2.90 | 3.28 | 2.94 | 2.80 |
| Speech rate (mean; syll/s) | 6.30 | 5.49 | 5.88 | 4.48 | 5.32 |
| Speech rate variation (syll/s) | 1.26 | 0.74 | 1.35 | 0.62 | 0.42 |
| Stimulus duration (s) | 13.64 | 12.76 | 13.97 | 12.71 | 12.10 |

ses, the median was calculated per stimulus. Only high pitch accents (H*, !H*, and ^H* accent tones; see Section 7.3 in Chapter 7 for more information on the DIMA annotation) were included in the analyses, which amounts to 4,573 pitch peaks and the exclusion of 669 low tones (L*).

The prominence ratio was measured as the number of prominent syllables (independent of prominence level or association with a pitch accent) divided by the total number of syllables in a phrase and normalized by phrase duration. The frequency of occurrence of the different prominence levels (weak, strong, emphatic; see Section 7.3) was measured as the number of such prominences within a phrase. The frequency was normalized by the duration of the stimuli (including pauses) to receive count per minute as a measuring unit. The prominences were only analyzed

for the long stimuli since there was more material and therefore more opportunity for prominent syllables.

Finally, speech rate and speech rate variability were measured in syllables per second. The speech rate was measured per phrase without pauses. The speech rate variability was only analyzed for the long stimuli initially used for pause- and breathing-related manipulations, and measured as the standard deviation of speech rate measurements across phrases.

10.3.2 Experiment design

This experiment takes the ratings of the previous two experiments and correlates them with the measurements of the unmodified stimuli.

There are different participant groups depending on stimuli and rated attributes, but they are identical with those in the previous chapters (see Table 8.4 in Section 8.3.5). This means, one group rated the short stimuli in terms of the charisma-adjacent attributes and the long stimuli in terms of charisma and familiarity. The second group then rated the long stimuli in terms of the charisma-adjacent attributes and the short stimuli in terms of charisma and familiarity.

Since only the unmodified stimuli were included (ORIG for the short stimuli, and MED_BR or MIX_BR for the long stimuli), there were a minimum of 50 responses for the long stimuli and the charisma-adjacent attributes, and around 200 ratings for each of the other stimulus-attribute combinations which were included in the mean ratings used for analysis. The higher number of ratings for the charisma-adjacent attributes and the short stimuli (200, versus the 50 with the long stimuli) comes from the fact that the original short stimuli were presented to the listeners once per attribute—meaning they were played four times for each participant—while there was no stimulus repetition with the long stimuli. Nonetheless, the repetitions were included in the current study as well.

10.3.3 Statistical analyses

This part of the investigation uses a combination of descriptive and inferential statistical methods. All figures were created with the `ggplot()` function of the `ggplot2` package (Wickham et al., 2014) in RStudio (RStudio Team, 2023). Since the sample size of the data is so small—only ten speakers, and only five per male/female/England/North America—the inferential statistics are performed with the full set of data, and observations regarding possible effects of speaker gender and speaker origin are reported based on descriptive statistics.

The first step of the analysis is descriptive. This step checks where the measurements of the stimuli fall within the distribution of measurements of all phrases in the full corpus. This is done with scatter plots of the stimulus measurements within

box plots of the measurements from the full data. This way it becomes obvious if the stimuli are actually representative of the speaker's general way of speaking (when the points fall within the box, meaning the 50 percent of data around the median), or if the stimuli are irregular for a speaker (when the points fall on the whiskers, which represent the top and bottom 25 percent, or outliers). In the latter case, the results should only be interpreted as relevant for the current sample, and more general results can be addressed in future studies. In these plots, the speakers are separated by gender, and arranged in a way that the English speakers are on the left and the North American speakers on the right, ordered within group alphabetically.

The second step of the analysis is inferential, in that the mean attribute ratings of the stimuli of each speaker were correlated with the mean measurement of the acoustic features using the `cor.test()` function included in base R. In the case of the long stimuli, a mean across the four phrases was used, unless specified otherwise. There are only ten data points (one for each speaker) and the sample size is therefore small, which is why correlations were calculated using Kendall's tau (τ), since this method also works with small sample sizes. The same method was used for the speakers and their vowel spaces in a previous study (Berger et al., 2023). The correlations were run separately for the short and long stimuli. The significance level for the correlations is set at $\alpha = .05$. However, some non-significant trends are also reported (meaning the p -value can be rounded to $.1$, meaning $p > .05$, but $< .15$).

The Kendall correlations cannot offer insights into possible differences between male and female speakers, as well as speakers from England or North America because that would mean splitting up the data sample even more and having an even smaller sample for which the Kendall method also seems to run into problems. But possible differences between the speaker groups can be investigated descriptively. In order to visualize this, a scatter plot and a line plot were combined representing the mean rating (including the standard deviation of the rating) for each speaker across all raters. The speakers are arranged according to the mean measurement of each specific acoustic feature in ascending order, with a regression line behind. That way, likely correlations come out visually. This method was also used by Niebuhr (2020) and Berger et al. (2023). It also means that speakers are arranged differently in the plot depending on the mean measurement of the respective acoustic feature.

In the case of the frequency of emphatic accents, results are only reported descriptively in the form of bar plots, in comparison to the full data. The frequency is also normalized to number per minute, so the results are relational, and not absolute numbers. In this case, the speakers are arranged by mean rating of the stimuli.

10.4 Results

10.4.1 Pitch

Looking at the results of the stimuli measurements in the context of the full corpus data (roughly 50 minutes of speech material) suggests that all pitch-related features have some stimuli that are outliers (see Figure 10.1). In most cases, the measurements of the short stimuli—the solid black dots—do fall within the 50 percent of data around the median, or just slightly outside. Generally, there seem to be more outliers with the long stimuli (unfilled points).

Since some of the stimuli measurements can be considered outliers, this also means that the perception results have to be seen as directly connected to the specific acoustic make-up of the stimuli. Generalization to the rest of the data (i.e., the speakers' voices in general) has to be done after further research in future studies.

All correlations of the pitch-related features are collected in Table 10.3. The measurements of the acoustic features are the means of the stimulus phrase in the case of the short stimuli, and the means across the four phrases for the long stimuli. In general, there were two significant results, and a few non-significant trends.

The only significant correlation with the responses to the short stimuli was between the *charismatic* ratings and the mean F0 of the stimuli, but there was a similar non-significant trend in the correlation between ratings and the median pitch as well. The correlation was negative, suggesting that the higher the mean F0 (and median pitch), the lower is the perceived charisma. This is also seen visually, especially for the mean pitch level (Figure 10.2A), where the downward trend from the lowest mean F0 on the left to the highest mean F0 on the right is fairly striking. One outlier is the speaker stimulus with the highest mean F0 (ZS) who is rated as similarly charismatic as the speakers with the lowest mean F0.

In Figure 10.2A and the correlation analysis, both male and female speakers were analyzed together. Unsurprisingly, the male speakers have lower mean (and median) F0 than the female speakers with the exception of speaker DH, who is male with a mean F0 in the “female” ranges of the sample. When the data set is split visually between male and female speakers, the negative correlation is clearer for the male speakers (right panel in Figure 10.2B). For the female speakers, the regression line in the background seems to suggest a slight positive correlation (left panel in Figure 10.2B). However, this is again influenced by speaker ZS with the highest pitch level. If the high F0 by ZS is left aside, the results for the female speakers also suggest a negative correlation between mean F0 and mean rating, mirroring the male speakers in the sample. In general, the charisma ratings for the female speakers were lower than those of the male speakers.

When splitting the data by speaker origin, the North American speakers also

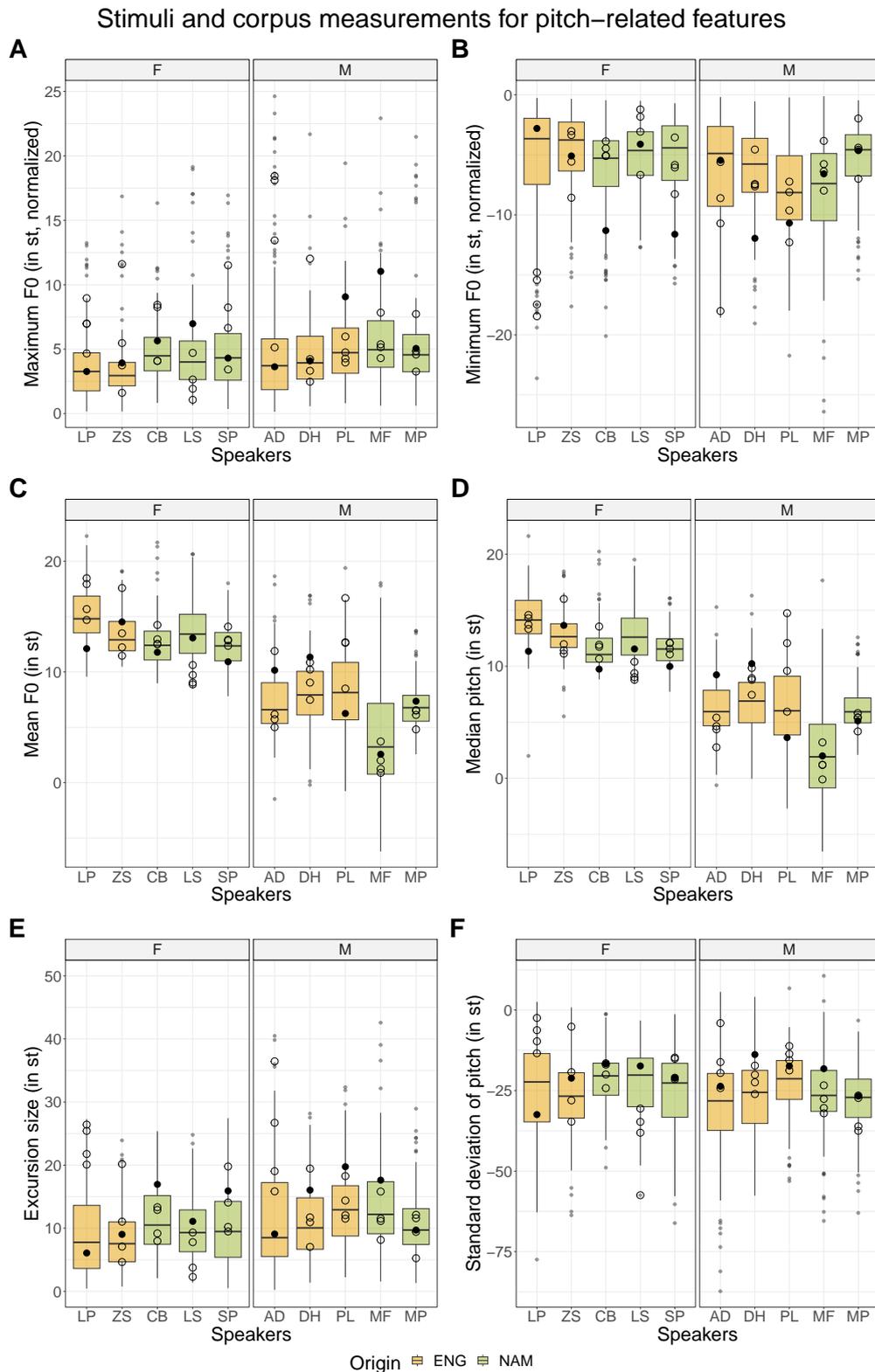


Figure 10.1: The measurements of the stimuli (short = solid point, long = unfilled point) situated within the measurements of the full annotated corpus (boxes, whiskers, and outliers). The six pitch-related features included in the investigation are presented: A) maximum F0, B) minimum F0, C) mean F0, D) median pitch, E) excursion size, and F) standard deviation of F0. The data are split by gender and colored by origin.

Table 10.3: The correlations of the pitch-related acoustic features (maximum, minimum, and mean F0, median pitch, excursion size, and pitch variability) with the mean attribute ratings (*charismatic* = CH, *authentic* = AU, *enthusiastic* = EN, *likable* = LI, *persuasive* = PE) for both short and long stimuli.

| Feature | Attr. | Short stimuli | | Long stimuli | | |
|-------------------|-------|---------------|-----|--------------|-------|-----|
| | | τ | p | τ | p | |
| Maximum F0 | CH | 0.3 | .24 | -0.09 | .72 | |
| | AU | -0.07 | .79 | -0.26 | .33 | |
| | EN | -0.31 | .21 | -0.26 | .31 | |
| | LI | 0.13 | .59 | -0.07 | .79 | |
| | PE | 0.02 | 1 | -0.02 | .93 | |
| Minimum F0 | CH | -0.2 | .42 | -0.14 | .59 | |
| | AU | -0.2 | .42 | 0 | 1 | |
| | EN | -0.04 | .86 | -0.31 | .23 | |
| | LI | -0.22 | .37 | -0.3 | .24 | |
| | PE | -0.24 | .38 | -0.02 | .93 | |
| Mean F0 | CH | -0.52 | .04 | * | 0.14 | .59 |
| | AU | 0.34 | .18 | | -0.15 | .56 |
| | EN | -0.18 | .47 | | 0.45 | .08 |
| | LI | -0.09 | .72 | | 0.4 | .12 |
| | PE | -0.02 | 1 | | -0.17 | .51 |
| Median pitch | CH | -0.43 | .09 | | 0.18 | .47 |
| | AU | 0.3 | .24 | | -0.2 | .44 |
| | EN | -0.13 | .59 | | 0.5 | .05 |
| | LI | -0.13 | .59 | | 0.35 | .17 |
| | PE | 0.07 | .86 | | -0.17 | .51 |
| Excursion size | CH | 0.25 | .32 | | 0.05 | .86 |
| | AU | 0.11 | .65 | | 0.05 | .85 |
| | EN | -0.04 | .86 | | 0.17 | .52 |
| | LI | 0.31 | .21 | | 0.21 | .41 |
| | PE | 0.11 | .73 | | -0.17 | .51 |
| Pitch variability | CH | 0.02 | .93 | | 0.09 | .72 |
| | AU | 0.39 | .13 | | -0.05 | .85 |
| | EN | -0.18 | .47 | | 0.31 | .23 |
| | LI | 0.22 | .37 | | 0.44 | .08 |
| | PE | 0.07 | .86 | | -0.27 | .3 |

Signif. codes: * .05, . .1

show a negative correlation, while this is not the case for the speakers from England (Figure 10.2C). Rather, the ratings of the English speakers suggest a dip in the ratings, with low mean pitch level on the one hand and high pitch level on the other hand being rated as more charismatic than medium pitch level. However, this could also indicate that male speakers are rated as more charismatic with lower pitch (the left three speakers are male), and the female speakers with higher pitch.

For the short stimuli, there additionally was a non-significant trend with the correlation between the *authentic* ratings and the pitch variability (see Table 10.3). This correlation was positive which suggests that the higher the pitch variability (i.e., standard deviation of pitch), the higher the perceived authenticity. This is visually also the case when the data set is split by speaker gender (Figure 10.3B) and speaker origin (Figure 10.3C), even though there are exceptions in the ratings, especially for the male speakers and the North American speakers. The main exception for both

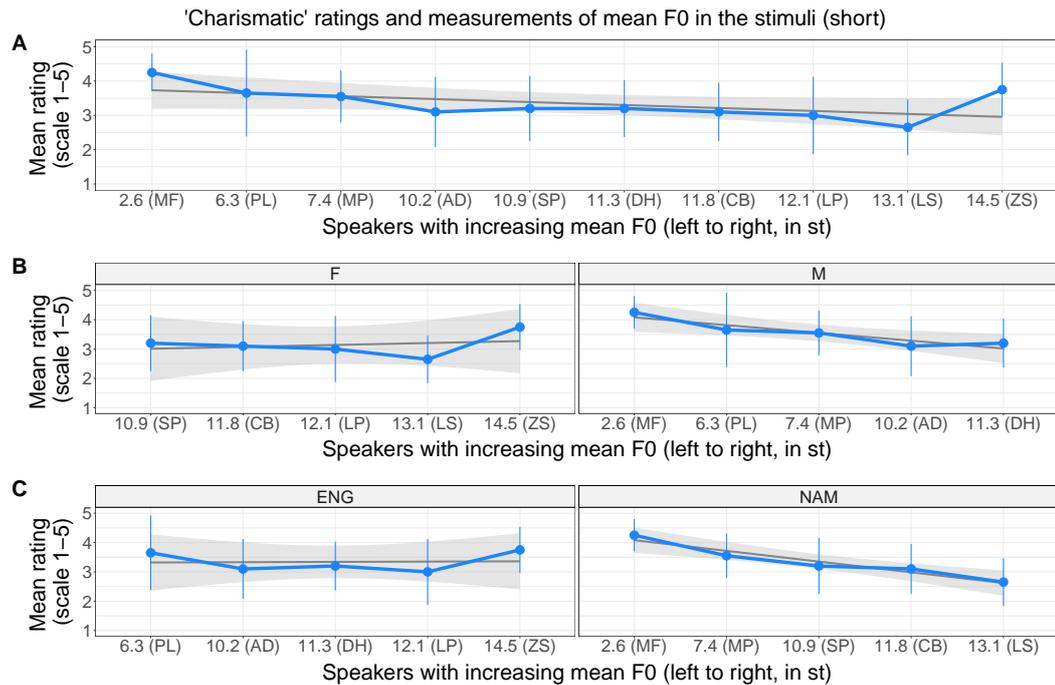


Figure 10.2: The results of the correlation of *charismatic* ratings and the mean F0 of the stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (here: mean F0) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating. The grey error bar shows if there was a stronger linear correlation (small grey area) or a weaker linear correlation (wider grey area).

groups is speaker MF, who was rated lower in terms of authenticity than the other (male, North American) speakers.

Another observation for the short stimuli is that the directions of the correlations of two pitch-related features—mean F0 as well as median pitch, which are inherently similar to each other—are the opposite between the charisma and authenticity ratings, irregardless of whether or not these correlations are significant. The correlations with the mean charismatic ratings were negative, and those with the authentic ratings were positive (see Table 10.3). This suggests that when the overall pitch level of a stimulus was higher, it was perceived as at the same time more charismatic and less authentic. This was reversed for the long stimuli.

There was one significant correlation for the long stimuli between the median pitch and the *enthusiastic* ratings. Note that the acoustic value is the mean of the median pitch across the four stimuli phrases per speaker. There was a similar non-significant trend for the *enthusiastic* as well as the *likable* ratings of the long stimuli and the mean F0. These correlations were positive, suggesting that the higher the median pitch (or mean F0), the more enthusiastic a speaker is perceived. The focus is on the significant correlation between median pitch and the *enthusiastic* ratings.

The visual inspection reveals the positive correlation, but also that there are out-

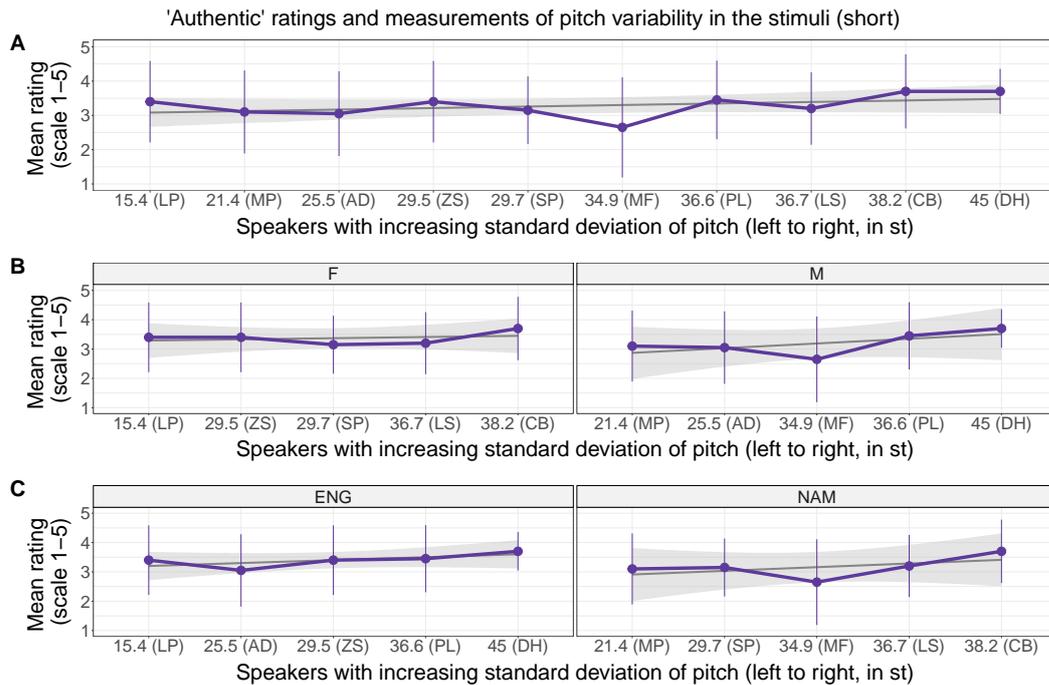


Figure 10.3: The results of the correlation of *authentic* ratings and the pitch variability (standard deviation of pitch in st) of the stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (here: pitch variability) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

liers, both negative and positive (Figure 10.4A). When splitting the data by speaker gender, the negative outlier becomes even clearer, as the male speaker with the second lowest median F0 had the by far lowest rating. In general the sample with the male speakers shows the positive correlation, but also a lot of variation mainly due to the negative outlier (see right panel in Figure 10.4B). On the other hand, the ratings for the female speakers (when arranged in ascending order of median pitch in the stimuli) follow a regression line fairly closely, albeit with a slight curve.

The outliers are equally present in the speakers from England (left panel in Figure 10.4C) which still suggest a positive correlation, but again with large variation. This is again different for the speakers from North America. Here, all ratings are actually roughly the same, suggesting no effect of median pitch whatsoever. These patterns are the same for the mean F0 of the stimuli. The North American stimuli were rated fairly consistently as neutral (the 3 on the y axis), while the speakers from England (with the exception of AD) were overall higher rated. This mirrors findings in the previous two chapters which found that there were at least consistent trends that suggest that English speakers in the sample were rated as overall more enthusiastic than the North American speakers in the sample.

Similar to the non-significant trend between mean F0 and the *likable* ratings, there was also a non-significant trend in the correlation between the pitch variability

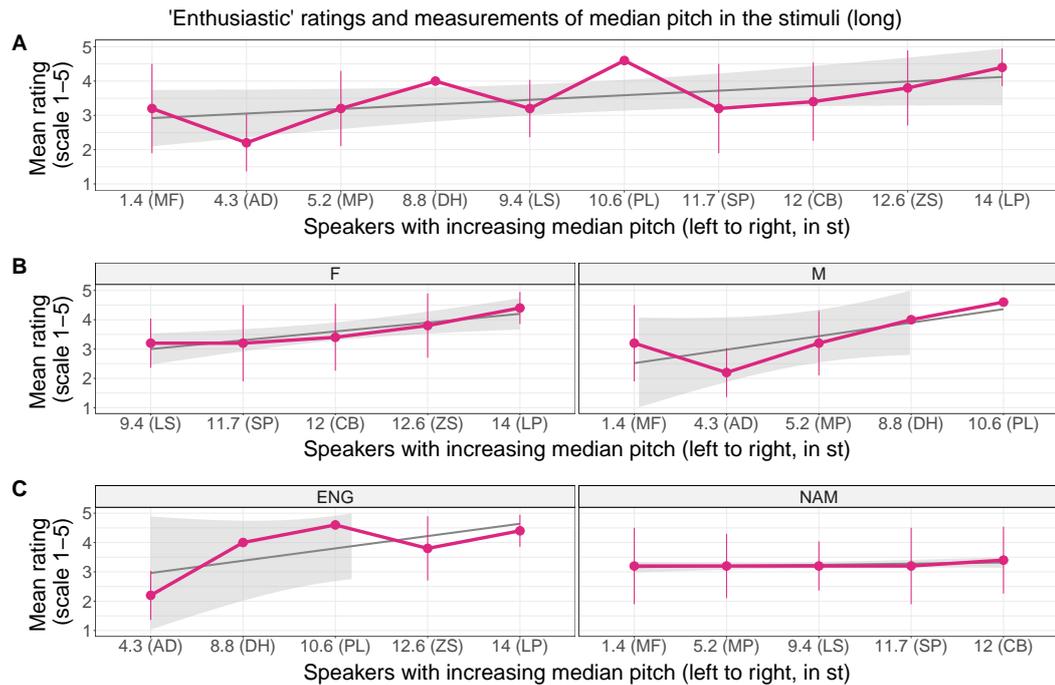


Figure 10.4: The results of the correlation of *enthusiastic* ratings and the median pitch of the long stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (here: median pitch) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating. The grey error bar shows if there was a stronger linear correlation (small grey area) or a weaker linear correlation (wider grey area). They are cut off because the variation could hypothetically go beyond the values of the rating scale.

measurements of the long stimuli and the likability ratings (see Table 10.3). This correlation was again positive, suggesting that more variable pitch was perceived as more likable (see also Figure 10.5A). The positive correlation is (visually) especially prominent for male speakers, and speakers from England. There were no obvious correlation patterns for female speakers and North American speakers (see Figures 10.5B and Figure 10.5C).

Finally, there is an observation for the charisma ratings and the mean F0 of the long stimuli. There was no significant correlation if all speakers were combined ($p = .59$). However, patterns emerge when the correlations are visualized and split by gender (see Figure 10.6). The ratings and mean F0 seem to have a slight correlation for the female speakers. This tentative correlation seems to be positive, which may also be present for the short stimuli (see above). Equally, the rating differences were small, adding to the tentative nature of the observation. For the male speakers, the pattern looks different. The visualization suggests a curve, with the highest charisma ratings for the stimuli with either the lowest or the highest mean F0. Those two stimuli actually received similarly high charisma ratings. The lowest rated stimulus is in the middle of the sample, though the mean F0 difference to the stimulus before is only slight. This pattern may also suggest a difference in speaker

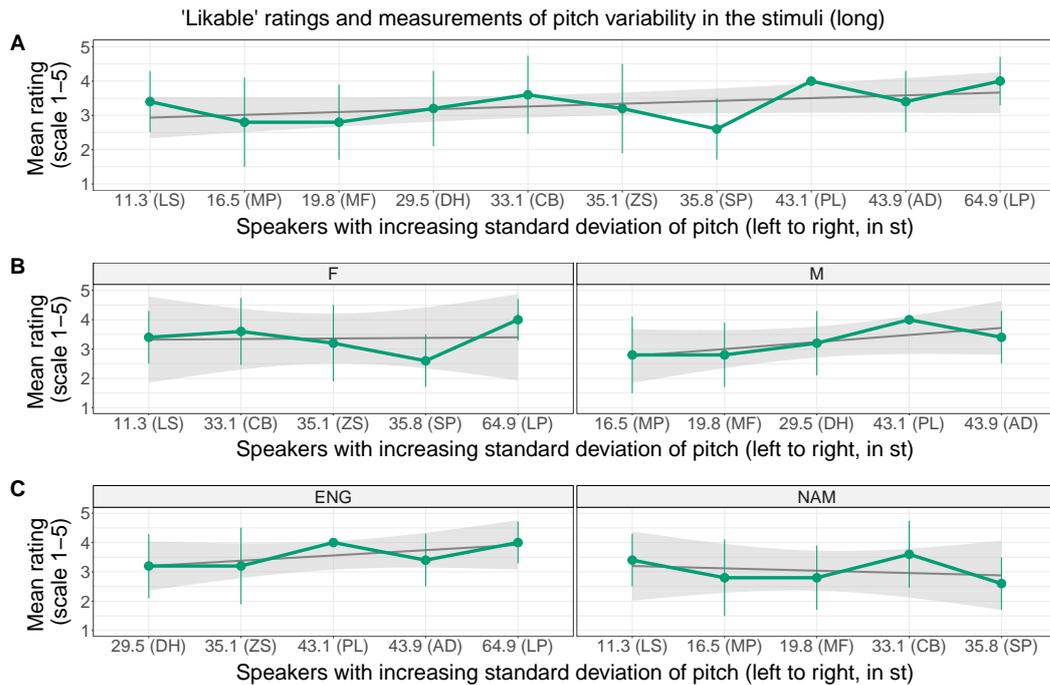


Figure 10.5: The results of the correlation of *likable* ratings and the pitch variability (standard deviation of pitch in st) of the long stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (here: pitch variability) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

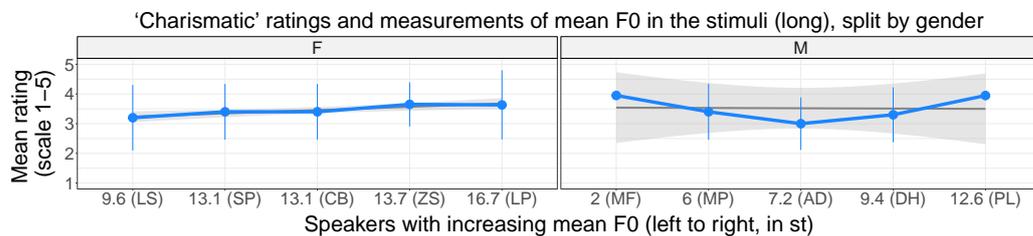


Figure 10.6: The results of the correlation of *charismatic* ratings and the mean F0 of the long stimuli, split by gender. The speakers and the respective mean value of the acoustic feature (here: mean F0) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating. The grey error bar shows if there was a stronger linear correlation (small grey area) or a weaker linear correlation (wider grey area).

origin, since the first two speakers have a downward/negative trajectory (and are the North American speakers), and AD, the lowest rated speaker, and the next two speakers, DH and PL, are from England and show a positive trend.

10.4.2 Intonation

Three intonation-related features were included in the analyses: the pitch peak timing, the prominence ratio, and the frequency of the different prominence levels.

Looking at the comparison between the stimulus measurements and the mea-

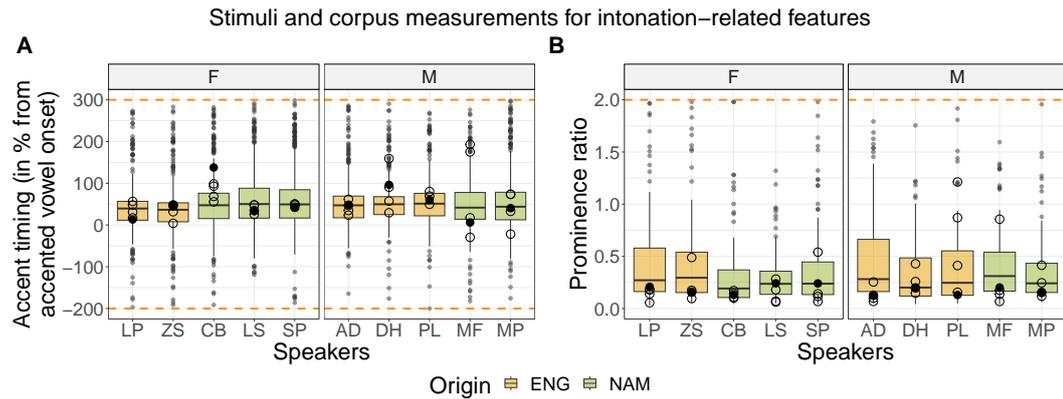


Figure 10.7: The measurements of the stimuli (short = solid point, long = unfilled point) situated within the measurements of the full annotated corpus (boxes, whiskers, and outliers). Two of the three intonation-related features included in the investigation are presented: A) peak timing (in % from accented vowel onset; note that the measurements of the stimuli are the median accent timing in each stimulus phrase), and B) prominence ratio. The data are split by gender and colored by origin. The dashed orange lines represent cut-off points of the data for the visualization.

measurements from the full amount of annotated speech material for pitch peak timing, the stimuli are mostly within the middle 50 percent of data (see Figure 10.7A). Note, though, that outliers above 300 percent from accented vowel onset and below -200 percent were excluded from the figure in order to make the figure more readable. However, peak timing of higher magnitudes (up to 600 percent after accented vowel onset) are not unrealistic, for example if the pitch peak associated with an accent occurs late in the following syllable. Nonetheless, the pitch peaks in the stimuli tend to occur within the accented vowel, with some exceptions (e.g., speaker MF). Also note that the stimuli measurements in the figure (and those that are used in the subsequent analyses) are based on the median accent timing per phrase, as this measure is less strongly affected by extreme timings.

As far as the prominence ratio is concerned, the majority of phrases in the corpus range between around 35 and 65 percent with the medians around 50 percent, meaning that most phrases have about the same amount of prominent and non-prominent syllables in the sample of YouTubers (see Figure 10.7B). There are quite a few stimuli measurements that are outside the boxes, suggesting that the interpretations of the results should be connected mostly to the sample. But there are no absolute outliers in the stimuli—the measurements tend to be close to the boxes and with that still on the whiskers of the box plot, suggesting many similar measurements.

Finally, there is the comparison of the frequency of prominent syllables and the different prominence levels (weak, strong, and emphatic; see Section 7.3 for details of the annotation and definition of each level). The frequency of the prominences is not absolute, but in count per minute (i.e., the absolute frequency is divided by

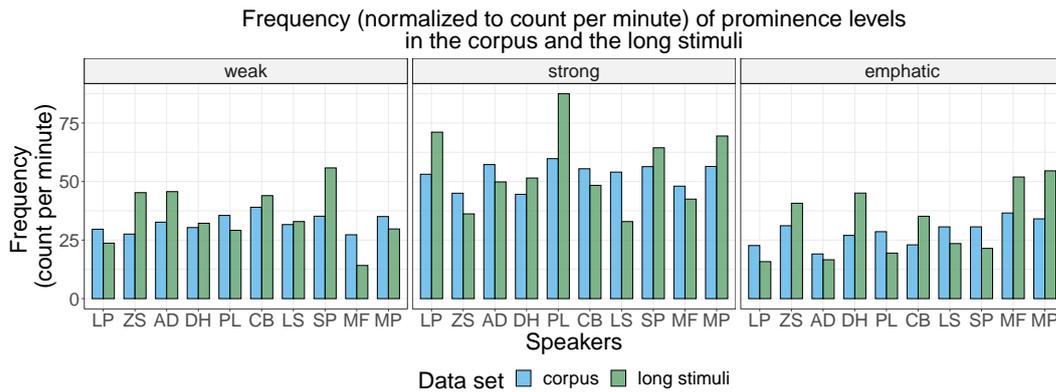


Figure 10.8: The frequency (normalized to count per minute) of the different prominence levels weak, strong, and emphatic in the corpus data as well as the long stimuli.

Table 10.4: The correlations of the pitch peak timing and prominence ratio with the mean attribute ratings (*charismatic* = CH, *authentic* = AU, *enthusiastic* = EN, *likable* = LI, *persuasive* = PE) for both short and long stimuli.

| Feature | Attr. | Short stimuli | | Long stimuli | |
|-------------------|-------|---------------|-------|--------------|-----|
| | | τ | p | τ | p |
| Pitch peak timing | CH | -0.07 | .78 | -0.18 | .47 |
| | AU | 0.52 | .04 * | -0.1 | .7 |
| | EN | 0.27 | .28 | 0.31 | .23 |
| | LI | 0.09 | .72 | 0.4 | .12 |
| | PE | -0.11 | .73 | -0.12 | .64 |
| Prominence ratio | CH | -0.2 | .42 | 0.41 | .1 |
| | AU | -0.11 | .65 | -0.1 | .7 |
| | EN | -0.67 | .01 * | 0.26 | .31 |
| | LI | 0.04 | .86 | -0.26 | .31 |
| | PE | 0.24 | .38 | 0.32 | .22 |

Signif. codes: * .05, . 1

the amount of speech material in the corpus/stimuli, including pauses). In general, there are many differences between corpus data and stimuli data (see Figure 10.8). The suggestion of the visualization is that for many speakers, there are many more prominent syllables in the stimuli than might be expected from the full amount of speech material (especially striking, for example, for speaker PL and the strong prominences). That means that generalizations can only be tentative.

Table 10.4 shows all results of the median pitch peak timing and the prominence ratio from the correlation analyses with the ratings of the short and long stimuli.

The results reveal a significant correlation for the peak timing in the short stimuli with their *authentic* ratings. This correlation is positive, suggesting that overall later pitch accents are perceived as more authentic. This is also shown by the visualization in Figure 10.9A. When the data are split up further, for example by gender, the positive correlation becomes visually even clearer for the male speakers, but there is the suggestion of a slight curve in the data of the female speakers which might hint that either earlier pitch peaks or very late pitch peaks could be perceived as

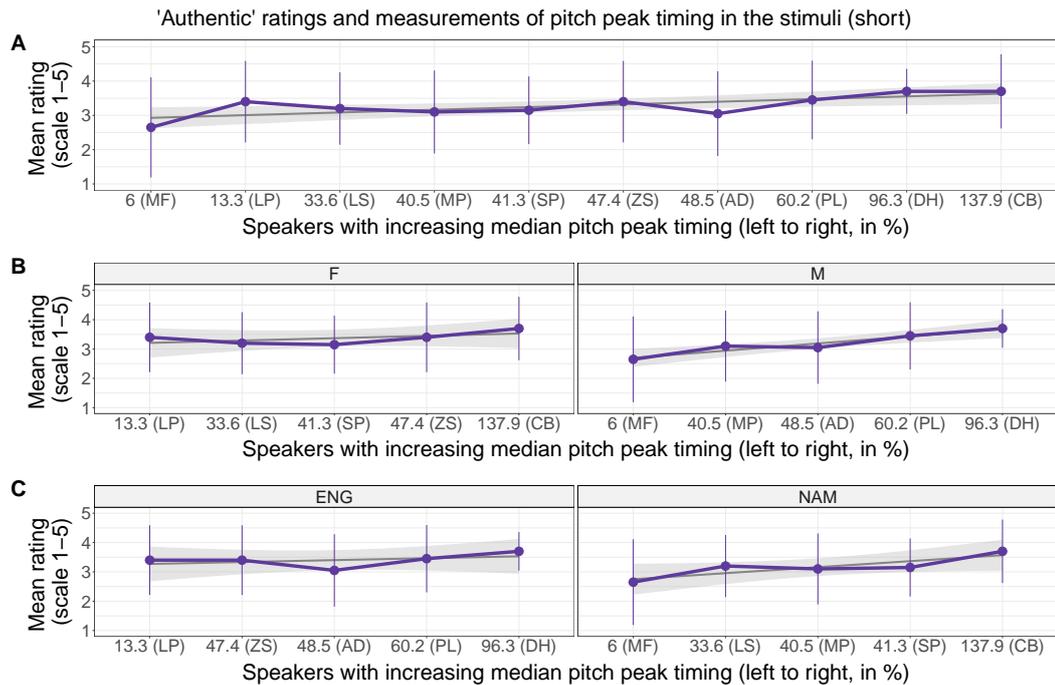


Figure 10.9: The results of the correlation of *authentic* ratings and the median pitch peak timing (in percent from accented vowel onset) of the short stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (here: median pitch peak timing) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

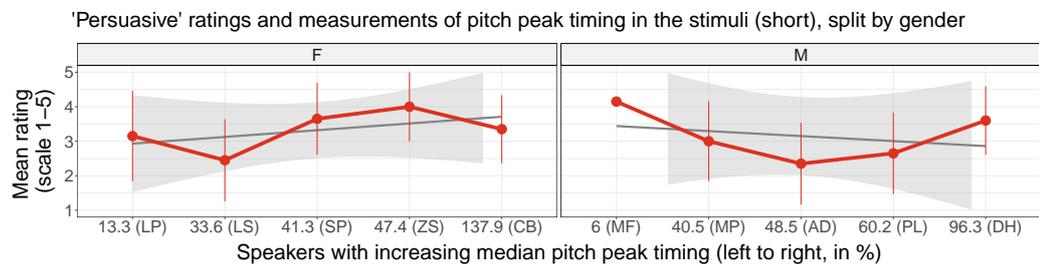


Figure 10.10: The results of the correlation of *persuasive* ratings and the median pitch peak timing (in percent from accented vowel onset) of the short stimuli, split by speaker gender. The speakers and the respective mean value of the acoustic feature (here: median pitch peak timing) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

more authentic depending on the speaker (Figure 10.9B). This may be similar when the data set is split by origin, where the positive correlation is clear for the North American speakers, but there is the suggestion of a rating dip for speakers from England (Figure 10.9C). This rating dip is also clearly visible for the accent timing and the *persuasive* ratings of the short stimuli, this time for the male speakers (see Figure 10.10) which was not a significant correlation or non-significant trend for the whole sample. There is no such correlation or trend with the long stimuli.

For the short stimuli, there was also a significant correlation between the promi-

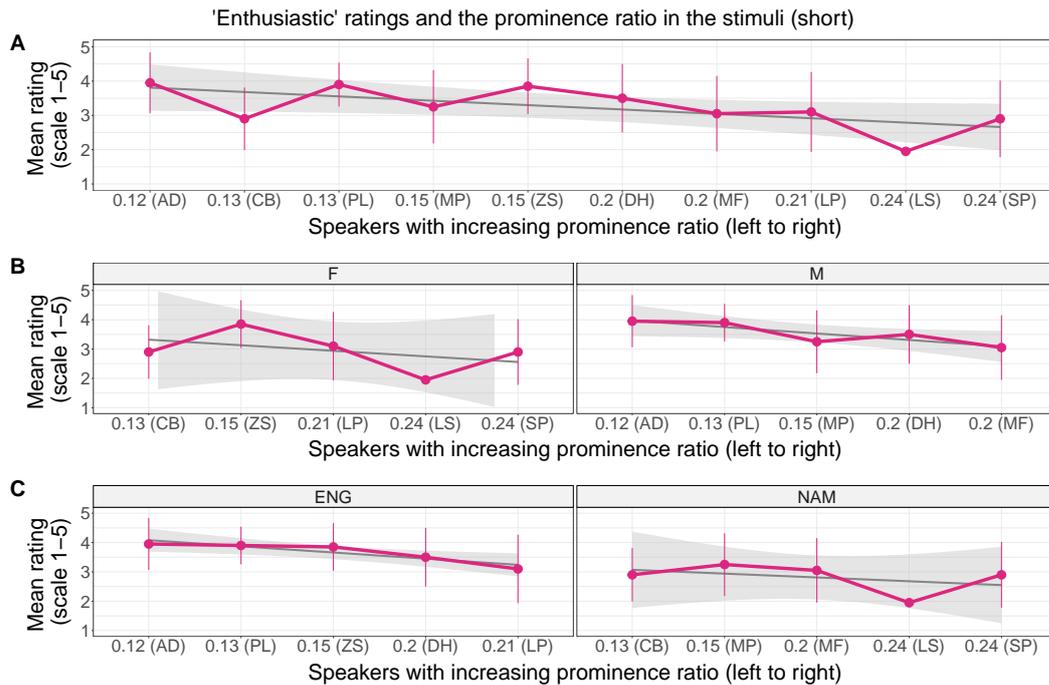


Figure 10.11: The results of the correlation of *enthusiastic* ratings and the prominence ratio of the short stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (prominence ratio) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

nence ratio and the *enthusiastic* ratings. This correlation was negative (see Table 10.4 and Figure 10.11A), suggesting that a larger amount of prominent syllables in a phrase was perceived as less enthusiastic which seems to be counter-intuitive when enthusiasm, expressiveness, and variation are connected. The correlation had positive and negative outliers, so there was variation in the correlation. Visually, the correlations were stronger for the male speakers and speakers from England, though also female speakers and North American speakers showed negative correlations (but with more variation; see Figures 10.11B and 10.11C).

For the long stimuli, there were two non-significant trends with peak timing and the prominence ratio. There was a non-significant trend with the long stimuli between peak timing and the *likable* ratings. This correlation was positive, which suggests that stimuli with a generally later pitch accent timing tended to be perceived as more likable. If the directions of this correlation and the others are considered (Table 10.4), and only looking at the direction without their magnitude, it seems like the *enthusiastic* and *likable* ratings pattern together with positive correlations, as opposed to the *authentic*, *persuasive*, and *charismatic* ratings, which—while the correlations are not significant—have a negative direction. The visuals for the *likable* correlation shows that there are positive and negative outliers of the patterns (Figure 10.12), and splitting the data set by gender or origin did not change the positive direction of the correlation.

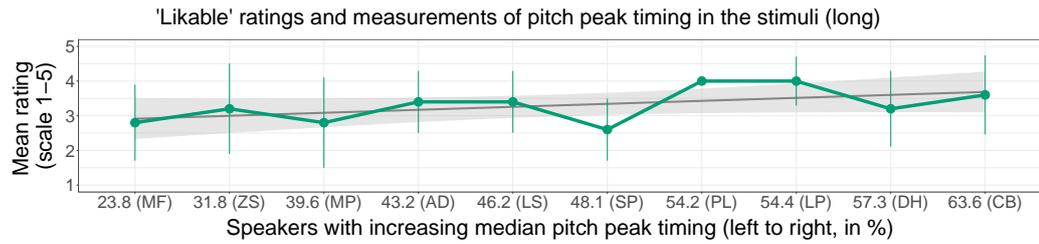


Figure 10.12: The results of the correlation of *likable* ratings and the median peak timing (in percent from accented vowel onset) of the long stimuli for all speakers. The speakers and the respective mean value of the acoustic feature (here: median peak timing) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

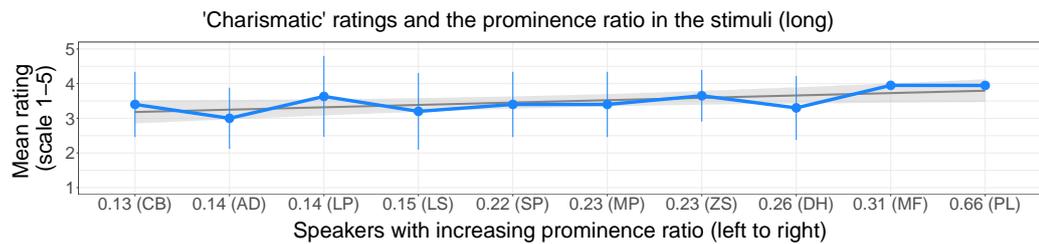


Figure 10.13: The results of the correlation of *charismatic* ratings and the mean prominence ratio of the long stimuli for all speakers. The speakers and the respective mean value of the acoustic feature (here: prominence ratio) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

Additionally, there was a non-significant trend for the correlation between the *charismatic* ratings of the long stimuli and the prominence ratio. This was a positive correlation (see Table 10.4 and Figure 10.13) which suggests that stimuli with more frequent prominent syllables tended to be perceived as more charismatic. Splitting up the data by gender or origin did not change the pattern, so only the correlation for all speakers combined is displayed.

Finally, there were some results for the frequency of prominences in the long stimuli. Table 10.5 shows all correlation results of the normalized frequency of prominent syllables of each prominence level (weak, strong, emphatic) with the mean rating response for the long stimuli. There were no significant correlations or trends in the data with strong prominences. In the data with the weak prominences, there was a non-significant trend, though, in the correlation with the *charismatic* ratings. This correlation was negative, suggesting that the more weak prominences there were in a stimulus, the less charismatic the speaker's voice behind the stimulus tended to be rated. Additionally, there was a significant correlation between the frequency of emphatic prominences in a stimulus and the *likable* ratings. This correlation was also negative, this time suggesting that more emphatic accents in a stimulus lead to the speaker being perceived as less likable, though there is also variation. Splitting up the data visually by gender or origin did not reveal addi-

Table 10.5: The correlations of the frequency of the different prominence levels (weak, strong, emphatic) with the mean attribute ratings (*charismatic* = CH, *authentic* = AU, *enthusiastic* = EN, *likable* = LI, *persuasive* = PE) for long stimuli.

| Feature | Attr. | Long stimuli | |
|----------------------|-------|--------------|-----|
| | | τ | p |
| Weak prominences | CH | -0.46 | .07 |
| | AU | 0 | 1 |
| | EN | -0.31 | .23 |
| | LI | -0.16 | .52 |
| | PE | -0.17 | .51 |
| Strong prominences | CH | 0.23 | .36 |
| | AU | 0.1 | .7 |
| | EN | 0.36 | .16 |
| | LI | 0.12 | .65 |
| | PE | -0.32 | .22 |
| Emphatic prominences | CH | 0.14 | .59 |
| | AU | -0.26 | .33 |
| | EN | -0.07 | .78 |
| | LI | -0.54 | .04 |
| | PE | 0.27 | .3 |

Signif. codes: * 0.05, . 0.1

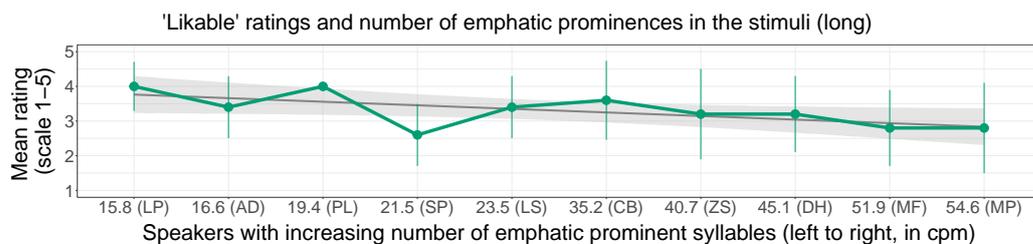


Figure 10.14: The results of the correlation of *likable* ratings and the frequency of emphatic accents (cpm) of the long stimuli for all speakers. The speakers and the respective mean value of the acoustic feature (here: frequency of emphatic accents) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

tional information, so all speakers are displayed together in Figure 10.14 for the correlation between *likable* ratings and emphatic prominences.

10.4.3 Duration and tempo

The final group of acoustic features included in this part of the analysis are four features that are connected to duration and tempo: phrase duration, speech rate, variation of speech rate, and the stimulus duration. Stimulus duration and the variation of speech rate are only investigated for the long stimuli, as they a) have (unlike the short stimuli) different phrase and stimulus durations, and b) consist of four phrases that allow for calculating the standard deviation of speech rate between the different phrases, which is the measure used for the speech rate variation.

When placing the measurements of the phrase duration and speech rate in the context of the rest of the speech material from the corpus, the measurements for the short stimuli mostly fall within the middle 50 percent of the corpus data, as

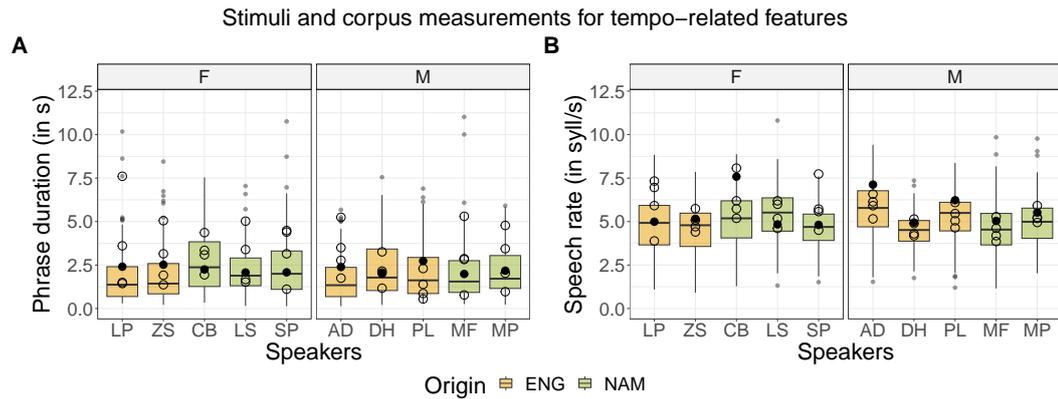


Figure 10.15: The measurements of the stimuli (short = solid point, long = unfilled point) situated within the measurements of the full annotated corpus (boxes, whiskers, and outliers). Two of the three tempo-related features included in the investigation are presented: A) phrase duration (in s), and B) speech rate (in syll/s). The data are split by gender and colored by origin.

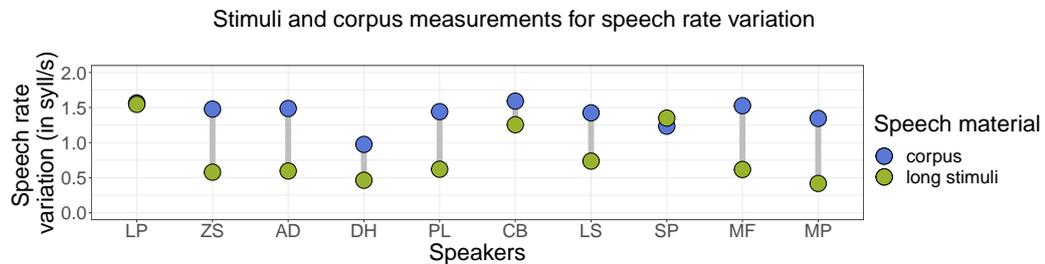


Figure 10.16: The variation of speech rate (standard deviation across phrases, in syll/s) in the long stimuli and in the full corpus data.

do those for the long stimuli, but there are a few outliers, but much fewer than with some of the other features in the previous sections (see Figure 10.15). That means it is likely possible to make tentative generalizations from the results of the investigation to the full data. As far as the variation of speech rate is concerned, there are some speakers—LP in particular, but also SP and CB—whose standard deviations line up well with those calculated across the entire corpus of speech material, but for the others, the standard deviation is much smaller in the long stimuli than the full data (see Figure 10.16). That means that generalization might not be possible overall. The stimulus duration cannot be situated in the greater context as it is directly connected to the stimuli and cannot be measured in the corpus.

The Kendall correlations overall revealed no significant correlations between the mean attribute ratings and the acoustic features of the stimuli (see Table 10.6). There were a few non-significant trends, though. The speech rate variation and the stimulus duration (both of the long stimuli) revealed no inferential results at all (see Table 10.7).

With the short stimuli, there were two non-significant trends: for both phrase

Table 10.6: The correlations of phrase duration and speech rate with the mean attribute ratings (*charismatic* = CH, *authentic* = AU, *enthusiastic* = EN, *likable* = LI, *persuasive* = PE) for both short and long stimuli.)

| Feature | Attr. | Short stimuli | | Long stimuli | |
|-----------------|-------|---------------|-----|--------------|-----|
| | | τ | p | τ | p |
| Phrase duration | CH | -0.02 | .93 | -0.14 | .59 |
| | AU | 0.2 | .43 | 0 | 1 |
| | EN | 0.4 | .11 | -0.36 | .16 |
| | LI | -0.22 | .37 | 0.07 | .79 |
| | PE | -0.24 | .38 | 0.02 | .93 |
| Speech rate | CH | 0.07 | .79 | -0.32 | .2 |
| | AU | 0.11 | .65 | 0.2 | .44 |
| | EN | 0.4 | .11 | -0.12 | .64 |
| | LI | -0.04 | .86 | 0.35 | .17 |
| | PE | -0.24 | .38 | -0.47 | .08 |

Signif. codes: * 0.05, . 0.1

Table 10.7: The correlations of speech rate variability and stimulus duration with the mean attribute ratings (*charismatic* = CH, *authentic* = AU, *enthusiastic* = EN, *likable* = LI, *persuasive* = PE) for long stimuli.)

| Feature | Attr. | Long stimuli | |
|-----------------------|-------|--------------|-----|
| | | τ | p |
| Speech rate variation | CH | 0.09 | .72 |
| | AU | 0 | 1 |
| | EN | 0.12 | .64 |
| | LI | 0.35 | .17 |
| | PE | 0.02 | .93 |
| Stimulus duration | CH | -0.18 | .47 |
| | AU | 0 | 1 |
| | EN | -0.26 | .31 |
| | LI | 0.11 | .65 |
| | PE | -0.12 | .64 |

Signif. codes: * .05, . 0.1

duration and speech rate, there were positive correlations with the *enthusiastic* ratings, and both correlations were exactly the same ($\tau = 0.4$, $p = .11$). For phrase duration, this suggests that longer phrases tended to be perceived as more enthusiastic in this sample. At the same time, the correlation with speech rate suggests that a higher speech rate tended to be perceived as more enthusiastic. Together, this may mean that longer phrases with more syllables in them (i.e., higher speech rate) evoke a sense of enthusiasm in tandem. Only the correlation for phrase duration is visualized here (Figure 10.17A). The visualization for the phrase duration with all speakers shows that there are outliers above and below the regression line, but when the data set is split up by gender (Figure 10.17B), the positive correlation becomes clearer for both female and male speakers. This is not the case when the data set is split by origin (Figure 10.17C).

The correlation between the *enthusiastic* ratings of the long stimuli and the phrase duration was not significant, but its direction was negative. It is therefore the opposite of the correlation in the short stimuli data set. Likewise, when splitting the by

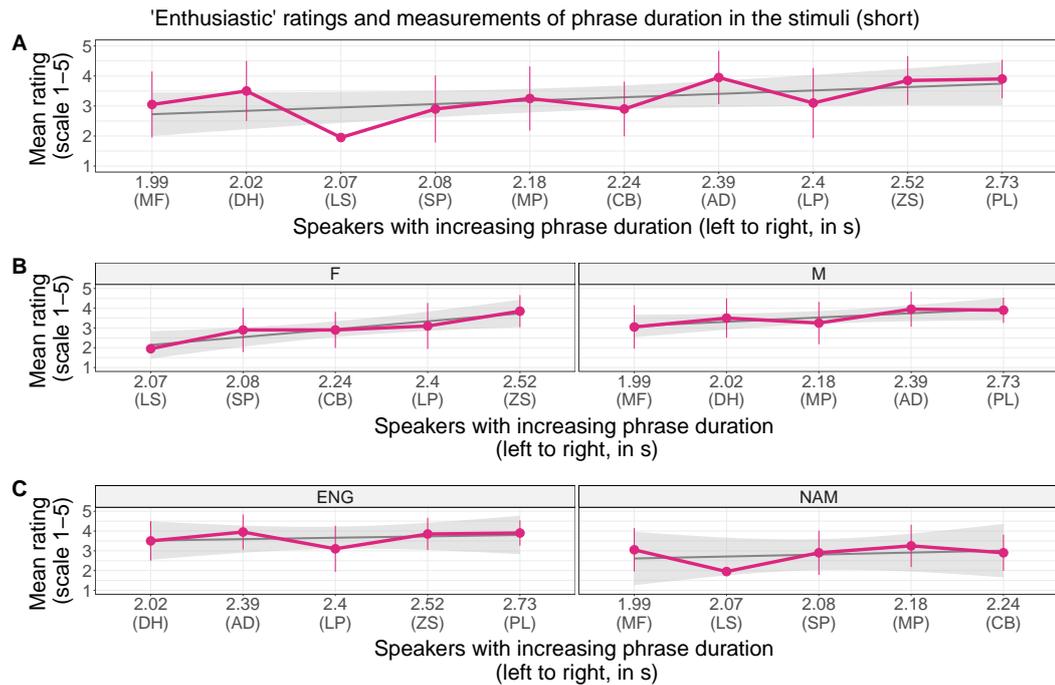


Figure 10.17: The results of the correlation of *enthusiastic* ratings and the phrase duration of the short stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (here: phrase duration) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

gender, the correlation seems to be negative and steep for the male speakers, with suggestions of a curved pattern for the female speakers. When the data set was split by origin, the lack of pattern for the North American speakers is especially striking: all speakers were rated roughly the same, and in general as less enthusiastic than almost all speakers from England. The visualization of this can be found in Figure K.1 in Appendix K.

There was also one non-significant trend with the ratings of the long stimuli, and this concerns the *persuasive* ratings, and the speech rate of the stimuli (note that this is the mean speech rate across all four phrases). This correlation was negative ($\tau = -0.47$, $p = .08$). It suggests that overall, speakers voices tend to be perceived as more persuasive when the speaker speaks with a lower speech rate. This is also visible when all speakers are considered together (Figure 10.18A), but this line is also slightly curvy.

When the data set is split by gender (Figure 10.18B), there seems to be a very steep negative correlation for the male speakers, but this arises from the extremely high ratings for the two speakers MF and DH with speech rates below five syllables per second, and the neutral and equal ratings for the other three male speakers with higher speech rates. For the female speakers, a negative correlation is subtle at best, rather there seems to be a curve in the ratings with the highest ratings for the speaker stimulus with a medium speech rate (when only looking at the speech rates

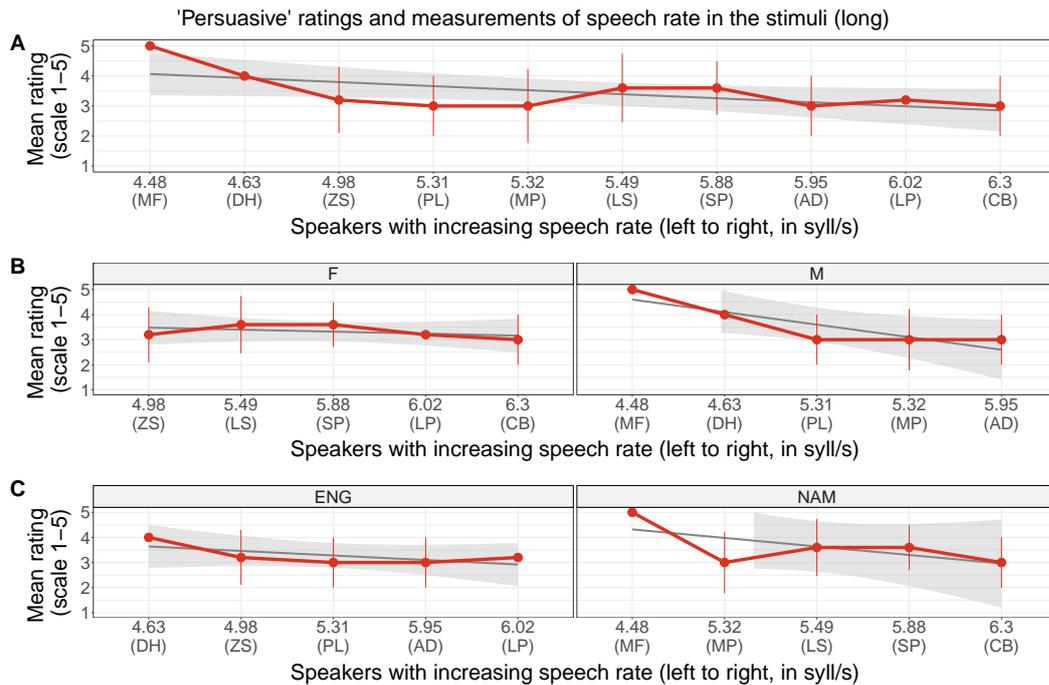


Figure 10.18: The results of the correlation of *persuasive* ratings and the speech rate of the long stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (here: speech rate) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

in the sample), and lower and higher speech rates receive lower persuasiveness ratings.

This is in a way similar when the data set is split by origin (Figure 10.18C). However, for the North American speakers, the regression line suggests a negative correlation, but there are such strong outliers that this would be an inconclusive result. Taking out the outlier with the highest rating and at the same time lowest speech rate (MF) would then end up with a similar curved pattern to the female speakers. There is also a curved pattern for the speakers from England, but compared to that of the female speakers (and the hypothetical one for the North American speakers), this pattern is inverted. That means that the lower and very high speech rates seem to receive higher *persuasive* ratings than the ones in the middle of the current sample, though in general all speakers are fairly close together.

While there were no other significant correlations or non-significant trends, visually splitting up the data by gender and origin revealed a couple of other observations. The curved patterns that were just described for the *persuasive* ratings also appeared for the *likable* ratings and their correlations with the stimulus duration of the long stimuli, both when the data set is split by gender (see Figure 10.19A) and by origin (see Figure 10.19B). The curves occur for the male speakers in the sample, as well as the speakers from England (which is likely connected, since three of the five speakers from England are male), and the pattern is similar to that of the

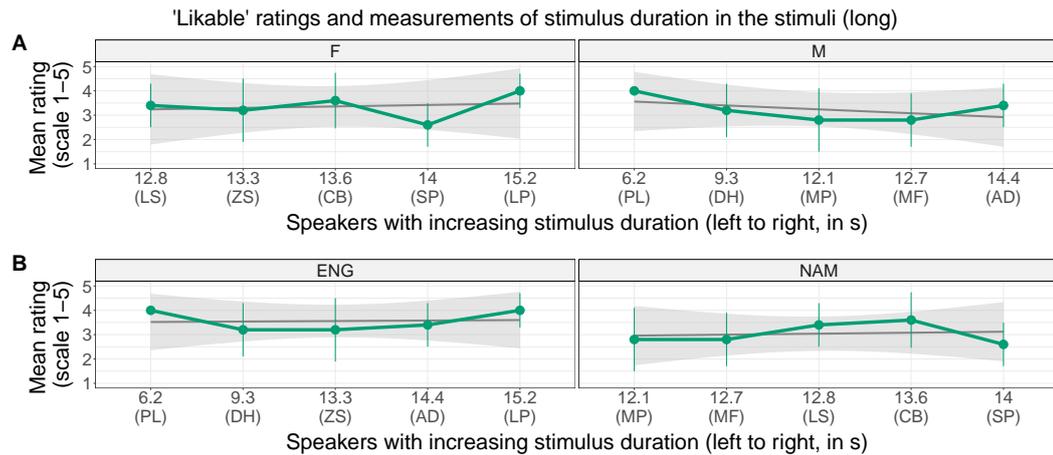


Figure 10.19: The results of the correlation of *likable* ratings and the total duration of the long stimuli, split by A) gender and B) origin. The speakers and the respective mean value of the acoustic feature (here: stimulus duration) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

English speakers with the *persuasive* ratings. That means that the speakers with the lowest and highest stimulus duration tend to be perceived as more likable, and the ones with medium stimulus duration receive lower ratings. For the female speakers, there seems to be no obvious pattern, but there might be the hint of an inverted curve (medium stimulus duration receiving higher ratings than shorter or higher stimulus duration) for North American speakers. The curved patterns for male and English speakers also occur with the *likable* ratings and the measurements of the mean phrase duration across the four phrases in the long stimuli (see Figure K.2 in Appendix K).

Finally, there was no significant correlation for the speech rate and the *authentic* ratings of the long stimuli ($\tau = 0.2$, $p = .44$). This lack of correlation is apparent when the ratings and measurements of all speakers are visualized (Figure 10.20A). When the data set is visually split by speaker gender (Figure 10.20B), there is still no obvious correlation for the female speakers. The male speakers show a clear positive correlation, suggesting that the higher the speech rate, the more authentic the speakers tended to be perceived.

10.5 Discussion

This chapter addresses the second research question in the project, namely whether or not speakers who are higher rated in the perception experiments use acoustic features that are connected to more charismatic speech according to previous research. The corresponding hypothesis (see Chapter 5 for both the research question and the main hypothesis) predicted that the experiment rating correlate with the ranking of the speakers/their stimuli that is created by the acoustic features

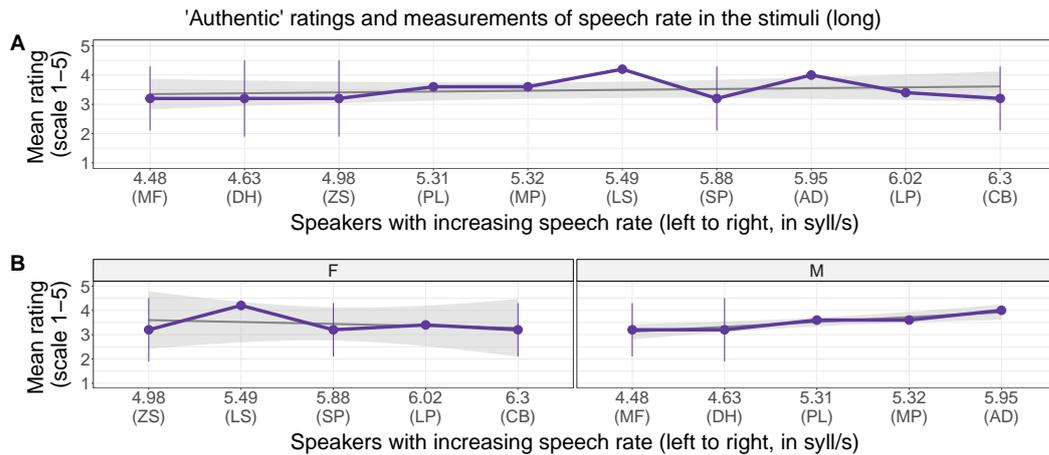


Figure 10.20: The results of the correlation of *authentic* ratings and the speech rate of the long stimuli, for A) all speakers, and split by B) gender. The speakers and the respective mean value of the acoustic feature (here: speech rate) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

connected to charismatic speech and the investigated related features. In general, this does not seem to be case in the current sample. There were a few significant results (though only one of them was directly with the *charismatic* rating), some trends, but mainly non-significant correlations.

Aside from the main hypothesis for the research question, several other hypotheses were put forward in this part of the project that directly connect to the acoustic features that were investigated (i.e., features related to pitch, intonation, and duration/tempo). These features are discussed in the sections below, after the results are first summarized. It is necessary to stress here again that this is only a small collection of acoustic features that are known or believed to be connected to charismatic speech. Including all relevant features would have a) been outside the scope of this investigation, and b) proven to be difficult to interpret since some of the features like intensity or voice quality are inherently affected by the audio quality, and on YouTube it is unknown what types of processing are implemented on the videos, and the audio signal is compressed for upload online (see Section 3.3.6).

10.5.1 Summary of the results

Table 10.8 (for the short stimuli) and Table 10.9 (for the long stimuli) provide an overview of the inferential results of this chapter.

The tables include information on the magnitude of the correlations. Different symbols are used for different magnitudes. A forward slash (“/”) signifies a negligible correlation ($\tau < 0.06$). Square brackets (“[]”) mark a weak correlation ($\tau \geq 0.06$ and < 0.26). Correlations of these two magnitudes are grayed out for easier readability. Parentheses (“()”) mark a moderate correlation ($\tau \geq 0.26$ and < 0.49), which

Table 10.8: Summary of the correlation results of all acoustic features with the the mean attribute ratings (*charismatic* = CH, *authentic* = AU, *enthusiastic* = EN, *likable* = LI, *persuasive* = PE) of the short stimuli. The acoustic features are maximum F0 (= MaxF0), minimum F0 (= MinF0), mean F0 (= MeanF0), median pitch (= MedP), excursion size (= Exc), pitch variability (= Var), pitch peak timing (= PkT), prominence ratio (= PrR), phrase duration (= PhrD), and speech rate (= SR).

| | MaxF0 | MinF0 | MeanF0 | MedP | Exc | Var | PkT | PrR | PhrD | SR |
|-----------|-------|-------|--------|------|-----|------|-----|-----|------|------|
| CH | (+) | [-] | -* | (-). | [+] | / | [-] | [-] | / | [+] |
| AU | [-] | [-] | (+) | (+) | [+] | (+). | +* | [-] | [+] | [+] |
| EN | (-) | / | [-] | [-] | / | [-] | (+) | -* | (+). | (+). |
| LI | [+] | [-] | [-] | [-] | (+) | [+] | [+] | / | [-] | / |
| PE | / | [-] | / | [+] | [+] | [+] | [-] | [+] | [-] | [-] |

Legend: * $p \leq .05$, . $p \leq .1$

is also where trends (p rounded to .1) can occur. Finally, if no modifier is used, it signifies a strong correlation ($\tau \geq 0.49$ and < 71), which are significant. These magnitude categories are based on suggestions from Wicklin (2023). In the tables, a plus sign (“+”) represents a positive correlation direction, and a minus sign (“-”) represents a negative correlation.

The overview shows that no attribute is encoded in the same way by all investigated features. Rather, each feature seems to evoke different positive or negative perceptions depending on the attribute. That means that the interplay of the prosodic features and the interaction of the different attributes is complex and warrants more targeted research.

10.5.2 Interpretation of the pitch-related findings

The first specific hypothesis of the study predicted that speakers with larger pitch variability and range, and higher minimum and maximum F0 would be perceived as more *charismatic*, *persuasive*, and *enthusiastic*, but less *authentic* and *likable* (H_{P31}). This was largely not supported by the data. This is mostly due to lack of significant correlations, but also when looking at the directions of the non-significant correlations, these directions are in the majority of cases opposite to what was expected. For example, the correlations with minimum and maximum F0 did not reveal significant results. Considering only the correlation directions, though, suggests that all attributes align with negative correlations (although maximum F0 in the short stimuli is an exception to this). This was not expected for charisma in particular. There was a non-significant trend in the correlation between the *likable* ratings and the pitch variability, which suggests that larger pitch variability was perceived as more likable. This was also opposite to the expectations of this project based on previous findings on charismatic speech, but would be in line with findings for German and the perception of likability (Weiss et al., 2021).

Table 10.9: Summary of the correlation results of all acoustic features with the the mean attribute ratings (*charismatic* = CH, *authentic* = AU, *enthusiastic* = EN, *likable* = LI, *persuasive* = PE) of the long stimuli. The acoustic features are maximum F0 (= MaxF0), minimum F0 (= MinF0), mean F0 (= MeanF0), median pitch (= MedP), excursion size (= Exc), pitch variability (= Var), pitch peak timing (= PkT), prominence ratio (= PrR), weak prominences (= WPr), strong prominences (=SPr), emphatic prominences (= EPr), phrase duration (= PhrD), stimulus duration (=StimD), speech rate (= SR), and speech rate variation (= SRV).

| | MaxF0 | MinF0 | MeanF0 | MedP | Exc | Var | PkT | PrR | WPr | SPr | EPr | PhrD | StimD | SR | SRV |
|----|-------|-------|--------|------|-----|-----|-----|-----|-----|-----|-----|------|-------|-----|-----|
| CH | [-] | [-] | [+] | [+] | / | [+] | [-] | (+) | (-) | [+] | [+] | [-] | [-] | (-) | [+] |
| AU | (-) | / | [-] | [-] | / | / | [-] | [-] | / | [+] | (-) | / | / | [+] | / |
| EN | (-) | (-) | (+) | +* | [+] | (+) | (+) | (+) | (-) | (+) | [-] | (-) | (-) | [-] | [+] |
| LI | [-] | (-) | (+) | (+) | [+] | (+) | (+) | (-) | [-] | [+] | -* | [+] | [+] | (+) | (+) |
| PE | / | / | [-] | [-] | [-] | (-) | [-] | (+) | [-] | (-) | (+) | / | [-] | (-) | / |

Legend: * $p \leq .05$, . $p \leq .1$

It could be that the lack of significant correlations comes from the extremely small data set with only ten speakers and therefore ten data points for correlation. The pooling of male and female speakers may have additionally influenced the analyses, though all measurements were made in semitones (and in the case of minimum and maximum F0 also normalized by the speaker's mean) and therefore are more comparable between genders than if the measurements were compared in Hz (see Graddol, 1986; Henton, 1995).

The next two hypotheses predicted there would be gender differences for mean and median pitch, namely that the male speakers would receive higher *charismatic*, *enthusiastic*, and *persuasive* ratings with higher mean and median pitch, but lower *authentic* and *likable* ratings (H_{P32}). The opposite was expected for female speakers (H_{P33}). This was generally also not the case. Not all correlations showed patterns, and the correlations could not be calculated, as mentioned above. However, there were a few patterns. For example, there was a negative correlation between the charisma ratings for the short stimuli and the mean F0 measurements. For the entire sample, this was significant, but it could not be interpreted since there were opposite expectations for male and female speakers. Visually, this negative correlation only seemed to be present for the male speakers where the opposite was expected. In this case it could be that the participants' understanding of charisma is connected more to associations with deeper voices and authority rather than higher voices. Equally, this result cannot be taken as the sole reason for a charismatic voice since it is established that mean F0 interacts with many of the other pitch-related features investigated here (e.g., higher and faster rising pitch accents, see Mixdorff et al., 2018; see also Section 3.2.1), so it can at most be part of the package. Future studies should also take other pitch-related features into account, like the pitch baseline, the difference between baseline and mean F0, and an adjusted pitch range (Mixdorff et al., 2018; Niebuhr et al., 2018a; Niebuhr and Skarnitzl, 2019).

There was another significant, this time positive correlation between median pitch and the *enthusiastic* ratings of the long stimuli. This correlation was positive for all speakers combined. The visual inspection of the data split by gender revealed that the correlations remained positive for both male and female speakers, but with slightly less variation for the female speakers. That means that within each speaker group, those speakers with higher median pitch were perceived as more enthusiastic. Therefore, at least for enthusiasm, this would argue against a gender difference, but rather suggest that the male and female speakers in this sample were perceived similarly based on this particular acoustic feature.

Additionally, the correlation between charismatic ratings of the long stimuli and the mean F0 measurements was not significant at all, but splitting the data by gender revealed a clear, but slight positive correlation for the female speakers, which goes against expectations. However, there was actually fairly little rating variation

between the different speakers, with finally may rather suggest no major influence of mean F0 on female speakers' perceived charisma. For the male speakers, there seems to be no linear correlation. The ratings are highest for speakers with the lowest and the highest mean F0, and dip for the stimulus with an intermediate pitch. This may suggest that the linear analysis methods used in this investigation may not be the most appropriate choice for all acoustic features and rating attributes, and the data should be revisited and re-examined with non-linear methods.

A final observation in the correlations of the pitch-related features concerns the (mostly non-significant) correlations mean F0 and median pitch with both the *charismatic* and *authentic* ratings. To begin with, the correlations in the short stimuli data were the opposite of what was expected. The expectation was that the correlations would be positive for charisma, but negative for authenticity (which was the case for the long stimuli). In the short stimuli sample, they were negative for charisma and positive for authenticity. As a second point, it was expected that authenticity and charisma as attributes would be coded differently in terms of pitch. This was found, since the correlations with the two attributes had differing directions in both data sets (short and long stimuli). This direction opposition between the two attributes was also found when investigating the vowel space size in terms of a mean range of the first formant for the same speaker sample and the same stimulus-rating combinations (Berger et al., 2023). Here, the correlations went in the expected direction for charisma, but at that point a larger vowel space was also assumed to be connected to higher authenticity ratings. However, this was not the case. The suggestion was made that on YouTube, charisma and authenticity may be different and perhaps clashing concepts that are both important for success, but are encoded acoustically in different ways. Charisma was thought to be more connected to vocal effort and therefore a larger vowel space, while authenticity was thought to be more connected to being relaxed and like in a friendly conversation, therefore encoded with less vocal effort to convey the speakers are approachable (see Berger et al., 2023).

In the current pitch-related analysis, this opposition may also be hinted at, but perhaps in a different way. Since—unlike with the vowel spaces—the correlation directions were negative for charisma, this could mean that the raters associate charisma more strongly with lower voices because of its connection to trustworthiness, and that authenticity (especially on YouTube) could be associated with excitement. Excited speech has been shown to have higher F0 levels, for example in horse race commentaries, where this speaking style may also be part of a professional routine that could use features like pitch level to keep a sense of spontaneity (Trouvain and Barry, 2000). This may also apply to YouTube and its largely unscripted, but still routinized performances. However, one also has to keep in mind that the calculated correlations take into account both male and female speakers

together as the sample is too small for separate calculations. The visual inspection of the data for mean F0 and charisma suggests that the negative correlation only applies to the male speakers (and the North American speakers if the data set is split by origin; see Figure 10.2). These results could therefore change with different data sets.

10.5.3 Interpretation of the intonation-related findings

Moving on to the discussion of the intonation-related results, the first hypothesis was related to pitch peak timing and predicted that speakers using later pitch peaks (relative to the accented vowel onset) would be perceived as more *charismatic*, *enthusiastic*, and *persuasive*, but less *authentic* and *likable* (H_{p34}). Only the correlations for the enthusiasm ratings seem to align, where there were positive (though non-significant) correlations, suggesting later accents may be perceived as more enthusiastic. This may suggest that enthusiasm and vocal effort could be connected as expected, as previous research proposes that later pitch accents have the same effect as higher pitch accents at making an accent sound more prominent, more listener-oriented, and require a longer pitch movement that can signal effort (Gussenhoven, 2002; Chen et al., 2002; Gussenhoven, 2016). The *charismatic* and *persuasive* ratings had negative, though non-significant, correlations in this sample; and the *authentic* and *likable* ratings had positive, in the case of authenticity even significant, correlations. These directions were—especially with the association of vocal effort and listener-orientedness—not expected. Rather, based on previous research (e.g., Biadys et al., 2007; Rosenberg and Hirschberg, 2009), later accents would be expected for charisma (and persuasiveness) because of the increased vocal effort. For authenticity (and likability), the opposite would make more sense as increased vocal effort could be interpreted as not being relaxed and approachable, which would be important on YouTube (Kyncl and Peyvan, 2017). Another way of looking at these results could be that the later accents and the association with being listener-oriented and clear, but also excited is more at the forefront for the *authentic* perception, which then would fit the positive correlation as well as the positive correlation for enthusiasm.

Additionally, it might be that the choice of the median accent timing across all pitch accents in a stimulus phrase is not the ideal measurement as it can cause extreme values to not be represented. It was initially chosen for this precise reason: to avoid the mean being influenced by extreme values. At the same time, the median still provided a way to use a central tendency measure in order to work with only one measurement per phrase. Future studies should therefore try out different measurements like the mean or the standard deviation—since it might be that variation of accent timing is important for the other attributes. Another option would

be to use the raw measurements, though then a different way of operationalizing the ratings would be needed, or a different set of stimuli that only contain one pitch accent. In general, pitch peak timing seems like a relevant topic for charismatic speech research that has not been conducted in detail so far. Since annotations are available for this corpus, future studies on YouTube data would be feasible to carry out.

There was another relevant visual result for the pitch peak timing in the short stimuli and their persuasiveness ratings, though the correlation was not significant. For the female speakers, there generally seems to be a positive trend as might be expected from previous research, though (if one outlier with the lowest median timing is left aside) the pattern may more resemble a curve with lowest ratings at the edges. For the male speakers, there was no linear correlation, but the figure shows a clear curve in the ratings: the speaker stimuli with the lowest and the highest pitch accent timing received the highest ratings, while the medium accent timing received the lowest rating (which would be the opposite to the female speakers if the pattern is interpreted as a curve). This could mean that especially for male speakers the extremes could be preferred—early accents or late accents—while accents that are centered in the accented vowel are perceived less positively. This could also hint at a negative influence of plateau contours. The label for downstepped high accent tones (!H*) can represent two peak types in the DIMA system: an actual pitch peak that is simply lower than the previous high tone, or a plateau-like and downstepped pitch accent. The latter one is always annotated at 50 percent of the vowel in the DIMA system (Kügler and Baumann, 2019a; see also Section 7.3). This needs further investigation, especially to find out to what degree the inclusion of these accents may have affected the measurements of the peak timing. In order to do this, the plateau-like and peak-like !H* accent tones need to be separated in an additional annotation pass. Additionally, and in relation to the !H* tones, pitch peaks (or accent tones) that are realized in the center of a vowel seem to be perceived as most persuasive for female speakers, while this vowel-medial peak timing seems to be rated least persuasive for male speakers. Previous research suggested that !H* is often connected to lists and rhetorics (Biadsky et al., 2007), which might be relevant for persuasion (at least for female speakers), but could also be perceived as boring in the context of YouTube (which may connect to the preference of earlier and later pitch peaks with male speakers).

The second hypothesis in this intonation-related group of features assumed that speaker stimuli with a higher ratio of prominent syllables are perceived as more *charismatic*, *enthusiastic*, and *persuasive*, but less *authentic* and *likable* (H_{P35}). With the long stimuli, charisma had a positive correlation (more prominent syllables tended to be perceived as more charismatic, as was expected) which was a non-significant trend. Similarly, the correlations with the *enthusiastic* and *persuasive* ratings were

positive for the long stimuli, and the *authentic* and *likable* correlations were negative (as predicted, though all correlations were not significant). Therefore, the long stimuli tended to be perceived as predicted. For the prominence ratio correlations in the long stimuli data, the difference in correlation direction then also suggests (similar to findings from mean F0 and median pitch) that charisma and authenticity may be connected to different prosodic productions, which again can be connected to vocal effort—more prominent syllables require higher or later pitch or longer segments which in turn means that a speaker chooses to put in effort to make their point come across (Gussenhoven, 2002). This is helpful for charisma, but many prominent syllables may end up sounding too performative and unnatural to be perceived as completely authentic.

Aside from the non-significant trend in the correlation with the *charismatic* ratings, the prominence ratio resulted in one significant correlation with the *enthusiastic* ratings of the short stimuli. The correlation was negative, which is against expectations. The opposite was the case for the long stimuli where—in line with expectations—the correlation was positive (though not significant), meaning that there were more prominent syllables compared to non-prominent syllables, which shows vocal effort, as discussed above, but might also prevent the utterance from becoming monotonous which would be negative for portraying enthusiasm. One possible reason for the difference in correlation direction between the two data sets is that the data sets use different measurement methods. For the short stimuli, the prominence ratio is simply the number of prominent syllables in an utterance divided by the total number of syllables in that same utterance. Since the short stimuli only consist of one utterance, this equals the measurement used for analyses. This is different for the long stimuli which consist of four phrases each. Here, the ratio was calculated per phrase and then the mean across the four phrases was used as the measurement for the analyses. Future studies should look into possible influences of the measurement method in more detail, as the results at this point are inconclusive. Additionally, it might also be that the difference comes from the participants of the experiment since the two data sets also go in hand with two groups of raters who may have by chance rated the stimuli similarly within their group, but differently to the participants of the other groups.

The final intonation-related hypothesis predicted that speaker stimuli with more emphatic accents would be perceived as more *charismatic*, *enthusiastic*, and *persuasive*, but less *authentic* and *likable* (H_{p36}). At the same time, correlations with the frequency of the other two prominence levels (weak and strong) was also investigated in an exploratory manner. There were no significant results for the strong prominences. There was a non-significant trend, though, in the correlation between weak prominences and the *charismatic* ratings. This correlation was negative, suggesting that the more weakly prominent syllables there are in a phrase, the less

charismatic the stimulus and the speaker are perceived. It suggests at the same time that the more non-weak prominences there are in an utterance, the higher the perceived charisma tends to be.

The directions of the correlations of the frequency of emphatic accents with the *charismatic* and *persuasive* ratings are positive which suggests that this may be tentative support for the hypothesis that more emphatic accents would lead to higher charisma and persuasiveness perceptions (though these correlations are not significant). The results of the *authentic* and especially the *likable* ratings are in line with the expectations though. The correlation with *likable* was even significant. These two correlations are negative, suggesting more emphatic accents are perceived as less likable (and authentic). It is likely that this result is due to the unnaturalness that can come into speech when too much of it is overly stressed, which happens easily with emphatic accents. While these types of accents (long vowels, long onset consonants, extremely high pitch or intensity, or a combination thereof) can create a rhythm, this can either be perceived as nice or annoying (see Beck, 2015; Hacker News, 2015), depending on the experience with the genre and perhaps also the duration of the presented material. Future studies will also look at differences between raters who follow vloggers on YouTube and regularly watch vlogs, and raters who do not. The data is available, but not included in the current project as it would be a different perspective on the topic (rater rather than speaker), and would have been outside the scope of this investigation.

It seems like the YouTubers in the sample used an extremely high amount of emphatic accents in their speech, as was expected from the features popular media outlets have reported for “YouTube voice” (Beck, 2015). The numbers used in the analysis are normalized to count per minute, since all stimuli were a) very short and b) of differing lengths. The amount of emphatic accents per minute ranges between 15.8 and 54.56 accents, which are both extremely high numbers (but especially the high end of the range). Note though that it is likely that not all phrases have as many emphatic accents, but this number is based on a stimulus with many emphatic prominences. In particular, it is based on a stimulus with a duration of 12.1 seconds (11.2 seconds without pauses). In this stimulus, the speaker (MP) produces eight emphatic prominences. This results in an extremely high number when normalized to one minute, but does not automatically mean that this degree of emphasis is continued throughout the video. In comparison, some speakers from the business world who were investigated for charismatic speech (Oprah Winfrey, Ginni Rometty, Meg Whitman, Mark Zuckerberg, Steve Jobs; see Novák-Tót, 2016; Novák-Tót et al., 2017; Niebuhr et al., 2020a) use significantly fewer emphatic accents. It seems like those investigations used absolute numbers in equal amounts of speech (20 minutes), so the number of emphatic accents mentioned in the investigations is divided by 20 minutes to also get a count per minute for comparison.

This suggests that the five speakers use between 1.35 and 6.55 emphatic accents per minute. This could suggest a few things. The speech material in the current study was annotated only by one person, so it might be that the annotator simply perceived more emphatic accents than others would. However, the difference between the two speaker groups seems so large that there is also a large margin of error, and it is unlikely that the annotation is tens of instances off. Rather it seems more likely that this is part of the YouTube (or at least vlogging) register or speaking style. A comparative analysis between the vloggers in this sample and CEOs from different international companies is in preparation (Berger, forthcoming). Additionally, this comparative analysis will also take the standard deviation of the raw emphatic prominence frequency per phrase into account to see if variation differs between YouTubers and CEOs.

In the context of YouTube, there might be an algorithm-related explanation for why a frequent use of emphatic accents in particular, but perhaps also prominent syllables in general, might be expected. According to Bishop (2018), videos where the closed captioning (CC) text work well with popular search engine terms are favored by the YouTube algorithm. Conversely, words with a high degree of articulatory reduction are often unrecognized and can therefore not be used as search terms. That is what Bishop calls “keyword stuffing”: to use frequently searched keywords in the spoken text which can be turned into CC text and then in turn can be used as search terms. Bishop suggests that “as closed captioning text is translated into written text, enunciating keywords ensures that they are readable text for a search engine” (2018, p. 80). And this is where emphatic accents like lengthened vowels and consonants come in: the lengthening already increases articulatory effort which also spreads to the rest of the word or syllable and strongly focuses the emphasized word. It therefore stands out from the rest of the speech around it which makes it easier for a CC system—but also for human listeners, of course—to understand. It might also be that using more emphatic accents (or at least more strong prominences) makes the speech clearer for non-native English speakers to understand. In order to maximize the chance of being understood, perhaps clear and emphasized speech is preferred in vlogs, as is suggested by the attention-paid-to-speech model (Labov, 1972; Lee, 2017). Additionally, it is unknown how much media training the speakers in the sample have received, but it is likely that most have taken some media training, especially when they are associated with talent agencies (see Bishop, 2018).

Future studies on charisma and emphatic accents should look into the type of words that receive emphatic prominence. It is likely that function words are not emphasized or only on rare occasions, but that it is rather the content words that are produced with emphatic accents (see, for example, Im et al., 2023), which would be relevant for turning speech into keywords for search engines. Equally, the rhythm

aspect of emphatic accents and charismatic speech should be investigated. It could be that the spacing between emphatic accents and prominences of other levels is important for a positive perception, or that the sequence or combination of different prominence levels is relevant. Similarly, the position of the accent within a phrase (towards the beginning, in the middle, at the end) might also be of relevance for future studies, which are being planned.

10.5.4 Interpretation of the tempo-related findings

The final category of acoustic features investigated here were tempo- and duration-related features. The first hypothesis in this category regards the phrase duration and predicted that speaker stimuli with shorter phrases would be perceived as more *charismatic, enthusiastic, persuasive, authentic* and *likable* (H_{P37}). That means that negative correlations would be expected for this feature. In terms of direction this can be found for likability and persuasiveness in the short stimuli, and charisma and enthusiasm with the long stimuli, though none of these correlations are significant. There were no correlations for charisma in the short stimuli as well as authenticity, likability, and persuasiveness in the long stimuli. The correlations with the *authentic* and *enthusiastic* ratings of the short stimuli went in the opposite direction than expected: they were positive suggesting longer phrases in a stimulus evoke higher ratings, which was not significant for authenticity, but a non-significant trend for enthusiasm. When splitting the data set by gender, the positive correlation becomes even clearer for both male and female speakers. However, the correlation seems stronger for the female speakers than the males, though this is likely mostly the case because the female speakers received overall more extreme ratings.

The phrase durations in the short stimuli range between 1.99 seconds and 2.73 seconds. This is overall higher in comparison to the five business speakers mentioned before (Mark Zuckerberg, Steve Jobs, Oprah Winfrey, Ginni Rometty, and Meg Whitman), whose mean phrase duration roughly ranges between 1.3 seconds and 1.93 seconds (see Novák-Tót, 2016; Niebuhr et al., 2020a). That suggests that the YouTubers in the sample produced longer phrases than speakers from the business context who are known for their varying degrees of charisma. For example, Steve Jobs was known as a very charismatic speakers, and his mean phrase duration in previous investigations was by far the shortest (1.25 seconds in speech directed at customers, and 1.39 seconds in speech directed at investors; see Niebuhr et al., 2020a). That suggests that there may be a difference between speakers on YouTube and in business. For enthusiasm, there was also a difference between the different data sets, though. The correlation in the data set with the long stimuli was not significant, but it was negative, as would be expected from previous research.

This is especially the case for the male speakers in the sample. Interestingly, the mean phrase duration of the by far highest rated male speaker in terms of enthusiasm (PL, mean phrase duration = 1.27 seconds) is roughly the same as that from Steve Jobs (1.25 seconds when speaking to customers).

It could be that the difference between the two data sets comes from the slightly different analysis methods. The short stimuli only consist of one phrase, so the absolute duration of that phrase is measured and used for analysis. The long stimuli consist of four phrases, so a mean across the phrases is calculated for analysis, similar to the methods in the investigations dealing with the speakers from the business context (Novák-Tót, 2016; Niebuhr et al., 2020a). A mean can be affected quite severely by outliers, so having a short mean phrase duration (as suggested) could also mean that phrases should overall be short without much variation, which is what could be behind longer mean durations. This should be investigated further for YouTubers, as it would suggest that having short phrases and therefore short pieces of information may be positive in vlog-style videos (as is also assumed for pauses by content creators on the platform, see The Film Theorists, 2020), but it might also mean that variability in phrase duration—which could also be connected to enthusiasm—is not preferred in this specific context.

Continuing on with the duration-related results, the next hypothesis predicted that longer stimuli (including pauses) would receive higher ratings (H_{P38}). This was not the case in the current data. Note that this analysis was only conducted for the long stimuli. The direction of the correlation with the *likable* ratings was positive as might be expected, but it was also far away from significance level. When the data set was split by gender and origin, curved patterns emerged for male speakers, as well as speakers from England, suggesting that there may be more to this, and that perhaps for some speaker groups having either very short or very long stimuli can affect the way a speaker is perceived in terms of likability. Similar visual patterns also occurred for the phrase duration in the long stimuli and the *likable* ratings (see Figure K.2 in Appendix K). It could be that the stimuli (and phrases, for that matter) were too short for an effect on the other attributes, as Jokisch et al. (2018) suggest that stimuli between 21 and 25 seconds are the amount of speech material to elicit the highest charisma ratings. Similarly, Caspi et al. (2019) found that first impressions solidify or change over time as content becomes more apparent. Their time frames are much larger than the ones in this study (the impression seems to solidify or change after about six minutes of speech; see Caspi et al., 2019). It is not improbable to assume, though, that content can have an effect, also with short stimuli, but perhaps not after the roughly six seconds of speech material that are presented for the highest rated speaker with the shortest stimulus duration. And perhaps the longer a rater has to listen to a speaker, the easier it may become to separate speaker and content when carrying out a rating task. This could be why

longer stimuli then received higher ratings again. This is just a theory and requires more targeted research, perhaps with several stimuli of the same speakers, lengthened one phrase at a time. Likewise, it seems like this is dependent on the speaker group, so this could also be of interest for future studies. Additionally, there is no information available on how familiarity with the speakers may play into the rating.

The next hypothesis predicted that speakers with higher speech rates would be perceived as more *charismatic*, *enthusiastic*, *persuasive*, *authentic* and *likable* (H_{P39}). Overall this was not the case. For example, the correlations with the *persuasive* ratings in both data sets, but especially in the long stimuli (it is a non-significant trend there) are negative, suggesting the opposite to the expectations: stimuli with higher speech rates were perceived as less persuasive. This seems to be mainly due to one outlier (highest rating, lowest speech rate). Likewise, visually there are hints that there are curved patterns again. For the female speakers it seems like the highest ratings are in the middle of the sample, between 5.5 and under 6 syllables per second. This would be much higher speech rates than the female speakers from the business context—Oprah Winfrey, Meg Whitman, and Ginni Rometty—who in previous investigations have mean speech rates between 4.28 and 5.01 syllables per second (Novák-Tót, 2016). In fact, the speech rates of the majority of YouTubers in the sample were higher than 5 syllables per second up to over 6 syllables per second, which reaches areas of Mark Zuckerberg’s speech rates that in the business context are said to be too high (Niebuhr et al., 2020a). Two of the male speakers had speech rates around the level of Steve Jobs or below (DH and MF, respectively), and those two are also the speakers with the highest ratings. It could therefore be that either YouTubers (mostly the male speakers in the sample) were perceived as most persuasive with lower speech rates (or rather: speech rates that are similar to charismatic business speakers) or that there may be a window of high but not too high speech rates (at least for the female speakers) that are preferred. A possible threshold for speech rate was also suggested in previous research (e.g., Berger, 2017; Berger et al., 2017; Niebuhr et al., 2020a).

There were non-significant trends of the *enthusiastic* ratings of the short stimuli with both speech rate and the phrase duration (see above), and both correlations were positive to the same degree. The similarity seems to suggest that in the current YouTube sample, the longer the phrase and at the same time the more information is presented in that amount of time, the more enthusiastic a speaker is perceived. This seems to go in hand with the general perception of YouTube—and especially vlogs as a video genre—to be fast-paced: information is presented with a string of utterances, separated by short or even absent pauses (i.e., cuts), and little audible breathing (see The Film Theorists, 2020). This also fits some viewer opinions (which are not always positive; this speaking style can polarize) in an online discussion

thread, where, for example, the “continual flow of fast information” that this way of speaking creates is mentioned (user *jodrellblank*, see Hacker News, 2015).

The final duration- and tempo-related hypothesis predicted that speakers’ stimuli with larger speech rate variation across phrases would be perceived as more *charismatic, enthusiastic persuasive, authentic* and *likable* (H_{p310}). This is overall also not the case in the current sample, despite variation of speech tempo being mentioned as a feature of “YouTube voice” (Beck, 2015). There are hints that larger speech rates variations may be perceived as slightly more enthusiastic and likable, but these correlations were not significant. Perhaps the amount of phrases in the stimuli to calculate the standard deviation of speech rate from was too small to get a representative result. Nonetheless, especially since previous research suggests that variation is one of the most important aspects of charismatic speech (Niebuhr et al., 2016b), this could be revisited in future studies with a larger data set.

10.5.5 Observations regarding gender, origin, and speaking style

Finally, this part of the investigation explored if there were gender or origin differences in the visual correlations between the acoustic features—aside from pitch level—and the rating attributes. As has become clear in the remainder of this discussion, there seem to be gender and origin differences. Many of the results visually show that there seem to be direction differences between male or female speakers, or speakers from North America or England. However, these direction differences are mainly a matter of a visual correlation in one group, and no obvious pattern in the other group. An example would be the visual correlations between mean speech rate and the *authentic* ratings for male and female speakers’ long stimuli, see Figure 10.20B). In particular, the gender- and origin-related results of this investigation are only based on descriptive statistics, since the current sample size was too small for accurate correlation calculations with separate data sets. This could be remedied and revisited with a larger data set in the future, especially because there has been very little (if any) research on charismatic speech and female speakers as well speakers from England, which should be expanded on. One correlation that suggests no difference between male and female speakers would be the median pitch with the *enthusiastic* ratings of the long stimuli, where the visual correlations for both groups were positive.

A final thought regards the presence of curved patterns in the visual results of some acoustic features and attributes when the data set is split either by origin or gender. For some analyses (e.g., the correlation of *persuasive* ratings with the pitch peak timing in the short stimuli data when they are visually split up by gender), the direction of the curve—meaning the stimuli with medium acoustic measurements compared to the rest of the sample rated receiving either highest or lowest

ratings—are reversed between the two groups, further suggesting gender-specific encoding of charisma and related attributes. This may suggest that the data should be re-analyzed using non-linear analysis methods. This would apply to the combination of acoustics and perception in this chapter, but also the direct perception results in the previous chapters. It may be that non-linear analysis methods could be better suited for the way that charisma and the charisma-adjacent attributes are perceived and may lead to more conclusive results.

Another general observation regards the *enthusiastic* ratings in the data set with the long stimuli for the North American speakers. These speakers were all similarly rated with a neutral response and without much variation between the speakers. Additionally, these ratings were overall lower than those for the English speakers. This is in line with the results from the previous two chapters which found statistically significant or at least non-significant trends for enthusiasm and speaker origin, also suggesting that North American speakers received lower enthusiasm ratings. This could be because all raters were originally from the British Isles and may have judged what they know more positively. It is possible that repeating this experiment with North American raters would change these results.

In general, there seem to be differences in how speakers on YouTube (at least as far as the current sample is concerned) and speakers from other genres like politics and business (whose results informed the expectations for this investigation) are perceived and how this perception can be related to the acoustic features. While these are only first indications, and very few significant results, it could lead future research in terms of what charisma and charismatic speech might mean for different speaking genres.

While the results of this study are mostly tentative (often based on correlation direction instead of statistically significant results) they point future research in new directions. Equally, the acoustic features included here are not the only ones relevant for charismatic speech, opening doors for further investigations. The analysis method of correlating acoustic measurements with mean ratings seems to offer interesting insights (especially if the sample size is increased) that may in the future be supplemented with other analysis methods, but overall help situate and connect stimuli and perception results more strongly.

Part IV

Discussion

Chapter 11

Discussion

11.1 Findings and implications

There were some indications and tendencies from the inferential statistics that will be used—together with findings from the descriptive statistics—to address the research questions of this investigation (Sections 11.1.1 to 11.1.3). Some additional findings relating to origin and gender differences, as well as vocal effort are also discussed.

11.1.1 Acoustic feature configuration for charisma perception

The first research question of the investigation asked how the acoustic features in a stimulus should be configured in order to be perceived as charismatic in the context of YouTube vlogs, which was referring to both the direct charisma ratings and the charisma-adjacent attributes (see RQ1 in Chapter 5). The main hypothesis associated with this research question is repeated below as H1, and while only charismatic as a perception is mentioned, all attributes are expected to go in similar directions.

H1: *Stimuli with larger pitch ranges, higher pitch level, medium speech rates, and non-rising phrase-final pitch contours on the one hand, and audible breathing and shorter pauses on the other hand are perceived as more charismatic.*

Generally, based on the perception experiments with manipulated stimuli, there seem to be at least tendencies for all of these expectations, though never for all attributes in the investigation. Larger pitch ranges and audible breathing noises seem to be perceived as more authentic, persuasive, and charismatic. Similarly, non-rising final contours seem to be preferred for the same three attributes, but only for female speakers. That suggests that the attributes authenticity, persuasiveness, and charisma may be produced with similar acoustic characteristics: pitch range, final contour direction, and audible breathing noises, which appear to be among the most important acoustic features for these attributes. This would then suggest that a larger pitch range may be perceived as more charismatic on YouTube, just

like it would be expected from business and politics (see, e.g., Berger, 2017; Mixdorff et al., 2018; Niebuhr et al., 2018a; Niebuhr and Skarnitzl, 2019). This is similar for fewer rising contours for female speakers (see Rosenberg and Hirschberg, 2009, who suggest the same for male speakers). However, the male speakers in the present investigation seem to be rated higher on the three attributes with rising final contours, contrary to previous research. Equally, the tentative preference for audible breathing noises is in line with suggestions from previous research (Michalsky and Niebuhr, 2019).

There were also effects of pitch level and speech rate with *enthusiastic* ratings, suggesting that these two acoustic features may be closer connected to encoding enthusiasm. When looking at the manipulation results, a higher pitch level seems to be perceived as more enthusiastic (this is only concerning male speakers, though) which is also in line with the expectations of the hypothesis. The speakers from England were also perceived as more charismatic with the stimuli with increased pitch level than the North American speakers, suggesting a similarity between the encoding of charisma and enthusiasm at least for this sample of speakers.

For the male speakers, enthusiasm and a medium to high speech rate also seem to be connected (this is similar for speakers from North America, where a high speech rate is preferred). This is also in line with the expectations, though the speech rates in the YouTube sample (and their manipulations) are in general much higher than those reported for at least business speakers: the speech rates in the medium speech rate category were between 5.6 and 6.6 syll/s, while speech rates of some business speakers tend to be around 4 to 5 syll/s (see Novák-Tót, 2016; Niebuhr et al., 2020a; see also the discussion of the speech rate findings in Section 10.5.4). Similarly, mean speech rate for American speakers in conversational speech was found to be around 4.8 syll/s (Syrdal, 1996). This suggests that enthusiasm on YouTube may be expressed mainly via an extremely high speech rate, which would be in line with the perception of YouTube vlogs and the internet as a whole of being incredibly fast-paced (see The Film Theorists, 2020). A medium speech rate also seems to be perceived as slightly more persuasive for female speakers, while male speakers may be preferred with low speech rates. That suggests that the female speakers need to do more and present more information quicker to be perceived as more charismatic, while male speakers may need to be a bit calmer.

It is not clear yet if the ratings between male and female speakers (for example in this case) match: that means, do male speakers with low speech rate and female speakers with medium speech rate receive similar charisma ratings? This is left as an open question for future investigations. However, the combination of the acoustics in the unchanged stimuli suggests that the YouTubers in the sample seem to be faster (i.e., have higher naturally occurring speech rates) in overall longer phrases than CEOs (compared to findings from Novák-Tót, 2016 and Niebuhr et

al., 2020a). Correlations with the ratings (especially with the *persuasive* ratings) suggest, though, that the relationship is more complicated: in the long stimuli, there was a negative trend with higher speech rates receiving lower persuasiveness ratings. For the male speakers in the sample this was visually striking; for the female speakers there was a slight curve in the pattern as medium speech rates were rated most persuasive and lower and higher rates as less persuasive. This might hint towards a sweet spot in a high, but not too high speech rate range (at least for the female speakers), which is similar to what is suggested in previous studies as well (specifically for male speakers, see Berger, 2017; Niebuhr et al., 2020a).

The idea of YouTube and the internet as being extremely fast (see The Film Theorists, 2020; see also online commentary in Hacker News, 2015) also fits with some pause duration results. First of all, there was a (visual) tendency that came out in the experiment results that suggests that pauses with different length categories in a stimulus (i.e., more variation) were perceived as more enthusiastic, as were cuts instead of pauses. The cuts could be seen as extremely short pauses, though since this actually means a complete lack of pauses, this seems to overshoot the expectations and may be a result that is specific to enthusiasm, and is a result that is specific to the YouTube context. In a speech on stage (such as in politics or business) there is no opportunity for cutting, and also no opportunity for excluding pauses from the stream of speech. Cuts never were the worst-rated stimuli, which suggests that this is accepted as part of speech on YouTube, though perception could be swayed by whether or not raters are familiar with speaking styles on YouTube. First visual observations based on additional participant information may suggest this (at least for *authentic* ratings), but this should be addressed in detail in a future investigation.

Equally, depending on the speaker group, and this time in line with the expectations from the hypothesis, shorter pauses seem to be perceived as more charismatic and persuasive on YouTube. This is perhaps the case for all speaker groups and the *charismatic* ratings—though the visual tendencies are very slight—and for the female speakers and North American speakers (note that the majority of North American speakers are female) and the *persuasive* ratings. For the *charismatic* ratings, this seems to also be corroborated by the results from the breathing-related part of the study, where the shorter of the two pause durations involved in the manipulations received higher ratings. The preference for shorter pauses seems to be in line with previous research (D’Errico et al., 2013; Niebuhr et al., 2020a), but also with the notion of a fast internet (The Film Theorists, 2020). The pauses in the YouTube sample of this investigation are even shorter than those suggested for business speakers (see Niebuhr et al., 2020a).

Overall, there were few patterns with the *likable* ratings and the manipulations. It may be that the features that were manipulated are not specific enough to relate

to likability. There was a trend suggesting that stimuli with higher mean F0 were perceived as more likable in the long stimuli of this investigation, and that this correlation is especially pronounced for the male speakers in the sample. Weiss et al. (2021) report that some previous research found that likability was correlated with higher mean pitch in German as in the current sample (Scherer, 1979) though this seems to be highly context-dependent. They did not find conclusive correlations in their study of German, though (Weiss et al., 2021). In general this may suggest that rating likability and connecting this to specific acoustic features can be difficult. Perhaps the stimuli were too short to gain enough of an impression, or the attribute is simply too subjective or complex to judge consistently.

The results have also shown that there are cultural differences in terms of speaker gender and speaker origin in the way the speakers are perceived by listeners from the British Isles. Gender and origin differences were not specifically expected, but they were included in the analyses. There is a lack of research on charismatic speech dealing with female speakers and speakers from England. However, this research has shown that there are differences that are worth looking into further (see above for some of these differences). In particular—and irrespective of the manipulation—speakers from North America were rated as less enthusiastic than speakers from England. At this point, this cannot be conclusively interpreted without also including raters from North America to see if that would then lead to higher enthusiasm ratings for North American speakers and lower ratings for speakers from England. If that would be the case, it might be an indication that raters tend to rate speaking styles and varieties closer to their own as more enthusiastic.

When looking at the *charismatic* ratings and the prosodic manipulations (Chapter 8), the male speakers also seem to be rated as consistently more charismatic than the female speakers, and no specific manipulation had median or mean ratings that were higher than those of the male speakers. This seems to be in line with previous research as well (e.g., Jokisch et al., 2018; Niebuhr et al., 2018a, 2019; Niebuhr, 2020; Gutnyk et al., 2019).

For the data in the current sample, the first research question can therefore be cautiously supported, though not for all attributes. The most alignment with the expectations is with the *charismatic* and *persuasive* ratings. Authenticity seems to behave similarly. Pitch level and speech rate seem to be mostly connected to enthusiasm, and likability is inconclusive at best.

11.1.2 Ratings and acoustic features

The second research question asked if speakers who were higher rated in the perception experiments would also use acoustic features that are connected to more

charismatic speech (according to previous literature; see RQ2 in Chapter 5). The hypothesis connected to this is repeated below as H2.

H2: *The ratings from the experiments correlate with the acoustic feature values known to be used in charismatic speech and related attributes.*

This research question is investigated with the results from the experiment combining the acoustics of the unchanged long and short stimuli and the mean ratings of the different attributes (see Chapter 10 for the investigation). That part of the project included several acoustic features: pitch level (in terms of mean F0 and median pitch, both expected to be higher), minimum and maximum F0 (both expected to be higher), pitch range and variability of pitch (both expected to be larger), pitch peak timing (later), number of emphatic accents (expected to be higher) and prominences of all levels, the prominence ratio (higher, though not investigated in previous research), speech rate (fast, not too fast), and the speech rate variability (larger), as well as phrase duration (shorter) and stimulus duration (longer). For the tempo- and duration-related features (phrase and stimulus duration, speech rate and its variability), the analysis-specific hypotheses in Chapter 10 expected no difference in the rating attributes; for the rest of the features, positive correlations were expected for the *charismatic*, *persuasive*, and *enthusiastic* ratings, and negative correlations for the *authentic* and *likable* ratings.

Overall, there seems to be some tentative evidence in line with the expectations especially for the enthusiasm ratings, and a few for the *likable* and *charismatic* ratings. There were also some results with the authenticity ratings, though they all were opposite to expectations. There were barely any correlations with the persuasiveness ratings. Additionally, when looking at the correlation directions, there were direction differences between the two data sets of short and long stimuli, which were more in line with predictions in the long stimuli data. That suggests that there may be a bigger impact of averaging across phrases in the long stimuli than initially expected. The expectations were based on previous research. The majority of previous research used averages across several phrases for analysis (see the overview in Chapter 3). This may be one explanation for the closer relationship between long stimuli and the predictions, as the methods were more similar than the lack of averaging for the short stimuli.

For example, in terms of mean F0 (and median pitch), the correlations go in the opposite direction of expectations: there was a negative correlation (significant) between the *charismatic* ratings and the mean F0 (and a non-significant trend with median pitch) in the short stimuli. Visually, this was especially the case for the male speakers and speakers from North America, but there seem to be no patterns for female and English speakers. When the correlations of mean F0 with the *charismatic* ratings (not significant) are split by speaker gender for the long stim-

uli, some patterns seem to emerge. There may be a slight positive correlation for female speakers which would go against expectations. For the male speakers, a curve seems to emerge suggesting that low and high mean F0 stimuli are rated as more charismatic than stimuli with medium mean F0, which seems to be both in line and against expectations and needs further research. A higher mean F0 across all speakers seems to be perceived as more enthusiastic (in line with expectations) and more likable (which was initially not expected).

There were some trends for the tempo- and duration-related features, though only for half of the attribute correlations. In particular, there were no correlations with the *charismatic* ratings. Stimulus duration and the variation of the speech rate across phrases had no effects on either of the rating attributes. Speech rate variation was mentioned as one of the features of YouTube voice (Beck, 2015) so this lack of correlation is surprising, but this could have been influenced by the choice of stimuli. It may be that the speech rates in the phrases were too similar in the stimuli to detect a difference for listeners. This was measured as the standard deviation of speech rate across four phrases, and the values ranged from 0.42 syll/s to 1.55 syll/s (see Table 10.2). However, this would be higher than the suggested 5 percent just noticeable difference for speech rate (Quené, 2007; for the lowest speech rate variation value in the sample, the difference to mean speech rate is 7.9 %, for the highest value it is a 26 % difference). It is also possible that future studies should look into quantifying speech rate variation differently, or that the stimuli were too short for participants to register speech rate variations.

The correlation between speech rate and the *persuasive* ratings of the long stimuli was not as expected: higher speech rates tended to be rated as less persuasive, though this was only a trend. Speech rate and phrase duration seem to mainly have an effect on the *enthusiastic* ratings, though. Speech rate was a relevant correlation in the data set of the short stimuli, and here it was a positive correlation as expected. An equal correlation was found for phrase duration and the *enthusiastic* ratings of the short stimuli which was again positive, though this was against expectations. Together this may suggest that more information (i.e., more syllables) in a longer phrase may be perceived as more enthusiastic. This is similar to previous research which found that “[the] more material presented to the subject, the more charismatic the speaker was perceived to be” (Rosenberg and Hirschberg, 2009, p. 645; the material in this case refers to the number of words in a stimulus).

The most results come from the intonation-related features including pitch peak timing, prominence ratio, and emphatic accent frequency (also the frequency of the other prominence levels, though). While not all of these features have been investigated before in relation to charismatic speech, they are mentioned as important for charismatic speech in business (especially emphatic accents; see Niebuhr et al., 2020a). These features also seem to be important especially in the context of You-

Tube, as the major characteristic of “YouTube voice” is said to be a combination of different emphasis strategies that are used together and in high frequency (see Beck, 2015). All the features investigated in this feature group are ways to emphasize, be it through length or lateness (of pitch accents; see Sections 3.2.2 and 4.3.4). The results go against the expectations for some of the attributes, though. For example, later accents were perceived as more authentic (significant with the short stimuli data) and likable (non-significant trend with the long stimuli data). In this case, the later peaks may be interpreted as more expressive and emotional (Berger et al., 2020), or unusual and effortful (Gussenhoven, 2002) which participants may associate with being more open and therefore authentic, but this needs to be addressed further in future studies. There was no effect on the *charismatic* and *persuasive* ratings.

The prominence ratio offers differing results depending on the data set. The prediction was that stimuli with more frequent prominent syllables would receive higher charisma, persuasiveness, and enthusiasm ratings because of the rhythm prominent syllables add to the melody, but that these stimuli would be perceived as less authentic and likable because the rhythm can also become less relaxed, approachable, and natural, which would likely be relevant for authenticity on YouTube. In the data set with the short stimuli, fewer prominent syllables were perceived as significantly more enthusiastic, but the opposite was the case with the long stimuli where the correlation (though not significant) was positive as expected. For the long stimuli, there was a non-significant trend for the correlation between the *charismatic* ratings. This correlation was equally positive, which suggests that within one data set, charisma and enthusiasm may be encoded similarly (for the short stimuli, this correlation was negative, again aligning with the *enthusiastic* correlation). In fact, the prominence ratio of the long stimuli was the only acoustic feature where the correlation directions of all attributes aligned with the predictions—the *charismatic*, *persuasive*, and *enthusiastic* ratings were higher with more frequent prominent syllables in the stimuli, and the *authentic* and *likable* ratings were lower. This should be revisited in future studies as this is an acoustic feature that has not been investigated in regards to charismatic speech so far, but may prove to be a relevant addition to the collection of features. Similarly, the difference in correlation direction depending on data set may need future research and could occur because of different measurement methods in the two data sets (as already mentioned above for the pitch-related features; see also Section 11.3.1 for a general discussion of the measurement methods).

Finally, the frequency with which different prominence levels arose revealed some trends and significant results. A larger amount of emphatic accents in a stimulus was perceived as significantly less likable, as was predicted. It is possible that the strong emphasis on a large portion of the words in an utterance creates a some-

what unnatural rhythm, that can seem annoying to many listeners. For example, a user in an online discussion thread mentions that “YouTube voice” (a major part of which is emphasis; see Section 4.3) was fine to listen to when it was sped up. They write that “most of the fluff becomes tolerable when it’s shorter by 33%” (user *maho*; Hacker News, 2015; third post from the top). The “shorter” can refer to segments, which would then relate to emphatic accents. It is possible that there may be differences in ratings depending on experiment participants who are accustomed to listening to YouTubers speak, and those who are not. The strong and frequent emphatic accents seem to be a specific part of speech on YouTube (Beck, 2015). In the sample of the long stimuli in this work, the number of emphatic accents ranges from 16 to 55 accents per minute. This is higher than the count per minute for the entire sample of speech material, which contains—depending on the speaker—18 to 36 emphatic accents per minute. These numbers are far from what has been reported for business speakers (roughly 1 to 7 emphatic accents per minute, see Novák-Tót, 2016, and Niebuhr et al., 2020a). Emphatic accents may be used so frequently in vlogs to act as keywords that can be understood by the speech-to-text algorithm and then be turned into search terms for better visibility (Bishop, 2018). All this taken together could potentially reduce the likability of speakers since the practice can also be interpreted as a performance for making revenue. Similarly, while not significant, the correlations for enthusiasm and, more importantly, authenticity were also negative, which may enhance the interpretation for likability. On the other hand—while there was no correlation between the frequency of emphatic accents and the *charismatic* ratings—there was a non-significant trend that suggests that more weak prominences in a stimulus tend to be perceived as less charismatic. This may then imply that the more non-weak prominences there are in a stimulus (i.e., strong or emphatic accents, which were separated for the current study, and assuming that the total number of prominences in a stimulus is similar which will be investigated in the future) the more charismatic a speaker may be perceived to be. This is just a theory at this point but will be investigated in further studies.

Because of the many correlation directions against predictions, and generally the few significant results, the results are inconclusive in regards to the research question. Further research is needed to gain more insights. In particular, there were few correlations with the charisma ratings, and those that were there were in the opposite direction than expected. Similarly, the results for the authenticity ratings were contrary to expectations. The correlations with the *enthusiastic* ratings seem to be most in line with expectations. This suggests two things. First, it is possible that the different attributes chosen for this investigation may be less strongly connected to charisma as a whole than initially expected. It may also be that the different attributes are encoded differently acoustically and then come together for

a charismatic impression. Second, it may be that charisma on YouTube and its acoustic encoding is simply different from charismatic speech in politics and business, which is where the majority of the expectations are from. To address this with certainty would require extensive research with more data and a lot more targeted stimuli, manipulations, and attributes.

An additional observation regards the connection that has been suggested between charismatic speech and increased vocal effort to signal that the listener is important (e.g., Michalsky and Niebuhr, 2019; for vocal effort in general see, e.g., Lindblom, 1990; Gussenhoven, 2002, 2016; Chen et al., 2002). There are some indications that increased vocal effort may also be linked to charisma in this study. For example, larger pitch ranges were perceived as more charismatic (and persuasive and authentic) in the perception experiment based on manipulations. Fewer weak prominences were also perceived as more charismatic, which may suggest that putting more articulatory effort into making important syllables stand out from the rest of the utterance clearly may be advantageous. Similarly, putting more effort into emphatic accents is perceived as less likable (and, to a smaller degree, less authentic), perhaps because being likable could be associated with also being relaxed which would be contrary to putting in extended amounts of vocal effort. Otherwise, the other acoustic features that may be considered to be connected to increased vocal effort (maximum F₀, standard deviation of pitch, pitch accent timing, and more prominent syllables in an utterance) seem to be mostly connected to authenticity.

A previous study with the same sample found that larger vowel spaces (which was equated to increased vocal effort) were perceived as more charismatic, but less authentic (specifically in the long stimuli; see Berger et al., 2023). It was assumed that this was the case because charisma and authenticity may be understood differently on YouTube, and authenticity would be connected to less vocal effort—especially because the participants were given a definition of what *authentic* means: “the person is not putting up an act”. This was predicted for the current study as well, but it seems to not be the case since the acoustic features connected with increased vocal effort also received the higher authenticity ratings. It could be that—despite the provided definition—participants rated the stimuli as authentic in the context of YouTube vloggers and saw an aspect of performance as authentically YouTube. This needs further investigation because there seems to be a difference between charisma and authenticity on YouTube that is at this point still inconclusive. Future studies could focus on this more closely in YouTube data, also in respect to other acoustic features (e.g., pitch peak height, coarticulation, etc.) in order to see exactly how theories like the Effort Code (Gussenhoven, 2002, 2016; Chen et al., 2002), the H&H Theory (Lindblom, 1990), or style shifting depending

on the video type (Labov, 1972; see also Section 4.3.4) may impact the balance between the perception of charisma and authenticity in YouTube videos.

11.1.3 Familiarity and charisma

The final research question in the investigation asked if there was a connection between the charisma ratings and the degree of familiarity that raters had with the speakers (RQ3). The hypothesis related to this is repeated as H3 below:

H3: *The more familiar a speaker is to the listeners, the more charismatic they are perceived.*

The investigation of this research question only deals with the direct charisma ratings for the manipulated stimuli, since the familiarity ratings were only elicited together with the direct charisma ratings.

The results confirm this hypothesis overall, with positive correlations in all data sets (which were subsets according to the different acoustic features). The linear mixed models suggest that unknown speakers are rated as less charismatic than all other degrees of familiarity for the short stimuli, but that there is no difference between familiar and known speakers in the long stimuli, especially for the male speakers.

The connection between charisma and familiarity is more complicated than simply “the more familiar, the more charismatic”, though. There are indications for some subsets and some speaker groups (e.g., final contour direction and speakers from England; pause duration/breathing noises and female speakers) that there is a rating difference between known and familiar speakers. That means that known speakers were rated as less charismatic than speakers who the experiment participants indicated sounded more familiar to them. At the same time, there was no rating difference between known speakers and speakers the participants indicated they did not know. That suggests that knowing a speaker can be just as detrimental to the charisma perception than not knowing a speaker. At this point, the attitude of a listener and the context of where the speaker is known from may come into play, pieces of information that were not elicited in the current study. It seems reasonable, though, that if a speaker is known or remembered specifically because they are not liked for whatever reason (perceived as annoying, connected to a scandal, prejudices, etc.), they are less likely to be perceived as charismatic as well because it is difficult to look past dislike and just focus on the voice.

11.2 Development of remote perception experiment method

Since the experiments of this project were carried out in the middle of the global Covid-19 pandemic, new methods for face-to-face experiments had to be developed in order to continue the research remotely. The developed set-up is discussed here, as this turned out to be a major part of the project, and recommendations for future remote perception studies are presented.

In general, the set-up with a Zoom call, screen sharing and remote control as well as an outsourcing of the metadata survey functions in a way that is not too dissimilar from face-to-face perception experiments. A face-to-face perception experiment with ExperimentMFC in Praat would always have the metadata survey in another place (online or on paper). The participants would work on a laptop with headphones that is owned either by the experimenter or the lab where the study takes place. However, the virtual experiment set-up described here is restricted to specific types of perception studies that can use ExperimentMFC, and therefore also individual sessions which could become problematic with larger sample sizes. Other programs might work as well, though no software was found that allowed the construction of non-web based or non-web saved perception experiments involving the rating of audio stimuli. ExperimentMFC is limited to identification and discrimination tasks, rating tasks with Likert scales (as in this investigation) and an additional goodness rating that can be used for a second rating on the same stimulus. It is therefore impossible to implement more than two ratings per stimulus presentation. If perception experiments fit those specifications, then the set-up with Zoom, screen sharing and remote control is a workable solution. The set-up allows for a continuation of perception research even in times of social distancing, working from home, and no possibility of travel.

However, there are quite a few negative aspects of this type of set-up. The first disadvantage that has to be kept in mind is that setting up takes a lot of time for each session (which would not happen in a lab where all equipment can remain in place) and also requires a lot of equipment in the home office. At least one desk is needed with a computer, webcam and microphone/headphones. The set-up used in this study called for a second desk off to the side because the decision was made that the participants should know that they were not being watched, but the author needed to be close to unmute the microphone in case questions were asked.

Another disadvantage for the participants is that the environment of their own home is not necessarily conducive to an experiment situation. The phone or doorbell could ring or the roommate or partner could walk through the room, which all introduce distractions that cannot be controlled and would not occur in a laboratory setting. At the same time, however, experiments in a familiar space can

sometimes be experienced as more natural and relaxed than in-person experiments in a laboratory (Leemann et al., 2020).

A solid background like a projector screen or a virtual (blurred) background are also recommended if no solid-colored wall is available. This is recommended because it increases the privacy of the experimenter's home office, while also limiting the chance of participants getting distracted or biased by background.

Remote experiment sessions are always at the mercy of technology. There were times when the computer would not recognize the headset or claim errors with the microphone. It is therefore recommended to have working back-up solutions ready and test the equipment before a session is supposed to start.

Additionally, the internet connection on both sides of the video call is a potential issue that needs to be addressed. There is always the possibility of one side (or both) not being able to connect, losing connection in the middle of the experiment, or individual stimuli not getting transmitted fully. Data loss because of problems with internet connectivity has to be expected. It is also recommended to include a response option the participants can click if the audio was not transmitted correctly or they briefly lost connection. This was done in the present study as well in order to be able to disregard specific data points in the statistical analyses without having to disregard the entire experiment session.

Another issue is that Zoom is notorious for data security problems, as the data are by default stored in the US where European data security policies do not apply, although during the duration of the experiment, Kiel University started running Zoom through their servers, making the used implementation of Zoom more secure (D. Geißler, p.c., March 2021). It is recommended to check with the data security experts at the respective university where the study takes place and discuss whether or not the method is allowed or what measures have to be taken to receive permission. In the present case, the use of Zoom was only permitted if personal information collection was outsourced to LimeSurvey whose servers store the collected data within the EU adhering to European data security policies. An anonymous code was also created. While this outsourcing of a personal information survey is not unheard of for perception experiments, it did result in having to distribute several links and pieces of information to the participants that had to be filled out in the correct order to also ensure that rights and consent for data storage were given before the experiment session started.

The final issue with the remote experiments was the extreme difficulty in finding participants. There are several possible reasons for this issue. It is very likely that the experiment still seemed too long for participants to partake in. One has to keep that in mind while planning virtual experiments. It is also very likely that potential participants are fatigued from "the new normal" regarding lectures and meetings being run on video conferencing systems. These issues have to be kept in mind

and the experiment has to be planned with an increased timeframe and schedule compared to in-person experiments.

Even though there are quite a few negative aspects to running experiments in the way described above, the method still allows for a continuation of research even during a global pandemic. Many plans need to be adjusted, many issues need to be considered and the general duration (in terms of weeks or months) might have to be increased to collect enough data to carry out reliable statistical tests. This does not mean that the method should be disregarded, but rather developed further.

Table 11.1 summarizes possible issues and some recommendations on what to keep in mind to deal with them. There is a major advantage that has not been mentioned yet: Collecting data virtually offers the opportunity to work with participants of the desired social background wherever they are—the only thing that has to be kept in mind is scheduling across time zones and access to devices. From that point of view, virtual experiments offer a cost-efficient data collection method that does not require traveling around the world to meet participants. However, with less cost for traveling and field work comes an added investment in equipment, costs, time and efforts which also have to be kept in mind. This study is only one of many studies to suggest this, to work with remote experiment methods (both for perception and production), and to continually develop these methods to build a collection of tools to be used remotely by researchers depending on study requirements (e.g., see for perception experiments: Yamamoto et al., 2021; Su et al., 2022; and for production experiments: Leemann et al., 2020; Freeman and De Decker, 2021; Ge et al., 2021; Zhang et al., 2021; Berger and Neitsch, 2023).

11.3 Limitations

11.3.1 Data treatment and measurements

The main limitation of the data treatment methods is that all data were annotated by only one person (the author). This means that all categorizations (like emphatic accent types, prominence levels, breathing noise duration, pause types, pitch accents, etc.) are based on only one person's perception, and the perception of these categories can be subjective (see, e.g., Roy et al., 2017; Brugos et al., 2023). Interrater agreement was carried out for part of the annotation—the DIMA annotation, which is in itself more prone to subjectivity in terms of prominence level and accent types. While the statistical interrater agreement was overall poor, this project is also one of the first investigations to apply DIMA to English speech data. That also means that the second annotator—while trained in DIMA annotation—was also used to annotating the intonation of different languages (in particular German, which would be

Table 11.1: Possible issues with remote perception experiments and recommendations on how to address them.

| Possible issue | Recommendations |
|-------------------------|---|
| Stimuli | – Use stimuli where the experimenter is the copyright holder to also be able to use web-based experiment services like LimeSurvey |
| Set-up | – Have everything in one place to easily move equipment (and furniture) around |
| Privacy | – Solid wall, projector screen, or virtual background (experimenter privacy) – Move away from the camera during experiment with muted microphone to not watch over the participant (participant privacy) |
| Technical malfunctions | – Test set-up (headset, microphone, screen sharing, etc.) well ahead of session – Have back-up solutions ready |
| Internet connection | – Turn off videos to get more stable audio signal – Prepare for data loss – Add response option "could not hear the audio" to avoid random responses when audio was not transmitted |
| Data security | – Store and elicit personal information/metadata on a different platform (i.e. on LimeSurvey) – Check for institution-specific regulations with data security expert |
| Participant recruitment | – Design the experiment as short as possible – Factor in increased timeframe for data collection – Factor in several rounds if participants do not respond or show up |

the case for the majority of DIMA-trained annotators), which could have an impact on prosody perception and its annotation. Therefore, the results of this project are a first step towards applying DIMA to other languages, but future studies should be based on annotations of more than one annotator or at least re-visit the interrater agreement analysis with additional annotators.

For the measurements, one of the two limitations is the difference in acoustic measurement methods in the different data sets of the short stimuli on the one hand, and the long stimuli on the other hand. Both data sets were used for the correlations with the mean ratings in Chapter 10. The mean or median measurement of each acoustic feature was taken for the phrase that made up the short stimulus of each speaker. For the long stimuli, the same was done for each of the four phrases within the stimulus, but then the mean across all phrases was used for the correlations. There are two aspects here that need to be considered. First, and tentatively speaking, it seems like there were slightly more correlation results that—at least in terms of correlation direction—aligned with the expectations derived from previous literature on charismatic speech in the long stimuli. This may suggest that averaging a measurement across phrases is more similar to previous studies in method and might therefore also come to similar findings (e.g., among many others, Novák-Tót, 2016; Niebuhr et al., 2020a). The ratings of the short stimuli may be strongly affected when their measurements can be considered outliers for

a specific speaker (see discussion in Section 10.5), which may deviate from the predictions. When averaging across several phrases is involved, as is the case with the long stimuli, such extreme values may be canceled out, which is also what would happen in previous research that also used averaging methods. Second, it could be that the short stimuli were simply too short to trigger a charismatic impression (see Jokisch et al., 2018; Caspi et al., 2019) and to elicit comparable results, but that the acoustic information in the long stimuli helped with that.

Likewise, the acoustic features included in this investigation are relatively frequently used, especially as far as the pitch measurements and the tempo-related features are concerned. The acoustic features that were collected in the intonation-related group—pitch peak timing, prominence ratio, and frequency of prominence levels (not just emphatic accent frequency)—are new to this investigation. As far as the author is aware, these three features have not been investigated in detail at the time of writing, and equally have not had extensive annotations as is the case here. The other features, however, are more frequently used and also easily influenced by each other (see the discussion of mean F0 and how it is influenced by other pitch features in Mixdorff et al., 2018; see also Section 3.2.1). One pitch feature that should be included in the future is the distance between a speaker's mean F0 in a phrase or stimulus from that speaker's baseline F0 (see Zellers and Schweitzer, 2017; Zellers, 2021, for baseline F0), calculated across the entirety of the speech material, since it has been found that male speakers in a business context are perceived as more charismatic when they speak higher than their baseline, and female speakers when they speak lower than their baseline (Niebuhr et al., 2018a). It may be that this is also relevant for charismatic speech on YouTube. Additionally, this investigation could only address a subset of the acoustic features that are known to be relevant for charismatic speech, mainly because the amount of data would have been outside the scope of the project, but also because some features, especially those related to intensity, are difficult to carry out with audio data from YouTube where the amount and types of post-processing are unknown.

Finally, explicit comparisons of the acoustics of the speakers and their stimuli to reference values from the literature not related to charismatic speech as well as speakers who were directly investigated for charisma could not be included in the present study. Some comparisons to other speakers from the business context were included for a small selection of investigated acoustic features in Chapter 10, but detailed comparisons for all features would have exceeded the capacities of the project. Additionally, not all features have reference values available as the investigations were new.

11.3.2 Experiment methodology

There were a few limitations to the project created by the experiment design itself, which should be adjusted for future data collection. The first one is that the experiment session itself was too long and repetitive, and it is likely that the participants got fatigued towards the end and perhaps did not listen carefully anymore or became more strict in their judgments as the experiment went on. This could be checked in a post-hoc test since the order of stimuli and therefore order of ratings is available in the raw data. The experiment sessions lasted about 60 minutes including instructions in the beginning and a short debriefing at the end. If the participants had come into a laboratory specifically for the experiment, such a duration probably would be less problematic, but the experiment had to be conducted online via Zoom because of the global Covid-19 pandemic. As a participant, sitting in front of a Zoom call concentrating while perhaps other things happen in their surroundings is extremely strenuous. For replications or continuations of this investigation, especially if the investigations are continued via remote methods, the experiment should be shortened in some capacity, at least in a way that that feels shorter for the participants.

One option for this could be to split the experiment session into two. This could be done two different ways: either the participants are asked to do both parts as in the current study (short stimuli with charisma-adjacent attributes and long stimuli with direct charisma ratings and familiarity; the second group does the opposite combinations), but come in on two days and do one part each day. That has the advantage that raters may forget what they did the last time and have a fresh start. It has the clear disadvantage, though, that double the amount of appointments (and opportunities for canceled sessions) have to be scheduled. An alternative would be to recruit participants for each of the four parts separately. That has the advantage of shorter sessions and only one session for each participant, but it means recruiting double the amount of participants. This proved to be extremely difficult for the remote method used in this investigation, as it took almost two years to recruit the minimum amount of participants needed for statistical analyses.

The second option to make the experiment feel shorter would be to include the possibility for breaks in the second part of the experiment that elicits the direct charisma ratings and the familiarity. Participants were able to take breaks between the different rating attributes (which was seen as a logical place for a break) in the first part of the experiment. In the second part of the experiment, there was only one rating attribute, which means there was no natural point for a break. In hindsight, at least two break opportunities should have been included nonetheless, since there were a lot of stimuli to rate, but no change in rating attribute to have some variety. Equally, this part tended to be longer in general, and it was quite

repetitive, so allowing for short breaks would have been helpful to keep the minds of the participant a bit fresher.

Another change that should be made in future continuations of the project and additional perception data collection is that the responses should be reduced to four (“Strongly agree”, “Agree”, “Disagree”, and “Strongly disagree”), leaving out the neutral answer and actually carrying out a forced choice experiment. In this investigation, the mean response, which is what statistical models like linear mixed effects models use for analysis, was pulled towards the neutral answer. That means that the results were mostly similar and that limited the possible interpretations of the results. If there is no neutral answer, the participants are forced to make a choice on their perception of an attribute. It is likely that that would result in more distinct mean responses for analysis. It could mean that clever participants then rather choose to not hear the audio (if there is a “Could not hear the audio” button included as in the current study). However, this is generally unlikely as most participants would actually make a choice if they have to—though this also has the potential to skew the results if participants are really undecided—and the risk is outweighed by the advantage the “Could not hear the audio” button brings in order to filter out audio that was not transmitted or participants indicating they did not listen to the audio because they got distracted. This happens, but a specific button they can use in that case can limit random choices.

Another limitation is that the experiment stimuli included both speakers from England and North America, but the raters were all originally from the British Isles. It is likely that this affected their ratings to some degree, mostly in that the speakers from England received higher ratings. Since the participant recruitment was so strenuous and long, there was no chance to include speakers from both cultural backgrounds, especially since this would have also meant increasing the sample significantly in order to have large enough speaker groups for statistical analysis. The rater sample is also not balanced for gender for the same reasons. However, there are the same amount of female, male, and diverse participants in each of the two experiment groups.

Finally, in terms of experiment design limitations, this investigation relies heavily on the manipulation of stimuli and therefore on the use of synthesized voices (Chapters 8 and 9). Some researchers mention the problems behind using synthesized voices for experiments:

The use of synthesized voices, for example, while achieving tight experimental control, sacrifices ecological validity and produces statistical estimates that may not generalize to normal communicative activity. (Burgoon et al., 1990, p. 141)

For the purposes of the study, the control that speech synthesis provides were

seen to be more important than the disadvantages that come with it. This is especially the case since there were similar results from both the manipulations and the acoustic measurements of the unchanged stimuli. Nonetheless, this leads to some stimulus-related limitations of the investigation, both in terms of piloting the stimuli, and the choice and manipulation of the stimuli.

The short stimuli were piloted for their naturalness and some stimuli were excluded. The excluded stimuli received a mean rating of 2 or lower on a scale from 1 (very unrealistic/unnatural) to 4 (very realistic/natural). With the possibilities of the TD-PSOLA algorithm (Charpentier and Stella, 1986; Moulines and Charpentier, 1990) in Praat (Boersma and Weenink, 2018) combined with the audio quality from YouTube, most of the stimulus manipulations actually only received ratings between two and three. That means that the manipulation methods should also be reconsidered, but it also means that raising the cut-off point for inclusion in the experiment would not have been practical in the current study. Thus, some degree of naturalness was sacrificed in order to have experimental control over specific acoustic-prosodic features. The long stimuli were not piloted, as there was less opportunity for unnaturalness. Piloting the stimuli would have been better though, in hindsight, as the breathing seems to be too short for some pause durations to be completely natural. Likewise, it would have been advantageous to run a pilot experiment to see if the differences in pause duration that were created were actually perceptible by listeners. This should be remedied in future studies. Furthermore, future stimuli should not solely be created by cutting and pasting pauses together. Rather, a combination of cutting and pasting (mainly: cutting out pauses if they are not supposed to be in the stimulus) and using the duration manipulation of the TD-PSOLA algorithm for stretching and compressing the pause durations, and to lengthen the natural breaths to fit longer pause durations (see Skarnitzl and Hledíková, 2022, for an explanation of how the TD-PSOLA algorithm's duration manipulation works).

Additionally, some of the stimuli should have been selected differently. In general, the stimuli (even the long ones) are quite short. That means that the results most likely show a first impression of the speakers. While it might be that the stimuli were short enough to not let the content influence the first impression too much, it could likewise be that the stimuli were even too short for a first impression (see Caspi et al., 2019) which could explain the overall lack of statistically significant results, as this may have led participants to opt for the neutral response rather than making a choice. Observations suggest that the raters may have been slightly more decisive in the long stimuli, though this did not translate into statistically significant results. It might suggest, however, that perhaps the longer exposure to the stimuli and therefore the speakers' voices may have had an influence on the rating certainty.

Equally crucial would be that future studies choose stimuli that fall within the middle 50 percent of data from the full speech material and are not outliers. This was not possible in the current study since the corpus was still being annotated when the stimuli were chosen. Now that all data are available, following studies could make sure to use stimuli that are representative for each speaker's voice, at least in the specific video. For a feature like speech rate, the stimuli had very high speech rates from the beginning, higher than average—ranging between 4.83 and 7.58 syll/s for the speakers in the YouTube sample, compared to an average for American English speakers of 4.8 syll/s (Syrdal, 1996; only two of the YouTube speakers were close to this average; see also Niebuhr et al., 2016b, 2020a for some average values from “normal” speakers as well as business speakers considered more or less charismatic). This could be a phenomenon specific to the speaking style of YouTube vlogs. It means, however, that the speech rates categorized as low in the manipulations (whether these were the unchanged speech rates or the manipulated ones) were already higher than average, and the high speech rates were way outside the norm. This should be controlled more closely in the future. Likewise, other aspects like the final contour shape were also not controlled in the long stimuli where this feature was not crucial for the analysis. In order to make all stimuli completely comparable, a host of different attributes that would not otherwise be investigated in the specific study should be kept as similar as is possible with the normal restrictions of variation in spoken language.

Additionally, the speaker and the experiment item (the utterance used as the basis for the stimulus manipulations) are the same in the current experiment design. That means that this study is making assumptions about charisma and the investigated related attributes for the speakers, but this is actually only the case in the context of that speaker's specific utterance. It means that all stimuli of each individual speaker have the same content. On the other hand, this makes the results comparable, which was prioritized in the current investigation.

Finally, not all rating attributes were specifically piloted, but it would have been advantageous to find out how participants understand the attributes and how they relate them to charisma. A pilot study used the attributes *charming*, *inspiring*, and *authentic*. The *authentic* rating was carried over, especially since it is also included in previous research (Rosenberg, 2010; Signorello et al., 2012a; Signorello et al., 2012b; D'Errico et al., 2013). The other two attributes seemed to be difficult to rate for the context of YouTube and in the short amount of speech material presented to the participants. *Charming* was therefore substituted by *likable*, and instead of *inspiring*, the attributes *enthusiastic* and *persuasive* were added together with the direct charisma rating in order to split up inspiration into two perhaps more immediate attributes that are also included in previous studies as parts of what it means to be charismatic (Rosenberg, 2010; Signorello et al., 2012a; Signorello et al., 2012b; D'Errico

et al., 2013). The results revealed that especially charisma, persuasiveness, and authenticity seem to be encoded similarly based on several acoustic features (pitch range, final contour direction, pause duration, pitch peak timing; see Chapters 8 through 10), and there were similarities between all attributes for other acoustic feature correlation directions (e.g., maximum and minimum F0; see Chapter 10); but overall more research into different aspects of charismatic speech is needed, perhaps also from the perspective of the raters and their specific understanding of the rating attributes.

11.3.3 Statistical methodology

Two final limitations of the investigation regard the statistical methods used for analysis. First of all, the linear mixed models that were constructed were knowledge based. That means that all variables that were deemed necessary for addressing the research questions of the project were kept in the models, and effects like overfitting were accepted to keep the models and results comparable throughout. This was chosen especially because previous research has shown that overfitting models is less problematic than it would be to underfit models (Barr et al., 2013). That also means, though, that it is likely that the models used are not the ones that would be the mathematically best-performing models.

Finally, all statistical analyses in this investigation use linear methods—linear mixed models, and Pearson and Kendall correlations. The results especially for the visual inspection of the correlations between acoustics and perception, as well as those related to familiarity and charisma, have shown, though, that there seem to be curved patterns in the data that cannot be captured by linear analysis methods. It is therefore likely that analyzing the data again with non-linear methods could lead to new results that perhaps are more conclusive to interpret. Possible analysis methods could include non-linear mixed models (Bates, 2011) or even a Bayesian version thereof (Williams et al., 2019) to try a more modern approach.

11.4 Future research

One possible future study could address other acoustic features that may be relevant for charisma specifically on YouTube. Of particular interest would be spectral features, in particular the so-called “Actor’s Formant” or “Speaker’s Formant” (see also Section 3.2.5). In a long-term average spectrum of voiced sections, the Speaker’s Formant is an energy peak in the spectrum that occurs between the 3 and 4 kHz frequency bands for male speakers, and might be higher for female speakers, though their usage is not proven yet (Bele, 2006; Master et al., 2008, 2012). It occurs often with actors who are projecting their voice and putting effort into it,

which would also apply to speakers on YouTube as the appearance on YouTube in front of a camera always has a performative aspect (see Hou, 2019; see also Chapter 4). Additionally, previous research has found that stronger speaker's formants in a (male) speaker's production leads to higher charisma ratings in the sense that raters were more likely to invest in a speaker's (imagined) company (Niebuhr et al., 2018a). Equally, the Speaker's Formant could potentially play into the difference in perception of charisma and authenticity that was noted in Chapter 10 and be a possible explanation for it. A study like this could also be paired with a perception experiment to investigate how speakers with overall stronger or weaker energy in the Speaker's Formant area of a spectrum are perceived, in particular in regards to charisma and authenticity.

A possible rater-oriented study for the future (instead of speaker-oriented, as the current project) is to see how the ratings of the speakers differ between experiment participants who indicated that they watch YouTube vlogs and/or follow internet personalities (which were close to 70% of the participants in this project) and those participants who did not indicate they watched vlogs. It is possible that the speaking style of YouTube vlogs may be rated more positively if listeners have been previously introduced to it and seem to like this speaking style, while listeners who are not accustomed to it may rate it significantly more negatively, considering that there are also many negative opinions about YouTube and the speaking style in vlogs (Beck, 2015; Hacker News, 2015; see Section 4.3). Similarly, future studies based on new material could target specific features. Two of the main differences between YouTubers and speakers from the business context were the number of emphatic accent occurrences and the high speech rate in fairly long phrases (see Section 11.1 above for a discussion). These features could be manipulated using the TD-PSOLA duration manipulation (Charpentier and Stella, 1986; Moulines and Charpentier, 1990) as in this project. For emphatic accents, this way the segmental lengthening could be investigated in a step-wise manner to also find out how much lengthening is too much lengthening for charisma and related attributes. Such an investigation could be paired with attribute ratings on the one hand, but also more indirect ratings on the other hand. Questions could include, for example, if participants would be inclined to a) watch a video of that YouTuber, and b) consider subscribing, plus perhaps reasons for their choice—with the task asking participants to decide only based on a speaker's voice. This would be a similar approach to the methodology presented in Niebuhr et al. (2018a). Should such a study be created completely new it would be helpful to also add more detailed questions about vlog consumption and vlog type preferences to get a much clearer picture of the listeners' YouTube backgrounds. This type of data was not collected in the current study.

Another possible avenue for future research that is even more focused on You-

Tube is to make use of the fact that YouTube is a visual platform, and the videos belonging to the audios that were analyzed in this project are available online. That means that studying YouTube videos and speech on YouTube basically lends itself to work on multi-modal communication. For example, it is known that acoustic prominences mostly also go in hand with larger or more distinct gestures (Ambrazaitis and House, 2022). Since the speech material in the current corpus revealed such an extremely high amount of prominences with their level annotations, this could be used to investigate if stronger acoustic prominence is accompanied by stronger gestures in YouTube vlogs (as is also suggested by Ambrazaitis and House, 2023, for news readers). Possible gestures to test this with could be eyebrow movements or head movements as both are generally visible in the video frame in YouTube vlogs. A previous study investigated the co-occurrence of pitch accents of different prominence levels and eyebrow movements in a subsample of the speaker sample in this investigation (Berger and Zellers, 2022). The study used an automated analysis of the gestures (OpenFace; see Baltrusaitis et al., 2018). The results suggested correlations between pitch height and strength of the eyebrow movement (Berger and Zellers, 2022), while the prominence level did not predict the occurrence of an eyebrow movement, but the presence of specific accent types did. However, there seems to be no time alignment between eyebrow movements and pitch accents, at least not with the analysis methods of the study, though the results overall were not conclusive (Berger and Zellers, 2022). Similar analyses could be carried out with different gestures like head movements to see if there is evidence for alignment in time or space (i.e., in strength of gesture and acoustics; see Ambrazaitis and House, 2023). The positive correlations in Berger and Zellers (2022) might suggest that stronger acoustic prominence (in terms of pitch height) and stronger gestures could appear together also in the YouTube sample, but also including the generally more noticeable head movements would provide relevant new information for the multi-modal encoding of prominence.

Finally, one of the main findings of this project is the relationship between charisma and familiarity, which future studies can address in further detail. The results of this project suggest that knowing a speaker can—at least in some speaker groups—be a similar disadvantage for the charisma perception as not knowing a speaker. This was not part of previous research which generally suggests that recognized speakers are perceived as more charismatic than unrecognized ones (see, e.g., Rosenberg and Hirschberg, 2005, 2009; Biadys et al., 2007; Jokisch et al., 2018; see also Lavan et al., 2016, for speaker identification and familiarity), and the correlations between charisma ratings and familiarity ratings were also consistently positive and significant in the present investigation. The current project already differs from other previous studies (e.g., Rosenberg and Hirschberg, 2005, 2009; Jokisch et al., 2018) in that it offered four different categories or degrees of familiarity. Pre-

vious studies seem to mostly ask raters if they recognize a speaker, or at least no further information is provided. The investigation of familiarity and charisma (and the other attributes, for that matter) could be designed in an even more fine-grained manner with very specific categories in order to find out how exactly the concepts play together. Future studies should additionally have the participants indicate if they know or remember a speaker because they like them or specifically dislike them, perhaps also with their reason (for example, if they find a speaker annoying this is likely going to influence their charisma rating, which might explain the results from this study), and from what context they know a speaker (YouTube, TV, other social media, etc.). This could also be included if the participants were simply familiar with a speaker or were generally unsure, but in this case they should indicate where they think they know the speaker from. Another step could be to ask the participants to name the speakers or channels on top of the other information.

Chapter 12

Conclusions

To conclude this project, charismatic speech on YouTube might be connected to a few acoustic-prosodic feature characteristics. It seems like having more stronger prominences, shorter pauses with audible breathing noises, and larger pitch ranges may be perceived as more charismatic in the context of YouTube, but also non-rising contours for female and rising contours for male speakers, and faster speech rates (in particular, faster than average speech rates). These seem to be similar features compared to charismatic speech in politics and business, but on a higher level especially in terms of prominence frequency, pause duration and speech rate.

Authenticity also seems to be important on YouTube, and was unexpectedly connected to more effortful acoustic productions. This may suggest that authenticity on YouTube may be—perhaps subconsciously—connected to performance which may be seen as authentic for being a YouTuber or vlogger. Authenticity, charisma, and persuasiveness seemed to align for several acoustic features throughout the investigation, which might suggest that these three attributes are quite closely linked, but that enthusiasm and likability may be something different.

Another main finding of the dissertation was that the more familiar listeners were with the speakers, the more charismatic the speakers were rated. However, the results also suggest that this correlation may not be linear, and that knowing a speaker and not knowing them can be equally problematic for charisma perception. This is where, in future studies, the attitude of the listeners towards a speaker can be included.

Nonetheless, there are many results that went in the opposite direction of expectations, and in particular for the manipulations there were few significant results. This together with the fact that many phrases of the stimuli acoustically fell outside the main amount of values in the full amount of speech material suggests that the results of this investigation can only be related to the stimuli used here. Generalization to the speakers in particular and “YouTube voice” as a whole—together with a better idea if this can be termed a consistent speaking style—will only be possible after several additional studies that revisit the data and findings of this investigation with different methods, stimuli, and experiment participants. This

study mainly made first steps towards ultimately understanding what charisma and related attributes might mean on YouTube.

Equally, this is the first step towards diversifying the research landscape of charismatic speech, since it is a) one of the few studies that include findings on female speakers, b) the first investigation on charismatic speech in speakers from England, and c) a new speaking genre. It is specifically important to also include other speaker groups in terms of origin and gender, but also other speaking genres (for example in terms of other professions who communicate with other people daily). With the YouTubers, this study moves slightly more towards every day, though exaggerated, language.

However, one always has to keep in mind that charisma can also be used for controversial or problematic motives. It is the speakers who are important, and it is not only the voice that plays a role. Listeners are therefore always required to be critical of presentations and not only be swayed by the voice, but also pay attention to what a speaker is saying.

As mentioned, this project mainly did preliminary work that can lead to future, more targeted studies. A part of this is the creation of a large corpus of annotated speech material (roughly 50 minutes) that is annotated on several levels both regarding interval-related studies and intonation alike. This corpus offers the opportunity for studies that are not only related to charisma and emotional speech. It also includes many features that could not be addressed yet, but will be in the future. For example, spectral analyses regarding the Speaker's Formant are possible, there are annotations for the types of pauses (though only on an impressionistic basis which might need subsequent revisions), the forms of filler particles, and many more intonation-related topics like boundary tones, use of intermediate phrases, and so on.

This was the first investigation to explicitly investigate charisma on YouTube as opposed to charismatic speech in politics or business. It is therefore the first step towards adding knowledge about a new genre of speech to the literature that can be explored further. In particular, the results of this project open doors to a large amount of possible future research that could offer further insights into a speaking style that bridges the gap between entertainment and business. The corpus annotation allows for more wide-spread investigations by hopefully many researchers.

Bibliography

- Abuljadail, M. (2018). Users and nonusers of YouTube and online video services. In L. Ha (Ed.), *The audience and business of YouTube and online videos* (pp. 17–28). Lexington Books.
- Alfie Deyes. (2013, April 21). *Draw My Life | PointlessBlog* [Video]. YouTube. <https://youtu.be/NJcaix-wLfo?t=123>
- Alfie Deyes Vlogs. (2018, January 18). *Am I happy making YouTube videos?* [Video]. YouTube. <https://youtu.be/pLYJvqKyFh4?si=viU-6fzGaPZpHmRh&t=613>
- Alfie Deyes Vlogs. (2020, April 19). *Answering questions you asked us* [Video]. YouTube. <https://youtu.be/JwB3gBvyRIE?si=d96XrmhHbdiwCNfE&t=55>
- AmazingPhil. (2013, February 01). *Draw My Life | AmazingPhil* [Video]. YouTube. <https://youtu.be/Mv1SLUjDGpA?t=143>
- AmazingPhil. (2018, November 15). *Why I Went To Hospital* [Video]. YouTube. <https://www.youtube.com/watch?v=IDrBKN9Lgk4>
- Ambrazaitis, G., & House, D. (2022). Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters. *Laboratory Phonology*, 24(1), 1–35. <https://doi.org/10.16995/labphon.6430>
- Ambrazaitis, G., & House, D. (2023). The multimodal nature of prominence: Some directions for the study of the relation between gestures and pitch accents. *Proceedings of the 13th International Conference of Nordic Prosody*, 262–273. <https://doi.org/10.2478/9788366675728-024>
- Andersen-Peters, J. (2016). Charitable YouTube discourse: Markiplier and the elements of online communication. In A. Richardson (Ed.), *Stylus: Knights Write Showcase Special Issue, Spring 2016* (pp. 54–62). University of Central Florida.
- Antonakis, J. (2012). Transformational and charismatic leadership. In D. V. Day & J. Antonakis (Eds.), *The Nature of Leadership (Second Edition)* (pp. 256–288). SAGE Publications.
- Antonakis, J. (2017). Charisma and the “new leadership”. In J. Antonakis & D. D. Day (Eds.), *The nature of leadership* (pp. 56–81). Sage London.
- Antonakis, J., Bastardo, N., Jacquart, P., & Shamir, B. (2016). Charisma: An ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organizational Behavior*, 3, 293–319.
- Antonakis, J., Fenley, M., & Liechti, S. (2011). Can charisma be taught? Tests of two interventions. *Academy of Management Learning & Education*, 10(3), 374–396.
- Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37(5), 715–727.
- Arthurs, J., Drakopoulou, S., & Gandini, A. (2018). Researching YouTube. *Convergence: The International Journal of Research into New Media Technologies*, 24(1), 3–15.
- Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40(3), 351–373.

- Arvaniti, A., & Atkins, M. (2016). Uptalk in Southern British English. *Proceedings of Speech Prosody 2016, Boston, Massachusetts*, 153–157. <https://doi.org/10.21437/speechprosody.2016-32>
- AudacityTeam. (2017). Audacity: Free audio editor and recorder [Computer application, version 2.3.1]. Retrieved December 7, 2023, from <https://audacityteam.org/>
- Awamleh, R., & Gardner, W. L. (1999). Perceptions of leader charisma and effectiveness: The effects of vision content, delivery, and organizational performance. *The Leadership Quarterly*, 10(3), 345–373.
- Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.
- Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. <https://doi.org/10.1109/fg.2018.00019>
- Banzina, E. (2021). Exploring phonetic cues to persuasive oral presentation: A study with British English speakers and English L2 learners. *Language Teaching Research*, 1–20. <https://doi.org/10.1177/13621688211037610>
- Banzina, E., & Niebuhr, O. (2023). How to pause in charismatic speeches: A case study of Barack Obama. *Proceedings of the 13th International Conference of Nordic Prosody*, 160–175. <https://doi.org/doi:10.2478/9788366675728-014>
- Barbosa, P., & Niebuhr, O. (2020). Persuasive speech is a matter of acoustics and chest breathing only. In M. Elmentaler & O. Niebuhr (Eds.), *An den Rändern der Sprache* (pp. 551–578). Peter Lang.
- Barr, D., Levy, R. P., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.31234/osf.io/39mhs>
- Bärtl, M. (2018). YouTube channels, uploads and views: A statistical analysis of the past 10 years. *Convergence: The International Journal of Research into New Media Technologies*, 24(1), 16–32. <https://doi.org/10.1177/1354856517736979>
- Bass, B. M. (1985). *Leadership and performance beyond expectations*. Free Press.
- Bass, B. M. (1988). Evolving perspectives on charismatic leadership. In J. A. Conger & R. N. Kanungo (Eds.), *Charismatic leadership: The elusive factor in organizational effectiveness* (pp. 40–77). Jossey-Bass.
- Bastardo, N. (2020). Signaling charisma. In J. P. Zúquete (Ed.), *Routledge International Handbook of Charisma* (pp. 313–323). Routledge.
- Bates, D. (2011). Mixed models in R using the lme4 package Part 6: Nonlinear mixed models. *Statistics*, 1, 1–9. Retrieved December 12, 2023, from <https://lme4.r-forge.r-project.org/slides/2011-01-11-Madison/6NLMMH.pdf>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2023). Package ‘lme4’. Retrieved December 7, 2023, from <https://github.com/lme4/lme4/>
- Beck, J. (2014). Experiment-MFC: Erstellung und Auswertung eines Perzeptions experiments in Praat. *Kieler Arbeiten in Linguistik und Phonetik (KALIPHO)*, 2, 81–113.
- Beck, J. (2015). The linguistics of ‘YouTube Voice’. *The Atlantic*, December 7, 2015. Retrieved December 7, 2023, from <https://www.theatlantic.com/technology/archive/2015/12/the-linguistics-of-youtube-voice/418962/>

- Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 9–54). OUP Oxford.
- Bele, I. V. (2006). The speaker's formant. *Journal of Voice*, 20(4), 555–578.
- Belk, R. W. (2013). Extended Self in a Digital World. *Journal of Consumer Research*, 40(3), 477–500. <https://doi.org/10.1086/671052>
- Bell, A. (1992). Hit and miss: Referee Design in the dialects of New Zealand television advertisements. *Language & Communication*, 12(3), 327–340. <https://eric.ed.gov/?id=ej453044>
- Benus, S., Enos, F., Hirschberg, J., & Shriberg, E. (2006). Pauses in deceptive speech. *Proceedings of Speech Prosody 2006, Dresden, Germany*, 1–4.
- Berger, S. (2017). *Winning over an audience – A perception-based analysis of prosodic features of charismatic speech* [Master thesis]. Kiel University. Kiel.
- Berger, S. (forthcoming). Charismatic speech on YouTube and business stage: The use of emphasis strategies in two public speaking styles. In L. Anderwald & E. Eggert (Eds.), *Linguistik 2.0 sprachdiskurse in interaktiven medien im internet*. Lang.
- Berger, S., & Neitsch, J. (2023). Investigating and comparing remote recording methods. *Proceedings of the 13th International Conference of Nordic Prosody*, 200–211. <https://doi.org/doi:10.2478/9788366675728-017>
- Berger, S., Niebuhr, O., & Brem, A. (2020). Of voices and votes: Phonetic charisma and the myth of Nixon's radio victory in his first 1960 TV debate with Kennedy. In M. Elmentaler & O. Niebuhr (Eds.), *An den Rändern der Sprache* (pp. 109–145).
- Berger, S., Niebuhr, O., & Peters, B. (2017). Winning over an audience – A perception-based analysis of prosodic features of charismatic speech. *Proceedings of the 43rd Annual Conference of the German Acoustical Society, Kiel, Germany*, 1454–1457.
- Berger, S., & Zellers, M. (2021). Pitch accent position, peak height, and prominence level relative to accented vowel onset on YouTube. *Proceedings of the 1st International Conference on Tone and Intonation (TAI)*, 137–141.
- Berger, S., & Zellers, M. (2022). Multimodal prominence marking in semi-spontaneous YouTube monologs: The interaction of intonation and eyebrow movements. *Frontiers in Communication*, 7(903015), 1–19. <https://doi.org/10.3389/fcomm.2022.903015>
- Berger, S., Zellers, M., & Niebuhr, O. (2023). "YouTube space" – A preliminary investigation of vowel spaces and charisma perception on YouTube. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)* (pp. 3770–3774). Guarant International.
- Biadisy, F., Hirschberg, J., Rosenberg, A., & Dakka, W. (2007). Comparing American and Palestinian perceptions of charisma using acoustic-prosodic and lexical analysis. *Eighth Annual Conference of the International Speech Communication Association*, 2221–2224.
- Biadisy, F., Rosenberg, A., Carlson, R., Hirschberg, J., & Strangert, E. (2008). A cross-cultural comparison of american, palestinian, and swedish perception of charismatic speech. *Proceedings of Speech Prosody 2008, Campinas, Brazil*, 579–582.

- Bishop, S. (2018). Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. *Convergence: The International Journal of Research into New Media Technologies*, 24(1), 69–84.
- Boersma, P. (2013). *ExperimentMFC*. Retrieved December 20, 2023, from <https://www.fon.hum.uva.nl/praat/manual/ExperimentMFC.html>
- Boersma, P. (2016). ExperimentMFC 2.1. The experiment file. Retrieved June 28, 2023, from https://www.fon.hum.uva.nl/praat/manual/ExperimentMFC_2_1_The_experiment_file.html
- Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [version 6.0.37]. <http://www.praat.org/>
- Bono, J. E., & Ilies, R. (2006). Charisma, positive emotions and mood contagion. *The Leadership Quarterly*, 17(4), 317–334.
- Bosker, H. R. (2017). The role of temporal amplitude modulations in the political arena: Hillary Clinton vs. Donald Trump. *Proceedings of Interspeech 2017*, 2228–2232.
- Brugos, A., Breen, M., Shattuck-Hufnagel, S., Veilleux, N., & Barnes, J. (2023). Marking prominence: Towards cue-based annotation of prosodic prominence. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences, Prague, Czech Republic* (pp. 1697–1701). Guarant International.
- Burgess, J., & Green, J. (2009a). The entrepreneurial Vlogger: Participatory culture beyond the professional-amateur divide. In P. Snickers & P. Vondreau (Eds.), *The YouTube Reader* (pp. 89–107).
- Burgess, J., & Green, J. (2009b). *YouTube: Online Video and Participatory Culture*. John Wiley & Sons.
- Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal behaviors, persuasion, and credibility. *Human Communication Research*, 17(1), 140–169.
- Caspi, A., Bogler, R., & Tzuman, O. (2019). “Judging a book by its cover”: The dominance of delivery over content when perceiving charisma. *Group & Organization Management*, 1–32.
- Charpentier, F., & Stella, M. (1986). Diphone synthesis using an overlap-add technique for speech waveforms concatenation. *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015–2018. <https://doi.org/10.1109/ICASSP.1986.1168657>
- Chen, A., Gussenhoven, C., & Rietveld, T. (2002). Language-specific uses of the Effort Code. *Proceedings of Speech Prosody 2002*, 1–4.
- Chi, C. (2021). YouTube algorithm: The constantly updated guide to YouTube’s updates & changes. <https://blog.hubspot.com/marketing/youtube-algorithm>
- Clark, M. (2020). Michael Clark: Convergence Problems. Retrieved December 7, 2023, from <https://m-clark.github.io/posts/2020-03-16-convergence/>
- Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, 39(2), 237–245.
- Cocker, H. L., & Cronin, J. (2017). Charismatic authority and the YouTuber: Unpacking the new cults of personality. *Marketing Theory*, 17(4), 455–472.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>

- Colleen Ballinger. (2017, December 31). *My experience with Netflix* [Video]. YouTube. <https://www.youtube.com/watch?v=KWM4DVRrdaA>
- Conde, M. A., Forteza-Martínez, A., & Andrade-Martínez, C. M. (2020). Análisis de la capacidad de liderazgo y el carisma de los principales youtubers españoles [Analysis of the leadership capacity and charisma of the main Spanish YouTubers]. *3C TIC: Cuadernos de desarrollo aplicados a las TIC*, 9(3), 17–41. <https://doi.org/10.17993/3ctic.2020.93.17-41>
- Conger, J. A., & Kanungo, R. N. (1987). Toward a behavioral theory of charismatic leadership in organizational settings. *Academy of Management Review*, 12(4), 637–647. <https://doi.org/10.5465/amr.1987.4306715>
- Conger, J. A., & Kanungo, R. N. (1988). Conclusion: Patterns and trends in studying charismatic leadership. In J. A. Conger & R. N. Kanungo (Eds.), *Charismatic leadership: The elusive factor in organizational effectiveness* (pp. 324–336). Jossey-Bass.
- Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language*, 46, 49–70. <https://www.degruyter.com/document/doi/10.1515/ijsl.1984.46.49/html>
- Crystal, D. (2003). *A dictionary of linguistics & phonetics (Fifth Edition)*. Blackwell Publishing.
- Daly, N., & Warren, P. (2001). Pitching it differently in New Zealand English: Speaker sex and intonation patterns. *Journal of Sociolinguistics*, 5(1), 85–96.
- Daniel Howell. (2017, October 12). *Daniel and Depression* [Video]. YouTube. <https://www.youtube.com/watch?v=Wp2TUPo5W0c>
- Daniel Howell. (2013, April 30). *Draw My Life - Dan Howell* [Video]. YouTube. <https://youtu.be/ypDWE-3kdgA?t=272>
- David JP Phillips. (2020, August 21). *Public speaker reacts to Markiplier* [Video]. YouTube. <https://www.youtube.com/watch?v=2chSEX5zt2A&t=424s>
- Davies, J. C. (1954). Charisma in the 1952 Campaign. *American Political Science Review*, 48(4), 1083–1102. <https://doi.org/10.2307/1951012>
- de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385–390. <https://doi.org/10.3758/BRM.41.2.385>
- Deller, R. A., & Murphy, K. (2020). ‘Zoella hasn’t really written a book, she’s written a cheque’: Mainstream media representations of YouTube celebrities. *European Journal of Cultural Studies*, 23(1), 112–132. <https://doi.org/10.1177/1367549419861638>
- D’Errico, F., Niebuhr, O., & Poggi, I. (2019). Humble voices in political communication: A speech analysis across two cultures. In S. Misra, O. Gervasi, B. Murgante, E. Stankova, V. Korkhov, C. M. Torre, A. M. A. C. Rocha, D. Taniar, B. O. Apduhan, & E. Tarantino (Eds.), *Computational science and its applications – ICCSA 2019: 19th International Conference, Saint Petersburg, Russia, July 1–4, 2019, Proceedings, Part II* (pp. 361–374). Springer.
- D’Errico, F., Signorello, R., Demolin, D., & Poggi, I. (2013). The perception of charisma from voice: A cross-cultural study. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 552–557.
- Donadeli, J. M., & Strapasson, B. A. (2015). Effects of monitoring and social reprimands on instruction-following in undergraduate students. *The Psychological Record*, 65, 177–188. <https://doi.org/10.1007/s40732-014-0099-7>

- Dredge, S. (2016). Why are YouTube stars so popular? *The Guardian*, February 3, 2016. Retrieved December 8, 2023, from <https://www.theguardian.com/technology/2016/feb/03/why-youtube-stars-popular-zoella>
- Edinburgh TV Festival. (2017, August 24). *Dan & Phil with Sue Perkins | The Great YouTube Take Off | EITF 2017* [Video]. YouTube. https://youtu.be/CwGFValHYxg?si=RJ0Gwu_2-Jp0iJIH&t=783
- Fischer, K., Niebuhr, O., & Asadi, A. (2022). The voice of creativity: Effects of pitch range. In *Studenttexte zur Sprachkommunikation (Vol. 103): Elektronische Sprachsignalverarbeitung 2022* (pp. 121–130, Vol. 103). TUBPress.
- Fischer, K., Niebuhr, O., Jensen, L. C., & Bodenhausen, L. (2019). Speech melody matters – How robots profit from using charismatic speech. *ACM Transactions on Human-Robot Interaction*, 9(1), 1–21. <https://doi.org/10.1145/3344274>
- Fisher, A., & Ha, L. (2018). What do digital natives watch on YouTube? In L. Ha (Ed.), *The audience and business of YouTube and online videos* (pp. 29–40). Lexington Books.
- Flora, C. (2005). The X-Factors of Success. *Psychology Today*. Retrieved December 8, 2023, from <https://www.psychologytoday.com/us/articles/200505/the-x-factors-success>
- Fox, J., Weisberg, S., & Price, B. (2023). Package 'car'. Retrieved 2023-12-08, from <https://r-forge.r-project.org/projects/car/>
- Freeman, V., & De Decker, P. (2021). Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps. *The Journal of the Acoustical Society of America*, 149(2), 1211–1223. <https://doi.org/10.1121/10.0003529>
- FreeTime. (2021). *FormatFactory* [Computer application]. Retrieved 2019, from <http://www.pcfreetime.com/formatfactory/index.php?language=de>
- Frese, M., Beigel, S., & Schoenborn, S. (2003). Action training for charismatic leadership: Two evaluations of studies of a commercial training module on inspirational communication of a vision. *Personnel Psychology*, 56(3), 671–698.
- Frøkjær-Jensen, B., & Prytz, S. (1976). Registration of voice quality. *Annual Report of the Institute of Phonetics University of Copenhagen*, 9, 237–251.
- Funk, M. (2020). How Many YouTube Channels Are There? Retrieved December 8, 2023, from <https://www.tubics.com/blog/number-of-youtube-channels>
- Ge, C., Xiong, Y., & Mok, P. (2021). How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements. *Proceedings of Interspeech 2021*, 3984–3988. <https://doi.org/10.21437/interspeech.2021-1122>
- Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society*, 15(8), 1348–1365. <https://doi.org/10.1177/1461444812472322>
- Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Language and Speech*, 4(4), 232–237. <https://doi.org/10.1177/002383096100400405>
- Grabe, E. (1997). Comparative intonational phonology: English and German. In A. Botinis (Ed.), *Intonation: Theory, Models, and Applications: Proceedings of an ESCA Workshop, September 18 - 20, Athens, Greece* (pp. 157–160).
- Grabe, E., Post, B., Nolan, F., & Farrar, K. (2000). Pitch accent realization in four varieties of British English. *Journal of Phonetics*, 28, 161–185. <https://doi.org/10.006/jpho.2000.0111>

- Grabo, A., Spisak, B. R., & van Vugt, M. (2017). Charisma as signal: An evolutionary perspective on charismatic leadership. *The Leadership Quarterly*, 28(4), 473–485.
- Graddol, D. (1986). Discourse specific pitch behavior. In C. Johns-Lewis (Ed.), *Intonation in Discourse* (pp. 221–237). Hill Press, Inc.
- Green, A. (2015). What is 'YouTube Voice'? A linguist breaks it down. *Mental Floss*, December 10, 2015. Retrieved December 8, 2023, from <https://www.mentalfloss.com/article/72291/what-youtube-voice-linguist-breaks-it-down>
- GTLive. (2019, March 27). *GTeaLive: Europe PASSED Article 13! Your Memes Are Banned?* [Video]. YouTube. <https://www.youtube.com/watch?v=wZi4N2JEr7M&t=1154s>
- Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and phonology. *Proceedings of Speech Prosody 2002, Aix-en-Provence, France*, 1–11.
- Gussenhoven, C. (2016). Foundations of Intonational Meaning: Anatomical and Physiological Factors. *Topics in Cognitive Science*, 8(2), 425–434. <https://doi.org/10.1111/tops.12197>
- Gutnyk, A., Niebuhr, O., & Gu, W. (2019). Differences in gender-specific charismatic speech across countries and languages. *1st International Seminar on the Foundations of Speech: Pausing, Breathing and Voice*, 27–29.
- Gutnyk, A., Niebuhr, O., & Gu, W. (2020). The role of audience gender in giving product presentations. *12th International Seminar on Speech Production*, 1–2.
- Guzman, M., Correa, S., Munoz, D., & Mayerhoff, R. (2013). Influence on spectral energy distribution of emotional expression. *Journal of Voice*, 27(1), 129e1–129e10.
- Ha, L. (2018a). Most popular YouTube channels. In L. Ha (Ed.), *The audience and business of YouTube and online videos* (pp. 135–160). Lexington Books.
- Ha, L. (2018b). YouTube as a global online video portal and an alternative to TV. In L. Ha (Ed.), *The audience and business of YouTube and online videos* (pp. 1–16). Lexington Books.
- Hacker News. (2015). *The linguistics of 'YouTube Voice' | Hacker News* [Online discussion; December 8, 2015]. Retrieved December 11, 2023, from <https://news.ycombinator.com/item?id=10693664>
- Hagi, S. (2017). The rise of the 'YouTube Voice' and why vloggers want it to stop. *Vice.com*, March 28, 2017. Retrieved December 11, 2023, from <https://www.vice.com/en/article/aepn94/the-rise-of-youtube-voice-and-why-vloggers-want-it-to-stop>
- Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *The Journal of the Acoustical Society of America*, 130(1), 508–513. <https://doi.org/10.1121/1.3598457>
- Heldner, M., & Włodarczak, M. (2016). Is breathing silence? *Fonetik 2016, Stockholm, Sweden, 13-15 June, 2016*, 35–38.
- Henderson, A. J., & Skarnitzl, R. (2022). "A better me": Using acoustically modified learner voices as models. *Language Learning & Technology*, 26(1), 1–21.
- Henton, C. (1995). Pitch dynamism in female and male speech. *Language & Communication*, 15(1), 43–61.
- Hiraga, Y. (2005). British attitudes towards six varieties of English in the USA and Britain. *World Englishes*, 24(3), 289–308.

- Hiroiyuki, T., & Rathcke, T. V. (2016). Then, what is charisma? The role of audio-visual prosody in L1 and L2 political speeches. *P & P 12: Proceedings of the Conference on Phonetics and Phonology in German-speaking countries*, 1–3.
- Holladay, S. J., & Coombs, W. T. (1993). Communicating visions: An exploration of the role of delivery in the creation of leader charisma. *Management Communication Quarterly*, 6(4), 405–427.
- Holladay, S. J., & Coombs, W. T. (1994). Speaking of visions and visions being spoken: An exploration of the effects of content and delivery on perceptions of leader charisma. *Management Communication Quarterly*, 8(2), 165–189.
- Hollister, S. (2021). Google sets all-time records as search and YouTube profits soar. *The Verge*, July 27, 2021. Retrieved December 11, 2023, from <https://www.theverge.com/2021/7/27/22596592/google-q2-2021-record-revenue-profit-youtube-ad-cloud-search>
- Hou, M. (2019). Social media celebrity and the institutionalization of YouTube. *Convergence: The International Journal of Research into New Media Technologies*, 25(3), 534–553. <https://doi.org/10.1177/1354856517750368>
- Howell, D., & Lester, P. (2015). *The Amazing Book is Not on Fire: The World of Dan and Phil*. Ebury Press.
- Howell, J. M., & Frost, P. J. (1989). A laboratory study of charismatic leadership. *Organizational Behavior and Human Decision Processes*, 43(2), 243–269.
- Howell, J. M., & Shamir, B. (2005). The role of followers in the charismatic leadership process: Relationships and their consequences. *Academy of Management Review*, 30(1), 96–112.
- Ilies, R., Curşeu, P. L., Dimotakis, N., & Spitzmuller, M. (2012). Leaders' emotional expressiveness and their behavioural and relational authenticity: Effects on followers. *European Journal of Work and Organizational Psychology*, 4–14. <https://doi.org/10.1080/1359432X.2011.626199>
- Im, S., Cole, J., & Baumann, S. (2023). Standing out in context: Prominence in the production and perception of public speech. *Laboratory Phonology*, 14(1), 1–62. <https://doi.org/10.16995/labphon.6417>
- Jennings, R. (2021). How should an influencer sound? *Vox*, July 13, 2021. Retrieved December 11, 2023, from <https://www.vox.com/the-goods/2021/7/13/22570476/youtube-voice-tiktok-influencer-sound>
- Johnson, S. K., & Dipboye, R. L. (2008). Effects of charismatic content and delivery on follower task performance. *Group & Organization Management*, 33(1), 77–106. <https://doi.org/10.1177/1059601106291072>
- Jokisch, O., Iaroshenko, V., Maruschke, M., & Ding, H. (2018). Influence of age, gender and sample duration on the charisma assessment of German speakers. *Studenten zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, 224–231.
- Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45, 326–347.
- Klein, K. J., & House, R. J. (1995). On fire: Charismatic leadership and levels of analysis. *The Leadership Quarterly*, 6(2), 183–198.
- Kohler, K. J. (2006). What is emphasis and how is it coded? *Proceedings of Speech Prosody, Dresden, Germany*, 1–4.
- Kügler, F., & Baumann, S. (2019a). *Annotationsrichtlinien DIMA; Version 4.0* [Annotation guidelines DIMA]. Retrieved December 11, 2023, from <http://www.dima.de/>

- // dima.uni-koeln.de/wp-content/uploads/2021/03/DIMA-Annotationsrichtlinien.V4.0-26jun2020.pdf
- Kügler, F., & Baumann, S. (2019b). *Kurzanleitung Annotationsrichtlinien DIMA; Version 4.0* [Short instructions annotation guidelines DIMA]. Retrieved December 11, 2023, from http://dima.uni-koeln.de/wp-content/uploads/2019/02/DIMA-Annotationsrichtlinien.V4.0_Kurzanleitung.pdf
- Kügler, F., Baumann, S., Andreeva, B., Braun, B., Grice, M., Neitsch, J., Niebuhr, O., Peters, J., Röhr, C. T., Schweitzer, A., & Wagner, P. (2019). Annotation of German intonation: DIMA compared with other annotation systems. *Proceedings of the International Congress of Phonetic Sciences (ICPhS) 2019*, 1–4.
- Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., Jannedy, S., Michalsky, J., Niebuhr, O., & Peters, J. (2015). DIMA: Annotation guidelines for German intonation. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, 1–5.
- Kyncl, R., & Peyvan, M. (2017). *Streamponks: How YouTube and the new Creators are transforming our lives*. Virgin Books.
- Labov, W. (1972). Some principles of linguistic methodology. *Language in Society*, 1(1), 97–120.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, 145(12), 1604–1614. <https://doi.org/10.1037/xge0000223>
- Lee, S. (2017). Style-shifting in Vlogging: An acoustic analysis of “YouTube Voice”. *Lifespans and Styles*, 3(1), 28–39.
- Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., & Messerli, J. (2020). Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard*, 6(s3), 1–16.
- Lennström, D., Lindbom, T., & Nykänen, A. (2013). Prominence of tones in electric vehicle interior noise. *42nd International Congress and Exposition on Noise Control Engineering, Innsbruck, Austria*, 1, 508–515.
- Lenth, R. (2023). Package ‘emmeans’. Retrieved December 11, 2023, from <https://github.com/rvlenth/emmeans>
- Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.
- Lewandowska-Tomaszczyk, B., & Wilson, P. A. (2021). Expressive and reserved cultural linguistic schemas: British and American pride clusters. In M. Sadeghpour & F. Sharifian (Eds.), *Cultural Linguistics and World Englishes* (pp. 261–293). Springer.
- Lilly Singh. (2013, June 28). *Draw My Life | Superwoman* [Video]. YouTube. https://youtu.be/yfTV3UV_WIY?t=322
- Lilly Singh Vlogs. (2017, December 7). *We need to have an honest talk* [Video]. YouTube. <https://www.youtube.com/watch?v=KjK81YmQEuY>
- Limesurvey GmbH. (2017). *LimeSurvey: An Open Source survey tool* [Computer application, Version 3]. Retrieved December 11, 2023, from <http://www.limesurvey.org>

- Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Springer. https://doi.org/10.1007/978-94-009-2037-8_16
- Louise Pentland. (2017, March 8). *I'm SO Sorry | IWD2017* [Video]. YouTube. <https://www.youtube.com/watch?v=sezwogn4zOE>
- Lundholm Fors, K. (2015). *Production and perception of pauses in speech* [PhD thesis]. University of Gothenburg. Gothenburg.
- Markiplier. (2013, May 05). *Draw My Life - Markiplier* [Video]. YouTube. <https://youtu.be/6Sl-1X58ObY?t=709>
- Markiplier. (2018, August 19). *Let's Be Completely Honest* [Video]. YouTube. <https://www.youtube.com/watch?v=T6cdM1kubk4>
- Master, S., de Biase, N., Chiari, B. M., & Laukkanen, A.-M. (2008). Acoustic and perceptual analyses of Brazilian male actors' and nonactors' voices: Long-term Average Spectrum and the "Actor's Formant". *Journal of Voice*, 22(2), 146–154. <https://doi.org/10.1016/j.jvoice.2006.09.006>
- Master, S., Grigolletto de Biase, N., & Madureira, S. (2012). What about the "Actor's Formant" in actresses' voices? *Journal of Voice*, 26(3), 147–154. <https://doi.org/10.1016/j.jvoice.2010.10.011>
- Mayer, J. (2017). *Phonetische Analysen mit Praat: Ein Handbuch für Ein-und Umsteiger*. <http://praatpfanne.lingphon.net/>
- Merriam-Webster. (n.d.). YouTuber. In *Merriam-Webster.com dictionary*. Retrieved January 20, 2024, from <https://www.merriam-webster.com/dictionary/YouTuber>
- Michalsky, J., & Niebuhr, O. (2019). Myth busted? Challenging what we think we know about charismatic speech. *Acta Universitatis Carolinae Philologica*, 2, 27–56. <https://doi.org/10.14712/24646830.2019.17>
- Mileva, M., Tompkinson, J., Watt, D., & Burton, A. M. (2020). The role of face and voice cues in predicting the outcome of student representative elections. *Personality & Social Psychology Bulletin*, 46(4), 617–625. <https://doi.org/10.1177/0146167219867965>
- Mingione, D. (2014). Hello Internet! An analysis of YouTuber greetings. *Explorations in Linguistics: An Online Journal of Undergraduate Research (Saint Joseph's University, Philadelphia, Pennsylvania)*, 1(1), 19–34.
- Mixdorff, H., Niebuhr, O., & Hönemann, A. (2018). Model-based prosodic analysis of charismatic speech. *Proceedings of Speech Prosody 2018, Poznan, Poland*, 814–818.
- Morris, M., & Anderson, E. (2015). 'Charlie is so cool like': Authenticity, popularity and inclusive masculinity on YouTube. *Sociology*, 49(6), 1200–1217.
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6), 453–467. [https://doi.org/10.1016/0167-6393\(90\)90021-z](https://doi.org/10.1016/0167-6393(90)90021-z)
- Netcraft. (2023). *Most visited websites*. Retrieved December 11, 2023, from <https://trends.netcraft.com/topsites>
- Niebuhr, O. (2010). On the phonetics of intensifying emphasis in German. *Phonetica*, 67(3), 170–198.
- Niebuhr, O. (2017). Clear speech – mere speech? How segmental and prosodic speech reduction shape the impression that speakers create on listeners. *Proceedings of Interspeech 2017, Stockholm, Sweden*, 894–898.

- Niebuhr, O. (2020). "Space fighters" on stage – How the F1 and F2 vowel-space dimensions contribute to perceived speaker charisma. In A. Wendemuth, R. Böck, & I. Siegert (Eds.), *Elektronische Sprachverarbeitung 2020* (pp. 265–277, Vol. 95). TUDPress.
- Niebuhr, O., Brem, A., Michalsky, J., & Neitsch, J. (2020a). What makes business speakers sound charismatic? A contrastive acoustic-melodic analysis of Steve Jobs and Mark Zuckerberg. *Cadernos de Linguística e Teoria da Literatura*, 1(1), 1–40.
- Niebuhr, O., Brem, A., Novák-Tót, E., & Voße, J. (2016a). Prosodic constructions of charisma in business speeches – A contrastive acoustic-prosodic analysis of Steve Jobs and Mark Zuckerberg. *Proceedings of Speech Prosody 2016, Boston, Massachusetts*, 79–81.
- Niebuhr, O., Brem, A., & Tegtmeier, S. (2017). Advancing research and practice in entrepreneurship through speech analysis – From descriptive rhetorical terms to phonetically informed acoustic charisma profiles. *Journal of Speech Sciences*, 6(1), 3–26.
- Niebuhr, O., & Gonzalez, S. (2019). Do sound segments contribute to sounding charismatic? Evidence from a case study of Steve Jobs' and Mark Zuckerberg's vowel spaces. *International Journal of Acoustics and Vibration*, 24(2), 343–355.
- Niebuhr, O., & Michalsky, J. (2019). Computer-generated speaker charisma and its effects on human actions in a car-navigation system experiment – or how Steve Jobs' tone of voice can take you anywhere. In S. Misra, O. Gervasi, B. Murgante, E. Stankova, V. Korkhov, C. M. Torre, A. M. A. C. Rocha, D. Taniar, B. O. Apduhan, & E. Tarantino (Eds.), *Computational Science and Its Applications – ICCSA 2019: 19th International Conference, Saint Petersburg, Russia, July 1–4, 2019, Proceedings, Part II* (pp. 375–390). Springer.
- Niebuhr, O., Reetz, H., Barnes, J., & Yu, A. C. L. (2020b). Fundamental Aspects in the Perception of f0. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 1–17). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198832232.013.3>
- Niebuhr, O., & Skarnitzl, R. (2019). Measuring a speaker's acoustic correlates of pitch – But which? A contrastive analysis based on perceived speaker charisma. *Proceedings of the International Congress of Phonetic Sciences (ICPhS) 2019*, 1774–1778.
- Niebuhr, O., Skarnitzl, R., & Tylečková, L. (2018a). The acoustic fingerprint of a charismatic voice – Initial evidence from correlations between long-term spectral features and listener ratings. *Proceedings of Speech Prosody 2018, Poznan, Poland*, 359–363.
- Niebuhr, O., Tegtmeier, S., & Schweisfurth, T. (2019). Female speakers benefit more than male speakers from prosodic charisma training – A before-after analysis of 12-weeks and 4-h courses. *Frontiers in Communication*, 4:12, 1–6. <https://doi.org/10.3389/fcomm.2019.00012>
- Niebuhr, O., Thumm, J., & Michalsky, J. (2018b). Shapes and timing in charismatic speech – Evidence from sounds and melodies. *Proceedings of Speech Prosody 2018, Poznan, Poland*.
- Niebuhr, O., Voße, J., & Brem, A. (2016b). What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice. *Computers in Human Behavior*, 64, 366–382.

- Niebuhr, O., & Wrzeszcz, S. (2019). A woman's gotta do what a woman's gotta do, and a man's gotta say what a man's gotta say – Sex-specific differences in the production and perception of persuasive power. *Proceedings of the 13th International Pragmatics Association Conference*, 1–4.
- Novák-Tót, E. (2016). *Charisma and women – An investigation of the acoustic-melodic profiles of business women in the light of gender bias* [Doctoral dissertation]. Utrecht University.
- Novák-Tót, E., Niebuhr, O., & Chen, A. (2017). A gender bias in the acoustic-melodic features of charismatic speech? *Proceedings of Interspeech 2017, Stockholm, Sweden*, 2248–2252.
- Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F₀ of voice. *Phonetica*, 41(1), 1–16.
- Pauser, S., & Wagner, U. (2018). “The dose makes the poison”: Investigating the optimum level of a salesperson's charisma. *Marketing ZFP*, 40(1), 35–47. <https://doi.org/10.15358/0344-1369-2018-1-35>
- Peters, J. (2015). *Intonation*. Universitätsverlag Winter.
- Potts, J. (2009). *A history of charisma*. Palgrave Macmillan.
- Prolific. (2022). *Prolific* [Website, April to October 2022]. <https://www.prolific.co>
- Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35(3), 353–362.
- Quené, H., Boomsma, G., & van Erning, R. (2016). Attractiveness of male speakers: Effects of voice pitch and of speech tempo. *Proceedings of Speech Prosody 2016, Boston, Massachusetts*, 1086–1089.
- Rao, L. (2016). YouTube CEO says there's “no timetable” for profitability. *Fortune*, October 19, 2016. Retrieved December 12, 2023, from <https://fortune.com/2016/10/18/youtube-profits-ceo-susan-wojcicki/>
- Raun, T. (2018). Capitalizing intimacy. *Convergence: The International Journal of Research into New Media Technologies*, 24(1), 99–113. <https://doi.org/10.1177/1354856517736983>
- Reh, S., van Quaquebeke, N., & Giessner, S. R. (2017). The aura of charisma: A review on the embodiment perspective as signaling. *The Leadership Quarterly*, 28(4), 486–507.
- Riggio, R. E. (1987). *The charisma quotient: What it is, how to get it, how to use it*. Dodd Mead.
- Riggio, R. E. (1998). Charisma. *Encyclopedia of Mental Health*, 1, 387–396.
- Rosenberg, A. (2010). AuToBI: A tool for automatic ToBI annotation. *Proceedings of Interspeech 2010, Makuhari, Japan*, 1–4.
- Rosenberg, A., & Hirschberg, J. (2005). Acoustic/prosodic and lexical correlates of charismatic speech. *Proceedings of Interspeech 2005, Lisbon, Portugal*, 513–516.
- Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Communication*, 51(7), 640–655.
- Roux, D. (2008). Consumers faced with telephone selling: Metacognition, resistance and strategies. *ACR North American Advances in Consumer Research*, 35, 467–474.
- Roy, J., Cole, J., & Mahrt, T. (2017). Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology*, 8(1). <https://doi.org/10.5334/labphon.108>
- RStudio Team. (2023). *RStudio: Integrated development environment for R* [Computer application]. RStudio, PBC. Boston, MA. <http://www.posit.co/>

- Scherer, K. R. (1979). Personality markers in speech. In K. R. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 147–209). Cambridge University Press.
- Searle, G. D., & Hanrahan, S. J. (2011). Leading to inspire others: Charismatic influence or hard work? *Leadership & Organization Development Journal*, 32(7), 736–754.
- Selting, M. (1994). Emphatic speech style: with special focus on the prosodic signalling of heightened emotive involvement in conversation. *Journal of Pragmatics*, 22(3/4), 375–408. [https://doi.org/10.1016/0378-2166\(94\)90116-3](https://doi.org/10.1016/0378-2166(94)90116-3)
- Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J. R., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzluft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., ... Uhmann, S. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2) [discourse analytic transcription system 2]. *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion*, 10, 353–402.
- Shamir, B., House, R. J., & Arthur, M. B. (1993). The motivational effects of charismatic leadership: A self-concept based theory. *Organization Science*, 4(4), 577–594.
- Shepherd, J. (2023). 22 essential YouTube statistics you need to know in 2023. *The Social Shepherd*. Retrieved December 12, 2023, from <https://thesocialshepherd.com/blog/youtube-statistics>
- Siegert, I., & Niebuhr, O. (2021a). Case report: Women, be aware that your vocal charisma can dwindle in remote meetings. *Frontiers in Communication*, 5, 1–7. <https://doi.org/10.3389/fcomm.2020.611555>
- Siegert, I., & Niebuhr, O. (2021b). Speech signal compression deteriorates acoustic cues to perceived speaker charisma. *Elektronische Sprachsignalverarbeitung*, 1–10.
- Signorello, R., & Demolin, D. (2013). The physiological use of the charismatic voice in political speech. *Proceedings of Interspeech 2013, Lyon, France*, 987–991.
- Signorello, R., D'Errico, F., Poggi, I., & Demolin, D. (2012a). How charisma is perceived from speech: A multidimensional approach. *2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, Amsterdam, Netherlands*, 435–440.
- Signorello, R., D'Errico, F., Poggi, I., Demolin, D., & Mairano, P. (2012b). Charisma perception in political speech: A case study. *International Conference on Speech and Corpora (GSCP 2012), Belo Horizonte, Brazil*, 343–348.
- Skarnitzl, R., & Hledíková, H. (2022). Prosodic phrasing of good speakers in English and Czech. *Frontiers in Psychology*, 13, 1–13. <https://doi.org/10.3389/fpsyg.2022.857647>
- Su, Z. H., Patel, S., Bredemeyer, O., FitzGerald, J. J., & Antoniades, C. A. (2022). Parkinson's disease deficits in time perception to auditory as well as visual stimuli – A large online study. *Frontiers in Neuroscience*, 16, 1–11. <https://doi.org/10.3389/fnins.2022.995438>
- Syrdal, A. K. (1996). Acoustic variability in spontaneous conversational speech of American English talkers. *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP96), Philadelphia, Pennsylvania*, 1–4.
- The Film Theorists. (2020, May 28). *Film Theory: Hey Fallon, you're doing it wrong!* [Video]. YouTube. <https://www.youtube.com/watch?v=4Omnex5tyfk>
- The Game Theorists. (2013, September 25). *Draw My Life - Game Theory, MatPat, and YOU!* [Video]. YouTube. https://youtu.be/8mkuIP_i3js?t=375

- The Game Theorists. (2022, December 20). *Game Theory: This is not my channel* [Video]. YouTube. <https://www.youtube.com/watch?v=dJfE5U6gn9g>
- Tomlinson, J. M., & Fox Tree, J. E. (2011). Listeners' comprehension of uptalk in spontaneous speech. *Cognition*, 119(1), 58–69. <https://doi.org/10.1016/j.cognition.2010.12.005>
- Towler, A. (2003). Effects of charismatic influence training on attitudes, behavior, and performance. *Personnel Psychology*, 56(2), 363–381. <https://doi.org/10.1111/j.1744-6570.2003.tb00154.x>
- Tremblay, A., & Ransijn, J. (2020). *Package 'LMERConvenienceFunctions'*. Retrieved December 12, 2023, from <https://cran.r-project.org/web/packages/LMERConvenienceFunctions/index.html>
- Trouvain, J., & Barry, W. J. (2000). The prosody of excitement in horse race commentaries. *ITRW on Speech and Emotion, Newcastle, Northern Ireland*, 1–4.
- Trouvain, J., & Werner, R. (2022). A phonetic view on annotating speech pauses and pause-internal phonetic particles. In C. Schwarze & S. Grawunder (Eds.), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion: Konzepte, Probleme, Lösungen* (pp. 55–73). Narr Francke Attempto Verlag.
- Trudgill, P. (2000). *Sociolinguistics: An introduction to language and society*. Penguin.
- Tur, B., Harstad, J., & Antonakis, J. (2022). Effect of charismatic signaling in social media settings: Evidence from TED and Twitter. *The Leadership Quarterly*, 33(5), 1–17. <https://doi.org/10.1016/j.leaqua.2020.101476>
- Tylečková, L., Prokopová, Z., & Skarnitzl, R. (2017). The effect of voice quality on hiring decisions. *Acta Universitatis Carolinae: Philologica*, 3, 109–120.
- Tyler, J. C. (2015). Expanding and mapping the indexical field: Rising pitch, the uptalk stereotype, and perceptual variation. *Journal of English Linguistics*, 43(4), 284–310. <https://doi.org/10.1177/0075424215607061>
- Weber, M. (1968). *Economy and society: An outline of interpretive sociology*. Eds: G. Roth & C. Wittich; transl. E. Fischoff et al. Bedminster Press.
- Weiss, B., Trouvain, J., & Burkhardt, F. (2021). Acoustic correlates of likable speakers in the NSC database. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers* (pp. 245–262). https://doi.org/10.1007/978-981-15-6627-1_13
- Wen, X. (2018). Comments on YouTube product review videos. In L. Ha (Ed.), *The audience and business of YouTube and online videos* (pp. 73–84). Lexington Books.
- White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35(4), 501–522.
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2014). *Package 'ggplot2'*. <https://ggplot2.tidyverse.org/>
- Wicklin, R. (2023). Weak or strong? How to interpret a Spearman or Kendall correlation. *SAS Blogs*, April 5, 2023. Retrieved July 31, 2024, from <https://blogs.sas.com/content/iml/2023/04/05/interpret-spearman-kendall-corr.html>
- Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *The Journal of the Acoustical Society of America*, 127(3), 1559–1569.

- Williams, D. R., Zimprich, D. R., & Rast, P. (2019). A Bayesian nonlinear mixed-effects location scale model for learning. *Behavior Research Methods*, 51(5), 1968–1986. <https://doi.org/10.3758/s13428-019-01255-9>
- Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.
- Winter, B., Oh, G. E., Hübscher, I., Idemaru, K., Brown, L., Prieto, P., & Grawunder, S. (2021). Rethinking the frequency code: A meta-analytic review of the role of acoustic body size in communicative phenomena. *Philosophical Transactions of the Royal Society B*, 376, 1–13. <https://doi.org/10.1098/rstb.2020.0400>
- Xu, Y. (2013). ProsodyPro – A tool for large-scale systematic prosody analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France*, 7–10.
- Yamamoto, H. W., Kawahara, M., & Tanaka, A. (2021). A web-based auditory and visual emotion perception task experiment with children and a comparison of lab data and web data. *Frontiers in Psychology*, 12, 1–13. <https://doi.org/10.3389/fpsyg.2021.702106>
- Zellers, M. (2021). An overview of forms, functions, and configurations of backchannels in Ruruuli/Lunyala. *Journal of Pragmatics*, 175, 38–52. <https://doi.org/10.1016/j.pragma.2021.01.012>
- Zellers, M., & Schweitzer, A. (2017). An investigation of pitch matching across adjacent turns in a corpus of spontaneous German. *Proceedings of Interspeech 2017, Stockholm, Sweden*, 2336–2340. <https://doi.org/10.21437/interspeech.2017-811>
- Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America*, 149(6), 3910–3916. <https://doi.org/10.1121/10.0005132>
- Zhang, F. R. (2018). Sponsored videos on YouTube. In L. Ha (Ed.), *The audience and business of YouTube and online videos* (pp. 107–118). Lexington Books.
- Zhang, F. R., & Bi, N. C. (2018). YouTube and other social media. In L. Ha (Ed.), *The audience and business of YouTube and online videos* (pp. 85–96). Lexington Books.
- Zoe Sugg. (2018, January 29). *2018 plans & re-united with my bestie* [Video]. YouTube. <https://youtu.be/xUBSWu5SloY?si=MetTm6Zvd4RXRidB&t=155>
- Zoella. (2013, April 4). *Draw My Life | Zoella* [Video]. YouTube. <https://youtu.be/qx6fwery65M?t=219>
- Zoom Video Communications. (2021). *Zoom* [Computer application, versions 5.4.2-5.12.2]. Retrieved October 21, 2021, from <https://zoom.us/>

Part V

Appendix

Appendix A

YouTube statistics

Table A.1: The subscriber counts (= Subs) of the channels with the video, as well as the views and likes of the investigated video in absolute numbers, and normalized (nSubs, nViews, nLikes) to 100 days since creation of channel/upload of the video. The statistics were collected and calculated on November 22, 2022.

| | Subs | Views | Likes | nSubs | nViews | nLikes |
|-----------|------------|-----------|---------|------------|------------|-----------|
| LP | 2,220,000 | 426,827 | 28,283 | 47,324.66 | 20,471.32 | 1,356.50 |
| ZS | 4,930,000 | 2,437,447 | 58,911 | 132,136.16 | 138,648.86 | 3,351.02 |
| AD | 3,640,000 | 443,716 | 27,085 | 78,482.10 | 25,082.87 | 1,531.09 |
| DH | 6,190,000 | 3,730,832 | 395,774 | 105,182.67 | 199,830.32 | 21,198.39 |
| PL | 3,930,000 | 1,576,378 | 142,463 | 64,090.02 | 107,382.70 | 9,704.56 |
| CB | 8,710,000 | 1,716,709 | 76,238 | 146,190.00 | 96,066.54 | 4,266.26 |
| LS | 2,740,000 | 688,828 | 55,820 | 68,792.37 | 38,035.78 | 3,082.27 |
| SP | 2,960,000 | 273,811 | 10,937 | 112,718.96 | 20,494.84 | 818.64 |
| MP | 2,960,000 | 273,811 | 10,937 | 112,718.96 | 20,494.84 | 818.64 |
| MF | 34,000,000 | 2,236,574 | 195,511 | 887,265.14 | 143,738.69 | 12,564.97 |

Appendix B

Measurements of short stimuli

Table B.1: An overview of the acoustic measurements of the stimuli in this study for the four manipulated features pitch level, pitch range, speech rate, and final contour shape. The * indicates that this was the measurement of the unchanged stimulus. The + marks issues with the measurements. The *na* signifies that the value could not be measured. The – indicates that the stimulus was excluded from the experiment after a naturalness pilot.

| | | Speakers (ENG) | | | | |
|----------------------------|---------|---------------------|----------------------|--------|---------|----------------------|
| | | LP | ZS | AD | DH | PL |
| Pitch level (in Hz) | LF0 | 170.00 | – | 151.60 | 158.41 | 117.32 |
| | ORIG | 200.00 | 232.44 | 177.81 | 189.44 | 144.75 |
| | HF0 | – | – | 212.19 | 224.03 | 169.44 |
| Pitch range (in st) | LF0R | 12.18 | 6.94 | 15.53 | 7.45 | 15.08 |
| | ORIG | 15.92 | 10.40 | 17.49 | 16.07 | 19.84 |
| | HF0R | 16.09 | 11.31 | 19.60 | 26.56 | 23.46 |
| Speech rate (in syll/s) | LSR | 5.01* | 5.18* | – | 4.97* | 5.24 |
| | MSR | 6.01 | 6.18 | 5.98 | 5.97 | 6.24* |
| | HSR | 7.01 | 7.18 | 7.12* | 7.00 | 7.24 |
| Final contour (in st) | falling | -9.73 | <i>na</i> | -4.74* | -16.78 | -5.08* |
| | plateau | -0.06* | 1.89* | -0.49 | 0.25* | -0.70 |
| | rising | 2.05 | 3.68 | 1.62 | 2.03 | 6.77 |
| | | Speakers (NAM) | | | | |
| | | CB | LS | SP | MF | MP |
| Pitch level (in Hz) | LF0 | 166.85 | – | – | 106.36 | 135.15 |
| | ORIG | 197.82 | 213.59 | 185.91 | 119.95 | 152.86 |
| | HF0 | 235.10 | 252.15 | – | 142.93 | 183.64 |
| Pitch range (in st) | LF0R | 5.27 | 4.47 | 15.72 | 9.30 | 5.66 |
| | ORIG | 10.20 | 8.79 | 17.22 | 16.24 | 9.50 |
| | HF0R | 22.50 | 12.95 | 19.60 | – | 15.60 |
| Speech rate (in syll/s) | LSR | 5.44 | 4.86* | 4.83* | 5.06* | 4.53 |
| | MSR | 6.42 | – | 5.83 | 6.05 | 5.54* |
| | HSR | 7.58* | 6.87 | – | 7.07 | 6.54 |
| Final contour (in st) | falling | -2.49 | -2.29 | -5.60* | -11.70* | -11.85 |
| | plateau | 1.32 ⁺ * | -1.84 ⁺ * | 0.39 | -0.36 | -4.62 ⁺ * |
| | rising | 0.99 ⁺ | 4.68 | 4.56 | 14.78 | 3.53 |

Appendix C

Experiment file (example)

```
1 "ooTextFile"
2 "Collection" 6
3
4 "ExperimentMFC 7" "soundcheck"
5
6 blankWhilePlaying? <no>
7 stimuliAreSounds? <yes>
8 stimulusFileNameHead = "sounds/"
9 stimulusFileNameTail = ".wav"
10 stimulusCarrierBefore = ""
11 stimulusCarrierAfter = ""
12 stimulusInitialSilenceDuration = 0.5 seconds
13 stimulusMedialSilenceDuration = 0
14 stimulusFinalSilenceDuration = 0.5 seconds
15 numberOfDifferentStimuli = 1
16   "MF-MED_BR" ""
17
18 numberOfReplicationsPerStimulus = 1
19 breakAfterEvery = 0
20 randomize = <PermuteBalancedNoDoublets>
21 startText = "Before the experiment properly begins,
22
23 you can adjust the audio volume to a comfortable level.
24
25 You may play the audio up to three times.
26
27 Afterwards, the experiment will begin,
28 but you will have time to ask some questions.
29
30 Click to begin."
31
```

```
32 runText = "Please adjust the volume to a comfortable loudness."
33 pauseText = ""
34 endText = "The soundcheck is finished.
35
36 Click to start the experiment.
37
38 You will first see instructions
39 and have time to ask questions."
40
41 maximumNumberOfReplays = 3
42 replayButton = 0.2 0.45 0.4 0.6 "Play again!" ""
43 okButton = 0 0 0 0 "" ""
44 oopsButton = 0 0 0 0 "" ""
45 responsesAreSounds? <no> "" "" "" "" 0 0 0
46 numberOfDifferentResponses = 1
47   0.55 0.8 0.4 0.6 "Next!" 25 "" "next"
48 numberOfGoodnessCategories = 0
49
50 "ExperimentMFC 7" "exp11-likeable-list1"
51
52 blankWhilePlaying? <no>
53 stimuliAreSounds? <yes>
54 stimulusFileNameHead = "sounds/"
55 stimulusFileNameTail = ".wav"
56 stimulusCarrierBefore = ""
57 stimulusCarrierAfter = ""
58 stimulusInitialSilenceDuration = 0.5 seconds
59 stimulusMedialSilenceDuration = 0
60 stimulusFinalSilenceDuration = 0.5 seconds
61 numberOfDifferentStimuli = 27
62   "AD0086_ORIG" "The speaker's voice sounds ##likeable#."
63   "CB0040_ORIG" "The speaker's voice sounds ##likeable#."
64   "DH0114_ORIG" "The speaker's voice sounds ##likeable#."
65   "LP0110_ORIG" "The speaker's voice sounds ##likeable#."
66   "LS0035_ORIG" "The speaker's voice sounds ##likeable#."
67   "MF0021_ORIG" "The speaker's voice sounds ##likeable#."
68   "MP0081_ORIG" "The speaker's voice sounds ##likeable#."
69   "PL0015_ORIG" "The speaker's voice sounds ##likeable#."
70   "SP0001_ORIG" "The speaker's voice sounds ##likeable#."
```

71 "ZS0023_ORIG" "The speaker's voice sounds ##likeable#."
72
73 "AD0086_LF0-3st" "The speaker's voice sounds ##likeable#."
74 "CB0040_LF0-3st" "The speaker's voice sounds ##likeable#."
75 "DH0114_LF0-3st" "The speaker's voice sounds ##likeable#."
76 "LP0110_LF0-3st" "The speaker's voice sounds ##likeable#."
77 "MF0021_LF0-3st" "The speaker's voice sounds ##likeable#."
78 "MP0081_LF0-3st" "The speaker's voice sounds ##likeable#."
79 "PL0015_LF0-3st" "The speaker's voice sounds ##likeable#."
80
81 "AD0086_rising" "The speaker's voice sounds ##likeable#."
82 "CB0040_rising" "The speaker's voice sounds ##likeable#."
83 "DH0114_rising" "The speaker's voice sounds ##likeable#."
84 "LP0110_rising" "The speaker's voice sounds ##likeable#."
85 "LS0035_rising" "The speaker's voice sounds ##likeable#."
86 "MF0021_rising" "The speaker's voice sounds ##likeable#."
87 "MP0081_rising" "The speaker's voice sounds ##likeable#."
88 "PL0015_rising" "The speaker's voice sounds ##likeable#."
89 "SP0001_rising" "The speaker's voice sounds ##likeable#."
90 "ZS0023_rising" "The speaker's voice sounds ##likeable#."
91
92 numberOfReplicationsPerStimulus = 1
93 breakAfterEvery = 0
94 randomize = <PermuteBalancedNoDoublets>
95 startText = "This is PART I of the experiment."
96
97 You will hear short excerpts from several speakers
98 and rate the voices on different statements.
99
100 First, rate the voices in terms of this statement:
101
102 The speaker's voice sounds ##likeable#.
103
104 In case you could not hear the audio
105 (e.g., because it did not play),
106 please choose the ""Could not hear the audio"" button.
107
108 You may ask questions now.
109
110 Otherwise, click to begin."

```
111 runText = "The speaker's voice sounds ##likeable#."
112 pauseText = ""
113 endText = "Click to continue."
114 maximumNumberOfReplays = 0
115 replayButton = 0 0 0 0 "" ""
116 okButton = 0 0 0 0 "" ""
117 oopsButton = 0 0 0 0 "" ""
118 responsesAreSounds? <no> "" "" "" "" 0 0 0
119 numberOfDifferentResponses = 6
120     0.02 0.21 0.3 0.6 "strongly agree" 25 "" "5"
121     0.21 0.38 0.3 0.6 "agree" 25 "" "4"
122     0.38 0.61 0.3 0.6 "neither agree nor disagree" 20 "" "3"
123     0.61 0.78 0.3 0.6 "disagree" 25 "" "2"
124     0.78 0.98 0.3 0.6 "strongly disagree" 25 "" "1"
125     0.38 0.61 0.15 0.25 "Could not hear the audio" 20 "" "no audio"
126 numberOfGoodnessCategories = 0
127
128 "ExperimentMFC 7" "exp11-persuasive-list1"
129
130 blankWhilePlaying? <no>
131 stimuliAreSounds? <yes>
132 stimulusFileNameHead = "sounds/"
133 stimulusFileNameTail = ".wav"
134 stimulusCarrierBefore = ""
135 stimulusCarrierAfter = ""
136 stimulusInitialSilenceDuration = 0.5 seconds
137 stimulusMedialSilenceDuration = 0
138 stimulusFinalSilenceDuration = 0.5 seconds
139 numberOfDifferentStimuli = 27
140     "AD0086_ORIG" "The speaker's voice sounds ##persuasive#."
141     "CB0040_ORIG" "The speaker's voice sounds ##persuasive#."
142     "DH0114_ORIG" "The speaker's voice sounds ##persuasive#."
143     "LP0110_ORIG" "The speaker's voice sounds ##persuasive#."
144     "LS0035_ORIG" "The speaker's voice sounds ##persuasive#."
145     "MF0021_ORIG" "The speaker's voice sounds ##persuasive#."
146     "MP0081_ORIG" "The speaker's voice sounds ##persuasive#."
147     "PL0015_ORIG" "The speaker's voice sounds ##persuasive#."
148     "SP0001_ORIG" "The speaker's voice sounds ##persuasive#."
149     "ZS0023_ORIG" "The speaker's voice sounds ##persuasive#."
150
```

```

151 "AD0086_HF0" "The speaker's voice sounds ##persuasive#."
152 "CB0040_HF0" "The speaker's voice sounds ##persuasive#."
153 "DH0114_HF0" "The speaker's voice sounds ##persuasive#."
154 "LS0035_HF0" "The speaker's voice sounds ##persuasive#."
155 "MF0021_HF0" "The speaker's voice sounds ##persuasive#."
156 "MP0081_HF0" "The speaker's voice sounds ##persuasive#."
157 "PL0015_HF0" "The speaker's voice sounds ##persuasive#."
158
159 "AD0086_plateau" "The speaker's voice sounds ##persuasive#."
160 "CB0040_falling" "The speaker's voice sounds ##persuasive#."
161 "DH0114_falling" "The speaker's voice sounds ##persuasive#."
162 "LP0110_falling" "The speaker's voice sounds ##persuasive#."
163 "LS0035_falling" "The speaker's voice sounds ##persuasive#."
164 "MF0021_plateau" "The speaker's voice sounds ##persuasive#."
165 "MP0081_falling" "The speaker's voice sounds ##persuasive#."
166 "PL0015_plateau" "The speaker's voice sounds ##persuasive#."
167 "SP0001_plateau" "The speaker's voice sounds ##persuasive#."
168 "ZS0023_falling" "The speaker's voice sounds ##persuasive#."
169
170 numberOfReplicationsPerStimulus = 1
171 breakAfterEvery = 0
172 randomize = <PermuteBalancedNoDoublets>
173 startText = "Rate the voices in terms of this statement:
174
175 The speaker's voice sounds ##persuasive#.
176
177 Click to continue."
178 runText = "The speaker's voice sounds ##persuasive#."
179 pauseText = ""
180 endText = "Click to continue."
181 maximumNumberOfReplays = 0
182 replayButton = 0 0 0 0 "" ""
183 okButton = 0 0 0 0 "" ""
184 oopsButton = 0 0 0 0 "" ""
185 responsesAreSounds? <no> "" "" "" "" 0 0 0
186 numberOfDifferentResponses = 6
187 0.02 0.21 0.3 0.6 "strongly agree" 25 "" "5"
188 0.21 0.38 0.3 0.6 "agree" 25 "" "4"
189 0.38 0.61 0.3 0.6 "neither agree nor disagree" 20 "" "3"
190 0.61 0.78 0.3 0.6 "disagree" 25 "" "2"

```

Appendix C Experiment file (example)

```
191 0.78 0.98 0.3 0.6 "strongly disagree" 25 "" "1"
192 0.38 0.61 0.15 0.25 "Could not hear the audio" 20 "" "no audio"
193 numberOfGoodnessCategories = 0
194
195 "ExperimentMFC 7" "exp11-enthusiastic-list1"
196
197 blankWhilePlaying? <no>
198 stimuliAreSounds? <yes>
199 stimulusFileNameHead = "sounds/"
200 stimulusFileNameTail = ".wav"
201 stimulusCarrierBefore = ""
202 stimulusCarrierAfter = ""
203 stimulusInitialSilenceDuration = 0.5 seconds
204 stimulusMedialSilenceDuration = 0
205 stimulusFinalSilenceDuration = 0.5 seconds
206 numberOfDifferentStimuli = 28
207 "AD0086_ORIG" "The speaker's voice sounds ##enthusiastic#."
208 "CB0040_ORIG" "The speaker's voice sounds ##enthusiastic#."
209 "DH0114_ORIG" "The speaker's voice sounds ##enthusiastic#."
210 "LP0110_ORIG" "The speaker's voice sounds ##enthusiastic#."
211 "LS0035_ORIG" "The speaker's voice sounds ##enthusiastic#."
212 "MF0021_ORIG" "The speaker's voice sounds ##enthusiastic#."
213 "MP0081_ORIG" "The speaker's voice sounds ##enthusiastic#."
214 "PL0015_ORIG" "The speaker's voice sounds ##enthusiastic#."
215 "SP0001_ORIG" "The speaker's voice sounds ##enthusiastic#."
216 "ZS0023_ORIG" "The speaker's voice sounds ##enthusiastic#."
217
218 "AD0086_LF0R" "The speaker's voice sounds ##enthusiastic#."
219 "CB0040_LF0R" "The speaker's voice sounds ##enthusiastic#."
220 "DH0114_LF0R" "The speaker's voice sounds ##enthusiastic#."
221 "LP0110_LF0R" "The speaker's voice sounds ##enthusiastic#."
222 "LS0035_LF0R" "The speaker's voice sounds ##enthusiastic#."
223 "MF0021_LF0R" "The speaker's voice sounds ##enthusiastic#."
224 "MP0081_LF0R" "The speaker's voice sounds ##enthusiastic#."
225 "PL0015_LF0R" "The speaker's voice sounds ##enthusiastic#."
226 "SP0001_LF0R" "The speaker's voice sounds ##enthusiastic#."
227 "ZS0023_LF0R" "The speaker's voice sounds ##enthusiastic#."
228
229 "CB0040_LSR" "The speaker's voice sounds ##enthusiastic#."
230 "DH0114_HSR" "The speaker's voice sounds ##enthusiastic#."
```

```

231 "LP0110_HSR" "The speaker's voice sounds ##enthusiastic#."
232 "LS0035_HSR" "The speaker's voice sounds ##enthusiastic#."
233 "MF0021_HSR" "The speaker's voice sounds ##enthusiastic#."
234 "MP0081_HSR" "The speaker's voice sounds ##enthusiastic#."
235 "PL0015_HSR" "The speaker's voice sounds ##enthusiastic#."
236 "ZS0023_HSR" "The speaker's voice sounds ##enthusiastic#."
237
238 numberOfReplicationsPerStimulus = 1
239 breakAfterEvery = 0
240 randomize = <PermuteBalancedNoDoublets>
241 startText = "Rate the voices in terms of this statement:
242
243 The speaker's voice sounds ##enthusiastic#.
244
245 Click to continue."
246 runText = "The speaker's voice sounds ##enthusiastic#."
247 pauseText = ""
248 endText = "Click to continue."
249 maximumNumberOfReplays = 0
250 replayButton = 0 0 0 0 "" ""
251 okButton = 0 0 0 0 "" ""
252 oopsButton = 0 0 0 0 "" ""
253 responsesAreSounds? <no> "" "" "" "" 0 0 0
254 numberOfDifferentResponses = 6
255 0.02 0.21 0.3 0.6 "strongly agree" 25 "" "5"
256 0.21 0.38 0.3 0.6 "agree" 25 "" "4"
257 0.38 0.61 0.3 0.6 "neither agree nor disagree" 20 "" "3"
258 0.61 0.78 0.3 0.6 "disagree" 25 "" "2"
259 0.78 0.98 0.3 0.6 "strongly disagree" 25 "" "1"
260 0.38 0.61 0.15 0.25 "Could not hear the audio" 20 "" "no audio"
261 numberOfGoodnessCategories = 0
262
263 "ExperimentMFC 7" "exp11-authentic-list1"
264
265 blankWhilePlaying? <no>
266 stimuliAreSounds? <yes>
267 stimulusFileNameHead = "sounds/"
268 stimulusFileNameTail = ".wav"
269 stimulusCarrierBefore = ""
270 stimulusCarrierAfter = ""

```

```
271 stimulusInitialSilenceDuration = 0.5 seconds
272 stimulusMedialSilenceDuration = 0
273 stimulusFinalSilenceDuration = 0.5 seconds
274 numberOfDifferentStimuli = 28
275 "AD0086_ORIG" "The speaker's voice sounds ##authentic#."
276 "CB0040_ORIG" "The speaker's voice sounds ##authentic#."
277 "DH0114_ORIG" "The speaker's voice sounds ##authentic#."
278 "LP0110_ORIG" "The speaker's voice sounds ##authentic#."
279 "LS0035_ORIG" "The speaker's voice sounds ##authentic#."
280 "MF0021_ORIG" "The speaker's voice sounds ##authentic#."
281 "MP0081_ORIG" "The speaker's voice sounds ##authentic#."
282 "PL0015_ORIG" "The speaker's voice sounds ##authentic#."
283 "SP0001_ORIG" "The speaker's voice sounds ##authentic#."
284 "ZS0023_ORIG" "The speaker's voice sounds ##authentic#."
285
286 "AD0086_HF0R" "The speaker's voice sounds ##authentic#."
287 "CB0040_HF0R" "The speaker's voice sounds ##authentic#."
288 "DH0114_HF0R" "The speaker's voice sounds ##authentic#."
289 "LP0110_HF0R" "The speaker's voice sounds ##authentic#."
290 "LS0035_HF0R" "The speaker's voice sounds ##authentic#."
291 "MP0081_HF0R" "The speaker's voice sounds ##authentic#."
292 "PL0015_HF0R" "The speaker's voice sounds ##authentic#."
293 "SP0001_HF0R" "The speaker's voice sounds ##authentic#."
294 "ZS0023_HF0R" "The speaker's voice sounds ##authentic#."
295
296 "AD0086_MSR" "The speaker's voice sounds ##authentic#."
297 "CB0040_MSR" "The speaker's voice sounds ##authentic#."
298 "DH0114_MSR" "The speaker's voice sounds ##authentic#."
299 "LP0110_MSR" "The speaker's voice sounds ##authentic#."
300 "MF0021_MSR" "The speaker's voice sounds ##authentic#."
301 "MP0081_LSR" "The speaker's voice sounds ##authentic#."
302 "PL0015_LSR" "The speaker's voice sounds ##authentic#."
303 "SP0001_MSR" "The speaker's voice sounds ##authentic#."
304 "ZS0023_MSR" "The speaker's voice sounds ##authentic#."
305
306 numberOfReplicationsPerStimulus = 1
307 breakAfterEvery = 0
308 randomize = <PermuteBalancedNoDoublets>
309 startText = "Rate the voices in terms of this statement:"
310
```

```

311 The speaker's voice sounds ##authentic#.
312
313 Click to continue."
314 runText = "The speaker's voice sounds ##authentic#."
315 pauseText = ""
316 endText = "PART I of the experiment is finished.
317
318 Click to begin PART II
319 when you are ready."
320 maximumNumberOfReplays = 0
321 replayButton = 0 0 0 0 "" ""
322 okButton = 0 0 0 0 "" ""
323 oopsButton = 0 0 0 0 "" ""
324 responsesAreSounds? <no> "" "" "" "" 0 0 0
325 numberOfDifferentResponses = 6
326     0.02 0.21 0.3 0.6 "strongly agree" 25 "" "5"
327     0.21 0.38 0.3 0.6 "agree" 25 "" "4"
328     0.38 0.61 0.3 0.6 "neither agree nor disagree" 20 "" "3"
329     0.61 0.78 0.3 0.6 "disagree" 25 "" "2"
330     0.78 0.98 0.3 0.6 "strongly disagree" 25 "" "1"
331     0.38 0.61 0.15 0.25 "Could not hear the audio" 20 "" "no audio"
332 numberOfGoodnessCategories = 0
333
334 "ExperimentMFC 6" "exp12-charisma+familiarity"
335
336 blankWhilePlaying? <no>
337 stimuliAreSounds? <yes>
338 stimulusFileNameHead = "sounds/"
339 stimulusFileNameTail = ".wav"
340 stimulusCarrierBefore = ""
341 stimulusCarrierAfter = ""
342 stimulusInitialSilenceDuration = 0.5 seconds
343 stimulusMedialSilenceDuration = 0 seconds
344 stimulusFinalSilenceDuration = 0.5 seconds
345 numberOfDifferentStimuli = 80
346     "AD-CUT" ""
347     "AD-LONG_BR" ""
348     "AD-LONG_NBR" ""
349     "AD-MED_BR" ""
350     "AD-MED_NBR" ""

```

351 "AD-MIX_BR" ""
352 "AD-MIX_L" ""
353 "AD-SHORT" ""
354
355 "CB-CUT" ""
356 "CB-LONG_BR" ""
357 "CB-LONG_NBR" ""
358 "CB-MED_BR" ""
359 "CB-MED_NBR" ""
360 "CB-MIX_BR" ""
361 "CB-MIX_L" ""
362 "CB-SHORT" ""
363
364 "DH-CUT" ""
365 "DH-LONG_BR" ""
366 "DH-LONG_NBR" ""
367 "DH-MED_BR" ""
368 "DH-MED_NBR" ""
369 "DH-MIX_BR" ""
370 "DH-MIX_L" ""
371 "DH-SHORT" ""
372
373 "LP-CUT" ""
374 "LP-LONG_BR" ""
375 "LP-LONG_NBR" ""
376 "LP-MED_BR" ""
377 "LP-MED_NBR" ""
378 "LP-MIX_BR" ""
379 "LP-MIX_L" ""
380 "LP-SHORT" ""
381
382 "LS-CUT" ""
383 "LS-LONG_BR" ""
384 "LS-LONG_NBR" ""
385 "LS-MED_BR" ""
386 "LS-MED_NBR" ""
387 "LS-MIX_BR" ""
388 "LS-MIX_L" ""
389 "LS-SHORT" ""
390

391 "MF-CUT" ""
392 "MF-LONG_BR" ""
393 "MF-LONG_NBR" ""
394 "MF-MED_BR" ""
395 "MF-MED_NBR" ""
396 "MF-MIX_BR" ""
397 "MF-MIX_L" ""
398 "MF-SHORT" ""
399
400 "MP-CUT" ""
401 "MP-LONG_BR" ""
402 "MP-LONG_NBR" ""
403 "MP-MED_BR" ""
404 "MP-MED_NBR" ""
405 "MP-MIX_BR" ""
406 "MP-MIX_L" ""
407 "MP-SHORT" ""
408
409 "PL-CUT" ""
410 "PL-LONG_BR" ""
411 "PL-LONG_NBR" ""
412 "PL-MED_BR" ""
413 "PL-MED_NBR" ""
414 "PL-MIX_BR" ""
415 "PL-MIX_L" ""
416 "PL-SHORT" ""
417
418 "SP-CUT" ""
419 "SP-LONG_BR" ""
420 "SP-LONG_NBR" ""
421 "SP-MED_BR" ""
422 "SP-MED_NBR" ""
423 "SP-MIX_BR" ""
424 "SP-MIX_L" ""
425 "SP-SHORT" ""
426
427 "ZS-CUT" ""
428 "ZS-LONG_BR" ""
429 "ZS-LONG_NBR" ""
430 "ZS-MED_BR_M" ""

431 "ZS-MED_NBR" ""
432 "ZS-MIX_BR_O" ""
433 "ZS-MIX_L" ""
434 "ZS-SHORT" ""
435
436 numberOfReplicationsPerStimulus = 1
437 breakAfterEvery = 0
438 randomize = <PermuteAll>
439 startText = "This is the second and final part of the experiment.
440
441 You will hear the same voices from PART I of the experiment once more,
442 but they talk for a longer time.
443
444 Please rate the voices in terms of how ##charismatic# they sound
445 and how ##familiar# the voice is to you from BEFORE THE EXPERIMENT.
446 Do not rate a speaker as familiar because you have heard them in the experiment
447 already.
448
449 In case you could not hear the audio (e.g., because it did not play),
450 please choose the "Could not hear the audio" button.
451
452 You may ask questions now.
453
454 Otherwise, click to begin."
455 runText = "##The speaker sounds charismatic.#
456 Rate the voice on the scale below from
457 "strongly agree" to "strongly disagree".
458
459
460
461
462
463
464
465
466
467
468
469
470 Afterwards, indicate how ##familiar# the speaker is to you

```

471 from before the experiment session."
472 pauseText = ""
473 endText = "The experiment is finished.
474
475 Please DO NOT close this window.
476
477 Indicate to the experimenter that you are finished
478 so we can continue.
479
480 Thank you for participating!"
481 maximumNumberOfReplays = 0
482 replayButton = 0 0 0 0 "" ""
483 okButton = 0 0 0 0 "" ""
484 oopsButton = 0 0 0 0 "" ""
485 responsesAreSounds? <no>
486 responseFileNameHead = "/"
487 responseFileNameTail = ""
488 responseCarrierBefore = ""
489 responseCarrierAfter = ""
490 responseInitialSilenceDuration = 0 seconds
491 responseMedialSilenceDuration = 0 seconds
492 responseFinalSilenceDuration = 0 seconds
493 numberOfDifferentResponses = 6
494 0.02 0.21 0.6 0.8 "strongly agree" 25 "" "5"
495 0.21 0.38 0.6 0.8 "agree" 25 "" "4"
496 0.38 0.61 0.6 0.8 "neither agree nor disagree" 20 "" "3"
497 0.61 0.78 0.6 0.8 "disagree" 25 "" "2"
498 0.78 0.98 0.6 0.8 "strongly disagree" 25 "" "1"
499 0.38 0.61 0.48 0.58 "Could not hear the audio" 20 "" "no audio"
500 numberOfGoodnessCategories = 5
501 0.02 0.28 0.18 0.28 "I know the speaker"
502 0.28 0.50 0.18 0.28 "They seem familiar"
503 0.50 0.72 0.18 0.28 "Unsure"
504 0.72 0.98 0.18 0.28 "I do not know the speaker"
505 0.38 0.61 0.05 0.15 "Could not hear the audio"
506
507 # I know the speaker = 1, They seem familiar = 2, Unsure = 3, I do not know the
508 speaker = 4, No audio = 5

```


Appendix D

Experiment instructions

NB: The participants were given instructions to read before the experiment as well. They were basically the same as the ones printed here and used for talking through the instructions, but less detailed.

Thank you for participating in my experiment! My name is Stephanie. I am a PhD student in Kiel, Germany, and this experiment is part of my research for my PhD thesis. As you read before, I investigate how speakers' voices are perceived by listeners. The speakers you will hear are social media personalities. You will hear different speakers (repeatedly) and you will be asked to rate their voices in terms of different attributes. Let's talk through the instructions. Please feel free to interrupt me if you have questions.

Try to pay attention to the VOICE while rating and disregard the content as much as possible. I know that won't always be possible, just try your best.

During the experiment you can mute yourself. You can also decide if you want to have your camera on or off. I will have my camera turned on but move off camera during the experiment and be muted (*NB: camera turned off if that is better for the internet connection*). I will not watch you make your decisions. I will be able to hear you, though, so you can ask questions and indicate to me when you are finished. Once we have talked through the instructions, I will share my screen with you and give you the ability to control that window so you can click through the experiment at your own pace.

If you at some point lose the connection to the Zoom call and cannot re-join, please message me (*NB: email or on Prolific*) and we will find a solution (so long as you still have an internet connection).

There are two parts to the experiment. Before, you will be able to check the audio and adjust the volume to a comfortable level. You can ask questions before the experiment and between the different parts (at points that the experiment will tell you), but not during each experiment part. Your responses will not be timed. Of course if you run into issues you can always talk to me.

Like I mentioned, you will rate voices in terms of different attributes or state-

ments. You will rate the voices on one statement at a time. The statements are constructed as “The speaker’s voice sounds...” and then one of four attributes. You will make all of your judgments on one attribute and then switch to another attribute and so on. A grey screen will inform you that the attribute you are rating is changing. The different response options are: “strongly agree”, “agree”, “neither agree nor disagree”, “disagree”, and “strongly disagree” and always in that order. Once you click on an answer, you will not be able to go back and make a correction.

If you lose the connection briefly and didn’t hear the audio or if the audio was not transmitted completely, please use the “Could not hear the audio” button that is below the other response options and do not just click anything.

Group 1

Experiment Part 1: You will hear several short utterances and rate the VOICES in terms of four attributes: likable (= friendly, approachable), authentic (= sincere, not putting up an act), persuasive (= convincing), enthusiastic (= engaging, lively). You will hear the voices and utterances several times, but they will always be slightly different. Please try and rate each utterance on its own and try not to compare too much. This might not always be possible, just try your best. Also: Please listen to the entire recording before you click your response!

Experiment Part 2: You will hear the same voices again, but they will talk for a longer time. Please listen until the end of each recording before clicking the first response! This is absolutely important! You will not be timed. If it happens by accident sometimes it is no big deal, but I might speak to you if it happens frequently. The audio excerpts are between 5 and 20 seconds long. You will hear the voices and utterances several times again, but they will always be slightly different. Nevertheless, you will be able to estimate how long you have to wait before clicking. You will rate the voices in terms of your more general impression of the speaker. This “general impression” is specified on screen before the experiment part begins, so you can ask questions then if you have any. After the general rating, you will also indicate how familiar you are with the speakers, choosing between “I know the speaker”, “They seem familiar”, “Unsure”, and “I do not know the speaker”. Both ratings happen in the same go, just one after the other. Again, it is important to listen to the entire audio before clicking the first response.

Group 2

Experiment Part 1: You will hear several utterances and rate the VOICES in terms of four attributes: likable (= friendly, approachable), authentic (= sincere, not putting up an act), persuasive (= convincing), enthusiastic (= engaging, lively).

Please listen until the end of each recording before rating the speakers' voices! This is absolutely important! You will not be timed. If it happens by accident sometimes it is no big deal, but I might speak to you if it happens frequently. The audio excerpts are between 5 and 20 seconds long. You will hear the voices and utterances several times, but they will always be slightly different. Nevertheless, you will be able to estimate how long you have to wait before clicking. Please try and rate each utterance on its own and try not to compare too much. This might not always be possible, just try your best.

Experiment Part 2: You will hear the same voices again, but they will talk for a shorter time. You will hear the voices and utterances several times again, but they will always be slightly different. You will rate the voices in terms of your more general impression of the speaker. This "general impression" is specified on screen before the experiment part begins, so you can ask questions then if you have any. After the general rating, you will also indicate how familiar you are with the speakers, choosing between "I know the speaker", "They seem familiar", "Unsure", and "I do not know the speaker". Both ratings happen in the same go, just one after the other. Again, it is important to listen to the entire audio before clicking the first response.

Do you have more questions?

So, I will share my screen and give you remote control. In a moment you can click once to take control. Please do not minimize the window because it is difficult to get back up. Once you start with the sound-check, please let me know if you hear the audio. Keep in mind: the experiment might need a moment to react to the click because of the remote control and possible lag, so again: take your time!

You can start whenever you are ready!

Debriefing:

- Do you have questions or comments?
- For participants between November 2020 and June 2021: I will donate £1 to PAPYRUS after the experiment sessions are finished.
- For participants between April and October 2022: I will send you another study on Prolific at the end of the week which includes the £7.50 compensation for this interview.

Appendix E

Linear mixed model codes (Perception

1)

Table E.1: LMM codes for the analyses of the pitch level manipulations and their effect on the different rating attributes (= Attr.) *authentic* (AU), *enthusiastic* (EN), *likable* (LI), *persuasive* (PE), and *charismatic* (CH). The ° marks models that showed singular fit, despite testing different optimizers. In some cases, the model did not converge before changing the optimizer, but still returned a singular fit.

| Pitch level | | |
|-------------|--|---|
| Attr. | Model structure | |
| AU | <code>lmer(response ~ Manipulation + Origin + Manipulation:Origin + (1 Speaker) + (1 + Origin Participant), data, control = lmerControl(optimizer = "nmkbw"))</code> | ° |
| EN | <code>lmer(response ~ Manipulation + Origin + Manipulation:Origin + (1 Speaker) + (1 + Origin Participant), data, control = lmerControl(optimizer = "bobyqa"))</code> | ° |
| LI | <code>lmer(response ~ Manipulation + Origin + Manipulation:Origin + (1 Speaker) + (1 + Origin Participant), data)</code> | • |
| PE | <code>lmer(response ~ Manipulation + Origin + Manipulation:Origin + (1 Speaker) + (1 + Origin Participant), data)</code> | • |
| CH | <code>lmer(response ~ Manipulation + Gender + Origin + Familiarity + Manipulation:Gender + Manipulation:Origin + Manipulation:Familiarity + Familiarity:Gender + Familiarity:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant) + (1 + Familiarity Participant), data, control = lmerControl(optimizer = "bobyqa"))</code> | ° |

Table E.2: LMM codes for the analyses of the pitch range manipulations and their effect on the different rating attributes (= Attr.) *authentic* (AU), *enthusiastic* (EN), *likable* (LI), *persuasive* (PE), and *charismatic* (CH). The ° marks models that showed singular fit, despite testing different optimizers. In some cases, the model did not converge before changing the optimizer, but still returned a singular fit.

| Pitch range | | |
|-------------|--|---|
| Attr. | Model structure | |
| AU | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data, control = lmerControl(optimizer = "nloptwrap", optCtrl = list(algorithm = "NLOPT_LN_NELDERMEAD")))</code> | • |
| EN | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | ° |
| LI | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | ° |
| PE | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | |
| CH | <code>lmer(response ~ Manipulation + Gender + Origin + Familiarity + Manipulation:Gender + Manipulation:Origin + Manipulation:Familiarity + Familiarity:Gender + Familiarity:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant) + (1 + Familiarity Participant), data, control = lmerControl(optimizer = "bobyqa"))</code> | ° |

Table E.3: LMM codes for the analyses of the final contour direction manipulations and their effect on the different rating attributes (= Attr.) *authentic* (AU), *enthusiastic* (EN), *likable* (LI), *persuasive* (PE), and *charismatic* (CH). The ° marks models that showed singular fit, despite testing different optimizers. In some cases, the model did not converge before changing the optimizer, but still returned a singular fit.

| Final contour direction | | |
|-------------------------|---|---|
| Attr. | Model structure | |
| AU | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data, control = lmerControl(optimizer = "bobyqa"))</code> | ° |
| EN | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data, control = lmerControl(optimizer = "nloptwrap", optCtrl = list(algorithm = "NLOPT_LN_NELDERMEAD")))</code> | |
| LI | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data, control = lmerControl(optimizer = "Nelder_Mead"))</code> | |
| PE | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | |
| CH | <code>lmer(response ~ Manipulation + Gender + Origin + Familiarity + Manipulation:Gender + Manipulation:Origin + Manipulation:Familiarity + Familiarity:Gender + Familiarity:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant) + (1 + Familiarity Participant), data)</code> | ° |

Table E.4: LMM code for the analyses of the speech rate manipulations and their effect on the different rating attributes (= Attr.) *authentic* (AU), *enthusiastic* (EN), *likable* (LI), *persuasive* (PE), and *charismatic* (CH). The ° marks models that showed singular fit, despite testing different optimizers. In some cases, the model did not converge before changing the optimizer, but still returned a singular fit.

| Speech rate | |
|-------------|---|
| Attr. | Model structure |
| AU | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> |
| EN | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> ° |
| LI | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> ° |
| PE | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data, control = lmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B")))</code> |
| CH | <code>lmer(response ~ Manipulation + Gender + Origin + Familiarity + Manipulation:Gender + Manipulation:Origin + Manipulation:Familiarity + Familiarity:Gender + Familiarity:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant) + (1 + Familiarity Participant), data, control = lmerControl(optimizer = "nmlme"))</code> ° |

Appendix F

Outputs: Linear mixed models (Perception 1)

Table F.1: The output of the LMM analysis for pitch level and the *likable* ratings. The intercept is the ORIG stimulus for male speakers from England. The independent variables are *Manipulation* (three levels: ORIG, LF0, HF0), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|-------------|-------|------|---------|-----------|---|
| (Intercept) | 2.95 | 0.39 | 7.52 | .002 | * |
| LF0 | 0.18 | 0.31 | 0.59 | .559 | |
| HF0 | -0.06 | 0.32 | -0.19 | .848 | |
| NAM | 0.33 | 0.62 | 0.53 | .625 | |
| LF0 × NAM | -0.57 | 0.50 | -1.15 | .252 | |
| HF0 × NAM | -0.09 | 0.50 | -0.17 | .861 | |

Signif. codes: * .05, . 1

Table F.2: The output of the LMM analysis for pitch level and the *persuasive* ratings. The intercept is the ORIG stimulus for male speakers from England. The independent variables are *Manipulation* (three levels: ORIG, LF0, HF0), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|-------------|-------|------|---------|-----------|---|
| (Intercept) | 2.87 | 0.42 | 6.76 | .003 | * |
| LF0 | 0.16 | 0.30 | 0.54 | .590 | |
| HF0 | -0.07 | 0.30 | -0.22 | .824 | |
| NAM | 0.71 | 0.67 | 1.06 | .353 | |
| LF0 × NAM | -0.00 | 0.48 | -0.00 | .998 | |
| HF0 × NAM | -0.32 | 0.48 | -0.68 | .500 | |

Signif. codes: * .05, . 1

Table F.3: The output of the LMM analysis for pitch range and the *authentic* ratings. The intercept is the ORIG stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: ORIG, LF0R, HF0R), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) |
|-------------|-------|------|---------|-----------|
| (Intercept) | 3.54 | 0.19 | 19.12 | < .001 * |
| LF0R | -0.25 | 0.34 | -0.73 | .466 |
| HF0R | -0.05 | 0.35 | -0.15 | .884 |
| male | -0.24 | 0.18 | -1.29 | .219 |
| NAM | -0.29 | 0.23 | -1.27 | .220 |
| LF0R × male | 0.22 | 0.35 | 0.64 | .520 |
| HF0R × male | -0.02 | 0.38 | -0.06 | .950 |
| LF0R × NAM | 0.06 | 0.38 | 0.17 | .867 |
| HF0R × NAM | 0.13 | 0.41 | 0.33 | .744 |

Signif. codes: * .05, . 1

Table F.4: The output of the LMM analysis for pitch range and the *likable* ratings. The intercept is the ORIG stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: ORIG, LF0R, HF0R), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) |
|-------------|-------|------|---------|-----------|
| (Intercept) | 3.20 | 0.35 | 9.04 | <.001 * |
| LF0R | -0.37 | 0.30 | -1.25 | .213 |
| HF0R | -0.31 | 0.31 | -1.01 | .315 |
| male | -0.10 | 0.37 | -0.27 | .793 |
| NAM | -0.05 | 0.38 | -0.13 | .897 |
| LF0R × male | 0.25 | 0.30 | 0.85 | .396 |
| HF0R × male | -0.01 | 0.33 | -0.04 | .967 |
| LF0R × NAM | 0.18 | 0.32 | 0.55 | .586 |
| HF0R × NAM | 0.07 | 0.35 | 0.19 | .848 |

Signif. codes: * .05, . 1

Table F.5: The output of the LMM analysis for pitch range and the *persuasive* ratings. The intercept is the ORIG stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: ORIG, LF0R, HF0R), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) |
|-------------|-------|------|---------|-----------|
| (Intercept) | 3.22 | 0.48 | 6.72 | <.001 * |
| LF0R | -0.30 | 0.32 | -0.93 | .353 |
| HF0R | -0.26 | 0.33 | -0.79 | .430 |
| male | -0.13 | 0.51 | -0.26 | .802 |
| NAM | 0.15 | 0.53 | 0.29 | .779 |
| LF0R × male | 0.24 | 0.33 | 0.73 | .464 |
| HF0R × male | 0.22 | 0.36 | 0.60 | .549 |
| LF0R × NAM | -0.08 | 0.36 | -0.23 | .818 |
| HF0R × NAM | 0.66 | 0.38 | 1.73 | .085 |

Signif. codes: * .05, . 1

Table F.6: The output of the LMM analysis for final contour direction and the *enthusiastic* ratings. The intercept is the falling stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: falling, plateau, rising), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|----------------|-------|------|---------|-----------|---|
| (Intercept) | 3.36 | 0.33 | 10.02 | <.001 | * |
| plateau | 0.00 | 0.27 | 0.01 | .994 | |
| rising | -0.23 | 0.36 | -0.65 | .514 | |
| male | 0.47 | 0.34 | 1.40 | .181 | |
| NAM | -0.61 | 0.33 | -1.86 | .085 | . |
| plateau × male | -0.17 | 0.29 | -0.58 | .562 | |
| rising × male | -0.09 | 0.36 | -0.24 | .810 | |
| plateau × NAM | -0.20 | 0.28 | -0.72 | .470 | |
| rising × NAM | 0.25 | 0.35 | 0.73 | .466 | |

Signif. codes: * .05, . 1

Table F.7: The output of the LMM analysis for final contour direction and the *likable* ratings. The intercept is the falling stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: falling, plateau, rising), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|----------------|-------|------|---------|-----------|---|
| (Intercept) | 3.15 | 0.41 | 7.61 | <.001 | * |
| plateau | 0.08 | 0.27 | 0.29 | .776 | |
| rising | -0.03 | 0.37 | -0.07 | .944 | |
| male | -0.16 | 0.41 | -0.39 | .706 | |
| NAM | -0.06 | 0.42 | -0.15 | .885 | |
| plateau × male | 0.20 | 0.29 | 0.69 | .492 | |
| rising × male | 0.21 | 0.35 | 0.60 | .548 | |
| plateau × NAM | 0.02 | 0.28 | 0.05 | .957 | |
| rising × NAM | 0.04 | 0.38 | 0.10 | .922 | |

Signif. codes: * .05, . 1

Table F.8: The output of the LMM analysis for speech rate and the *likable* ratings. The intercept is the LSR stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: LSR, MSR, HSR), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|-------------|-------|------|---------|-----------|---|
| (Intercept) | 3.17 | 0.39 | 8.04 | <.001 | * |
| MSR | -0.20 | 0.30 | -0.66 | .512 | |
| HSR | -0.03 | 0.30 | -0.11 | .913 | |
| male | -0.20 | 0.43 | -0.46 | .654 | |
| NAM | -0.07 | 0.44 | -0.16 | .877 | |
| MSR × male | 0.46 | 0.34 | 1.36 | .175 | |
| HSR × male | 0.12 | 0.33 | 0.37 | .714 | |
| MSR × NAM | -0.13 | 0.32 | -0.41 | .681 | |
| HSR × NAM | -0.08 | 0.34 | -0.24 | .808 | |

Signif. codes: * .05, . 1

Table F.9: The output of the LMM analysis for speech rate and the *persuasive* ratings. The intercept is the LSR stimulus for female speakers from England. The independent variables are *Manipulation* (three levels: LSR, MSR, HSR), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|-------------|-------|------|---------|-----------|---|
| (Intercept) | 3.20 | 0.48 | 6.69 | <.001 | * |
| MSR | -0.06 | 0.32 | -0.20 | .842 | |
| HSR | -0.14 | 0.33 | -0.41 | .679 | |
| male | -0.14 | 0.52 | -0.27 | .794 | |
| NAM | 0.04 | 0.53 | 0.07 | .948 | |
| MSR × male | 0.13 | 0.36 | 0.35 | .725 | |
| HSR × male | 0.05 | 0.36 | 0.13 | .899 | |
| MSR × NAM | 0.12 | 0.34 | 0.35 | .729 | |
| HSR × NAM | 0.35 | 0.37 | 0.93 | .356 | |

Signif. codes: * .05, . 1

Table F.10: The output of the LMM analysis for speech rate and the *charismatic* ratings. The intercept is the LSR stimulus for female speakers from England who are known to the participants. The independent variables are *Manipulation* (three levels: LSR, MSR, HSR), *Gender* (two levels: female, male), *Origin* (two levels: ENG, NAM), and *Familiarity* (four levels: I know the speaker, They seem familiar, Unsure, I do not know the speaker).

| | Est. | SE | t value | Pr(> t) | |
|----------------------------------|-------|------|---------|-----------|---|
| (Intercept) | 3.27 | 0.40 | 8.08 | <.001 | * |
| MSR | -0.14 | 0.33 | -0.42 | .675 | |
| HSR | -0.50 | 0.33 | -1.51 | .132 | . |
| male | 0.74 | 0.40 | 1.83 | .076 | . |
| NAM | 0.12 | 0.47 | 0.25 | .807 | |
| They seem familiar | 0.51 | 0.41 | 1.26 | .215 | |
| Unsure | -0.16 | 0.40 | -0.40 | .688 | |
| I do not know the speaker | -0.72 | 0.36 | -2.01 | .053 | * |
| MSR × male | 0.28 | 0.19 | 1.46 | .146 | . |
| HSR × male | 0.21 | 0.19 | 1.12 | .262 | |
| MSR × NAM | 0.02 | 0.19 | 0.11 | .913 | |
| HSR × NAM | -0.07 | 0.19 | -0.35 | .726 | |
| MSR × They seem familiar | 0.30 | 0.39 | 0.76 | .445 | |
| HSR × They seem familiar | 0.74 | 0.39 | 1.91 | .056 | . |
| MSR × Unsure | 0.33 | 0.39 | 0.86 | .390 | |
| HSR × Unsure | 0.60 | 0.41 | 1.47 | .143 | . |
| MSR × I do not know the speaker | 0.09 | 0.33 | 0.26 | .796 | |
| HSR × I do not know the speaker | 0.48 | 0.33 | 1.44 | .151 | . |
| male × They seem familiar | -0.59 | 0.35 | -1.68 | .096 | . |
| male × Unsure | -0.24 | 0.37 | -0.66 | .508 | |
| male × I do not know the speaker | -0.35 | 0.32 | -1.11 | .268 | |
| NAM × They seem familiar | -0.45 | 0.42 | -1.05 | .296 | |
| NAM × Unsure | 0.06 | 0.44 | 0.13 | .897 | |
| NAM × I do not know the speaker | 0.09 | 0.41 | 0.21 | .833 | |

Signif. codes: * .05, . 1

Appendix G

Output: Estimated marginal means (Perception 1)

Table G.1: The output of the EMM analysis for the interaction between pitch range manipulation and familiarity rating. The pairwise comparisons between known, familiar, and unknown speakers are presented. The variable *Manipulation* has three levels (ORIG, LFOR, HF0R), and *Familiarity* has four levels (I know the speaker, They seem familiar, Unsure, I do not know the speaker), three of which are displayed.

| contrast | estimate | SE | t.ratio | p.value |
|----------------------------------|----------|------|---------|---------|
| known × ORIG - unknown × ORIG | 0.67 | 0.36 | 1.861 | .757 |
| known × ORIG - unknown × LFOR | 0.78 | 0.36 | 2.16 | .605 |
| familiar × ORIG - unknown × ORIG | 0.87 | 0.22 | 3.95 | .016 * |
| familiar × ORIG - unknown × LFOR | 0.98 | 0.22 | 4.46 | .004 * |
| familiar × ORIG - unknown × HF0R | 0.61 | 0.22 | 2.76 | .237 |
| unknown × ORIG - known × LFOR | 0.08 | 0.37 | 0.21 | 1.000 |
| unknown × ORIG - familiar × LFOR | -0.5 | 0.22 | -2.29 | .501 |
| unknown × ORIG - familiar × HF0R | -0.92 | 0.22 | -4.14 | .008 * |
| known × LFOR - unknown × LFOR | 0.03 | 0.37 | 0.08 | 1.000 |
| known × LFOR - unknown × HF0R | -0.34 | 0.38 | -0.90 | .997 |
| unknown × LFOR - known × HF0R | -0.66 | 0.38 | -1.74 | .816 |
| unknown × LFOR - familiar × HF0R | -1.02 | 0.22 | -4.65 | .002 * |
| unknown × LFOR - unknown × HF0R | -0.37 | 0.11 | -3.37 | .038 * |
| known × HF0R - unknown × HF0R | 0.29 | 0.39 | 0.77 | .999 |
| familiar × HF0R - unknown × HF0R | 0.66 | 0.22 | 2.94 | .161 |

Signif. codes: * .05, .1

Appendix H

Measurements of long stimuli

Table H.1: The measurements of the phrase (= Phr), pause (= P), and breathing (= Br) durations in the stimuli used in the experiment. The phrase durations are only provided for the CUT stimuli as their duration only differed minimally in the other conditions due to manual annotations. In MIX_BR, one of the pauses did not include breathing, indicated by a -. All measurements are provided in seconds.

| Stim. | Int. | Speakers (ENG) | | | | | Speakers (NAM) | | | | |
|----------|------|----------------|------|------|------|------|----------------|------|------|------|------|
| | | LP | ZS | AD | DH | PL | CB | LS | SP | MF | MP |
| CUT | Phr1 | 3.65 | 5.04 | 1.74 | 2.14 | 0.62 | 4.36 | 1.52 | 4.37 | 5.27 | 0.95 |
| | Phr2 | 7.59 | 1.31 | 5.18 | 3.22 | 0.82 | 1.95 | 5.03 | 3.14 | 2.80 | 3.39 |
| | Phr3 | 1.42 | 1.85 | 2.79 | 1.05 | 1.34 | 3.38 | 1.63 | 4.49 | 2.72 | 4.70 |
| | Phr4 | 1.45 | 3.12 | 3.48 | 2.04 | 2.27 | 3.07 | 3.41 | 1.09 | 0.79 | 1.98 |
| LONG_BR | Br1 | 0.28 | 0.33 | 0.23 | 0.16 | 0.17 | 0.26 | 0.39 | 0.27 | 0.12 | 0.13 |
| | Br2 | 0.30 | 0.24 | 0.26 | 0.16 | 0.15 | 0.27 | 0.25 | 0.18 | 0.16 | 0.24 |
| | Br3 | 0.27 | 0.27 | 0.24 | 0.13 | 0.14 | 0.29 | 0.20 | 0.11 | 0.15 | 0.21 |
| | P1 | 0.77 | 0.92 | 0.70 | 0.60 | 0.61 | 0.69 | 0.81 | 0.85 | 0.50 | 0.68 |
| | P2 | 0.59 | 0.64 | 0.91 | 0.73 | 0.57 | 0.74 | 0.61 | 0.78 | 0.78 | 0.70 |
| | P3 | 0.72 | 0.64 | 0.72 | 0.89 | 0.58 | 0.79 | 0.56 | 0.67 | 0.65 | 0.78 |
| LONG_NBR | P1 | 0.86 | 0.82 | 0.75 | 0.80 | 0.80 | 0.82 | 0.94 | 0.90 | 0.61 | 0.61 |
| | P2 | 0.56 | 0.69 | 0.77 | 0.67 | 0.71 | 0.74 | 0.69 | 0.66 | 0.78 | 0.76 |
| | P3 | 0.69 | 0.74 | 0.74 | 0.67 | 0.70 | 1.10 | 0.69 | 0.66 | 0.59 | 0.58 |
| MED_BR | Br1 | 0.27 | 0.33 | 0.23 | 0.14 | 0.17 | 0.23 | 0.35 | 0.28 | 0.12 | 0.11 |
| | Br2 | 0.27 | 0.25 | 0.26 | 0.17 | 0.17 | 0.20 | 0.25 | 0.17 | 0.13 | 0.22 |
| | Br3 | 0.29 | 0.25 | 0.23 | 0.16 | 0.13 | 0.24 | 0.20 | 0.10 | 0.15 | 0.21 |
| | P1 | 0.34 | 0.48 | 0.35 | 0.25 | 0.22 | 0.28 | 0.43 | 0.32 | 0.20 | 0.23 |
| | P2 | 0.30 | 0.35 | 0.40 | 0.28 | 0.30 | 0.26 | 0.32 | 0.22 | 0.47 | 0.39 |
| | P3 | 0.36 | 0.38 | 0.35 | 0.21 | 0.20 | 0.30 | 0.29 | 0.24 | 0.23 | 0.24 |
| MED_NBR | P1 | 0.35 | 0.48 | 0.37 | 0.23 | 0.38 | 0.32 | 0.45 | 0.40 | 0.25 | 0.32 |
| | P2 | 0.29 | 0.38 | 0.40 | 0.29 | 0.40 | 0.27 | 0.32 | 0.36 | 0.53 | 0.41 |
| | P3 | 0.30 | 0.32 | 0.35 | 0.25 | 0.40 | 0.25 | 0.32 | 0.43 | 0.24 | 0.32 |
| MIX_BR | Br1 | 0.27 | 0.33 | 0.24 | 0.14 | - | - | - | - | 0.12 | 0.12 |
| | Br2 | 0.26 | 0.25 | - | 0.17 | 0.17 | 0.22 | 0.25 | 0.17 | - | - |
| | Br3 | - | - | 0.24 | - | 0.14 | 0.25 | 0.20 | 0.09 | 0.13 | 0.22 |
| | P1 | 0.33 | 0.47 | 0.37 | 0.23 | 0.22 | 0.23 | 0.43 | 0.32 | 0.19 | 0.23 |
| | P2 | 0.30 | 0.35 | 0.38 | 0.27 | 0.32 | 0.26 | 0.32 | 0.30 | 0.51 | 0.35 |
| | P3 | 0.32 | 0.32 | 0.34 | 0.35 | 0.21 | 0.32 | 0.28 | 0.25 | 0.24 | 0.25 |
| MIX_L | P1 | 0.17 | 0.18 | 0.31 | 0.17 | 0.80 | 0.80 | 0.88 | 0.90 | 0.18 | 0.15 |
| | P2 | 0.30 | 0.35 | 0.88 | 0.35 | 0.35 | 0.19 | 0.40 | 0.18 | 0.36 | 0.65 |
| | P3 | 0.69 | 0.74 | 0.19 | 0.72 | 0.19 | 0.34 | 0.14 | 0.35 | 0.61 | 0.43 |
| SHORT | P1 | 0.19 | 0.19 | 0.11 | 0.16 | 0.15 | 0.18 | 0.20 | 0.20 | 0.17 | 0.20 |
| | P2 | 0.17 | 0.17 | 0.19 | 0.20 | 0.19 | 0.15 | 0.13 | 0.19 | 0.19 | 0.20 |
| | P3 | 0.16 | 0.19 | 0.14 | 0.19 | 0.19 | 0.16 | 0.14 | 0.17 | 0.18 | 0.19 |

Appendix I

Linear mixed model codes (Perception 2)

Table I.1: LMM codes for the analyses of the pause duration manipulations and their effect on the different rating attributes (= Attr.) *authentic* (AU), *enthusiastic* (EN), *likable* (LI), *persuasive* (PE), and *charismatic* (CH). The ° marks models that showed singular fit, despite testing different optimizers. In some cases, the model did not converge before changing the optimizer, but still returned a singular fit. The × shows that a model did not converge despite testing other optimizers, and is therefore not interpretable.

| Pause duration | | |
|----------------|---|---|
| Attr. | Model structure | |
| AU | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | ° |
| EN | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | × |
| LI | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | |
| PE | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | × |
| CH | <code>lmer(response ~ Manipulation + Gender + Origin + Familiarity + Manipulation:Gender + Manipulation:Origin + Manipulation:Familiarity + Familiarity:Gender + Familiarity:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant) + (1 + Familiarity Participant), data)</code> | |

Table I.2: LMM codes for the analyses of the manipulation of the presence or absence of breathing noises and their effect on the different rating attributes (= Attr.) *authentic* (AU), *enthusiastic* (EN), *likable* (LI), *persuasive* (PE), and *charismatic* (CH). The ° marks models that showed singular fit, despite testing different optimizers. In some cases, the model did not converge before changing the optimizer, but still returned a singular fit. The × shows that a model did not converge despite testing other optimizers, and is therefore not interpretable.

| Breathing | | |
|-----------|---|---|
| Attr. | Model structure | |
| AU | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | ° |
| EN | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data, control = lmerControl(optimizer = "Nelder_Mead"))</code> | |
| LI | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data)</code> | × |
| PE | <code>lmer(response ~ Manipulation + Gender + Origin + Manipulation:Gender + Manipulation:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant), data, control = lmerControl(optimizer = "nlminbwrap"))</code> | |
| CH | <code>lmer(response ~ Manipulation + Gender + Origin + Familiarity + Manipulation:Gender + Manipulation:Origin + Manipulation:Familiarity + Familiarity:Gender + Familiarity:Origin + (1 Speaker) + (1 + Gender Participant) + (1 + Origin Participant) + (1 + Familiarity Participant), data, control = lmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B")))</code> | ° |

Appendix J

Outputs: Linear mixed models (Perception 2)

Table J.1: The output of the LMM analysis for pause duration and the *authentic* ratings. The intercept is the LONG stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG, MED, SHORT, CUT, MIX_L), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|--------------|-------|------|---------|-----------|---|
| (Intercept) | 3.80 | 0.34 | 11.32 | <.001 | * |
| MED | 0.23 | 0.27 | 0.84 | .401 | |
| SHORT | -0.29 | 0.49 | -0.60 | .551 | |
| CUT | -0.14 | 0.49 | -0.30 | .768 | |
| MIX_L | -0.61 | 0.50 | -1.21 | .231 | |
| male | -0.19 | 0.34 | -0.55 | .589 | |
| NAM | -0.46 | 0.40 | -1.16 | .256 | |
| MED × male | -0.22 | 0.29 | -0.76 | .449 | |
| SHORT × male | 0.08 | 0.47 | 0.16 | .875 | |
| CUT × male | -0.10 | 0.47 | -0.22 | .826 | |
| MIX_L × male | 0.10 | 0.48 | 0.22 | .828 | |
| MED × NAM | -0.08 | 0.29 | -0.26 | .793 | |
| SHORT × NAM | 0.55 | 0.60 | 0.92 | .366 | |
| CUT × NAM | 0.35 | 0.60 | 0.59 | .560 | |
| MIX_L × NAM | 0.78 | 0.60 | 1.30 | .207 | |

Signif. codes: * .05, . 1

Table J.2: The output of the LMM analysis for the manipulation of breathing noises and the *enthusiastic* ratings. The intercept is the LONG_NBR stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG_NBR, LONG_BR, MED_NBR, MED_BR, MIX_BR), *Gender* (two levels: female, male), and *Origin* (two levels: ENG, NAM).

| | Est. | SE | t value | Pr(> t) | |
|----------------|-------|------|---------|-----------|---|
| (Intercept) | 3.58 | 0.47 | 7.62 | <.001 | * |
| LONG_BR | 0.36 | 0.53 | 0.68 | .505 | |
| MED_NBR | 0.73 | 0.53 | 1.39 | .181 | |
| MED_BR | 0.46 | 0.33 | 1.40 | .165 | |
| MIX_BR | 0.47 | 0.53 | 0.89 | .387 | |
| male | -0.05 | 0.47 | -0.10 | .922 | |
| NAM | -0.65 | 0.40 | -1.60 | .132 | . |
| LONG_BR × male | -0.19 | 0.49 | -0.39 | .702 | |
| MED_NBR × male | -0.28 | 0.49 | -0.57 | .573 | |
| MED_BR × male | -0.29 | 0.35 | -0.82 | .416 | |
| MIX_BR × male | -0.44 | 0.49 | -0.89 | .382 | |
| LONG_BR × NAM | 0.38 | 0.37 | 1.05 | .302 | |
| MED_NBR × NAM | 0.45 | 0.37 | 1.23 | .226 | |
| MED_BR × NAM | -0.02 | 0.35 | -0.06 | .953 | |
| MIX_BR × NAM | 0.03 | 0.36 | 0.08 | .936 | |

Signif. codes: * .05, . .1

Table J.3: The output of the LMM analysis for the manipulation of pause duration and the *charismatic* ratings. The intercept is the LONG stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG, MED, SHORT, CUT, MIX_L), *Gender* (two levels: female, male), *Origin* (two levels: ENG, NAM), and *Familiarity* (four levels: I know the speaker, They seem familiar, Unsure, I do not know the speaker).

| | Est. | SE | t value | Pr(> t) | |
|-----------------------------------|-------|------|---------|-----------|---|
| (Intercept) | 3.64 | 0.30 | 12.27 | <.001 | * |
| MED | 0.10 | 0.17 | 0.59 | .557 | |
| SHORT | 0.25 | 0.21 | 1.18 | .238 | |
| CUT | 0.26 | 0.21 | 1.27 | .206 | |
| MIX_L | 0.23 | 0.21 | 1.10 | .272 | |
| male | 0.51 | 0.27 | 1.91 | .067 | . |
| NAM | 0.14 | 0.32 | 0.44 | .659 | |
| They seem familiar | 0.08 | 0.23 | 0.36 | .719 | |
| Unsure | 0.02 | 0.25 | 0.08 | .935 | |
| I do not know the speaker | -0.58 | 0.27 | -2.18 | .035 | * |
| MED × male | 0.03 | 0.11 | 0.27 | .785 | |
| SHORT × male | -0.11 | 0.13 | -0.80 | .423 | |
| CUT × male | -0.13 | 0.13 | -0.96 | .337 | |
| MIX_L × male | -0.02 | 0.13 | -0.15 | .882 | |
| MED × NAM | -0.06 | 0.11 | -0.51 | .609 | |
| SHORT × NAM | -0.10 | 0.14 | -0.72 | .473 | |
| CUT × NAM | 0.02 | 0.14 | 0.12 | .901 | |
| MIX_L × NAM | 0.03 | 0.14 | 0.20 | .845 | |
| MED × They seem familiar | -0.12 | 0.21 | -0.57 | .570 | |
| SHORT × They seem familiar | 0.00 | 0.26 | 0.00 | .998 | |
| CUT × They seem familiar | 0.20 | 0.26 | 0.76 | .448 | |
| MIX_L × They seem familiar | -0.15 | 0.26 | -0.58 | .565 | |
| MED × Unsure | -0.13 | 0.20 | -0.65 | .514 | |
| SHORT × Unsure | -0.21 | 0.25 | -0.85 | .396 | |
| CUT × Unsure | -0.29 | 0.24 | -1.20 | .232 | |
| MIX_L × Unsure | -0.29 | 0.25 | -1.14 | .254 | |
| MED × I do not know the speaker | -0.02 | 0.18 | -0.10 | .921 | |
| SHORT × I do not know the speaker | 0.10 | 0.22 | 0.46 | .643 | |
| CUT × I do not know the speaker | -0.10 | 0.21 | -0.47 | .641 | |
| MIX_L × I do not know the speaker | -0.18 | 0.22 | -0.83 | .405 | |
| male × They seem familiar | -0.32 | 0.21 | -1.54 | .123 | . |
| male × Unsure | -0.67 | 0.20 | -3.29 | .001 | * |
| male × I do not know the speaker | -0.87 | 0.18 | -4.77 | <.001 | * |
| NAM × They seem familiar | 0.18 | 0.25 | 0.73 | .465 | |
| NAM × Unsure | -0.25 | 0.25 | -0.99 | .322 | |
| NAM × I do not know the speaker | 0.15 | 0.24 | 0.61 | .542 | |

Signif. codes: * .05, . .1

Table J.4: The output of the LMM analysis for the manipulation of breathing noises and the *charismatic* ratings. The intercept is the LONG_NBR stimulus for female speakers from England. The independent variables are *Manipulation* (five levels: LONG_NBR, LONG_BR, MED_NBR, MED_BR, MIX_BR), *Gender* (two levels: female, male), *Origin* (two levels: ENG, NAM), and *Familiarity* (four levels: I know the speaker, They seem familiar, Unsure, I do not know the speaker).

| | Est. | SE | t value | Pr(> t) | |
|-------------------------------------|-------|------|---------|-----------|---|
| (Intercept) | 3.73 | 0.34 | 10.99 | <.001 | * |
| LONG_BR | 0.15 | 0.25 | 0.61 | .544 | |
| MED_NBR | 0.18 | 0.24 | 0.76 | .450 | |
| MED_BR | 0.16 | 0.25 | 0.65 | .516 | |
| MIX_BR | 0.31 | 0.25 | 1.24 | .217 | |
| male | 0.32 | 0.31 | 1.01 | .320 | |
| NAM | -0.02 | 0.37 | -0.07 | .948 | |
| They seem familiar | -0.08 | 0.31 | -0.27 | .787 | |
| Unsure | -0.25 | 0.33 | -0.76 | .451 | |
| I do not know the speaker | -0.72 | 0.30 | -2.36 | .024 | * |
| LONG_BR × male | -0.18 | 0.16 | -1.10 | .270 | |
| MED_NBR × male | -0.21 | 0.16 | -1.31 | .189 | |
| MED_BR × male | 0.09 | 0.16 | 0.57 | .565 | |
| MIX_BR × male | -0.00 | 0.16 | -0.02 | .981 | |
| LONG_BR × NAM | 0.01 | 0.16 | 0.06 | .949 | |
| MED_NBR × NAM | -0.04 | 0.16 | -0.23 | .821 | |
| MED_BR × NAM | -0.09 | 0.16 | -0.54 | .586 | |
| MIX_BR × NAM | -0.04 | 0.16 | -0.27 | .787 | |
| LONG_BR × They seem familiar | -0.00 | 0.31 | -0.01 | .991 | |
| MED_NBR × They seem familiar | -0.07 | 0.31 | -0.23 | .817 | |
| MED_BR × They seem familiar | -0.16 | 0.31 | -0.51 | .611 | |
| MIX_BR × They seem familiar | 0.07 | 0.31 | 0.24 | .810 | |
| LONG_BR × Unsure | 0.04 | 0.30 | 0.12 | .904 | |
| MED_NBR × Unsure | -0.08 | 0.29 | -0.26 | .796 | |
| MED_BR × Unsure | -0.15 | 0.30 | -0.51 | .612 | |
| MIX_BR × Unsure | -0.24 | 0.30 | -0.81 | .420 | |
| LONG_BR × I do not know the speaker | -0.07 | 0.25 | -0.27 | .789 | |
| MED_NBR × I do not know the speaker | -0.01 | 0.25 | -0.04 | .972 | |
| MED_BR × I do not know the speaker | -0.06 | 0.26 | -0.25 | .802 | |
| MIX_BR × I do not know the speaker | -0.14 | 0.25 | -0.57 | .571 | |
| male × They seem familiar | -0.10 | 0.24 | -0.39 | .695 | |
| male × Unsure | -0.36 | 0.24 | -1.50 | .135 | . |
| male × I do not know the speaker | -0.51 | 0.21 | -2.42 | .016 | * |
| NAM × They seem familiar | 0.38 | 0.29 | 1.29 | .202 | |
| NAM × Unsure | 0.00 | 0.31 | 0.01 | .995 | |
| NAM × I do not know the speaker | 0.28 | 0.28 | 0.98 | .328 | |

Signif. codes: * .05, . .1

Appendix K

Additional figures (Perception 3)

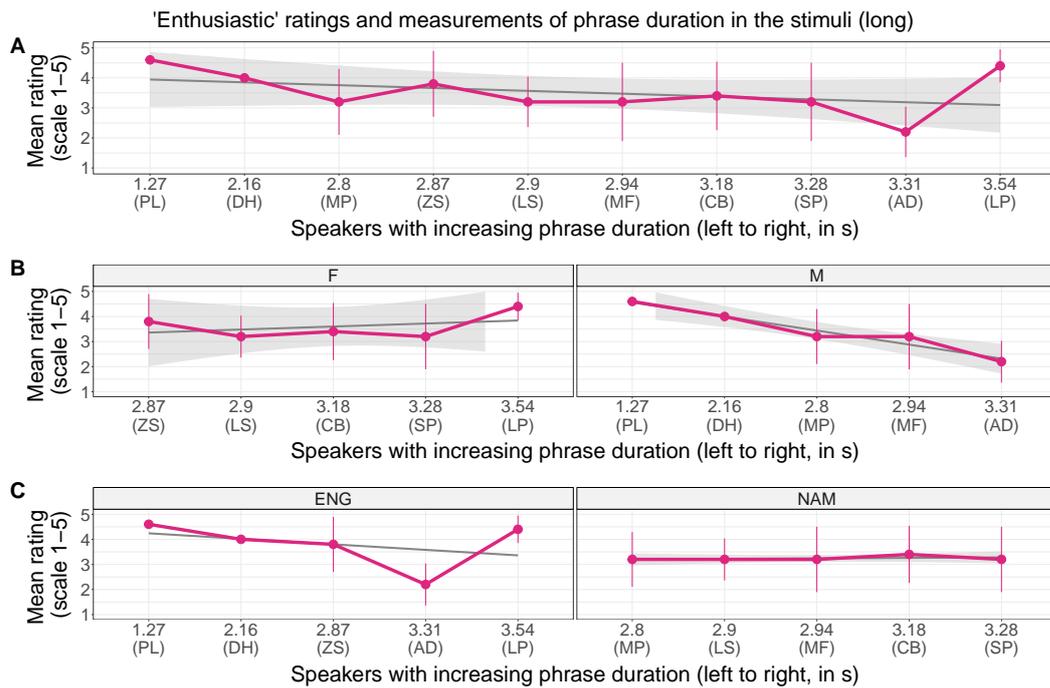


Figure K.1: The results of the correlation of *enthusiastic* ratings and the phrase duration of the long stimuli, for A) all speakers, and split by B) gender and C) origin. The speakers and the respective mean value of the acoustic feature (here: phrase duration) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

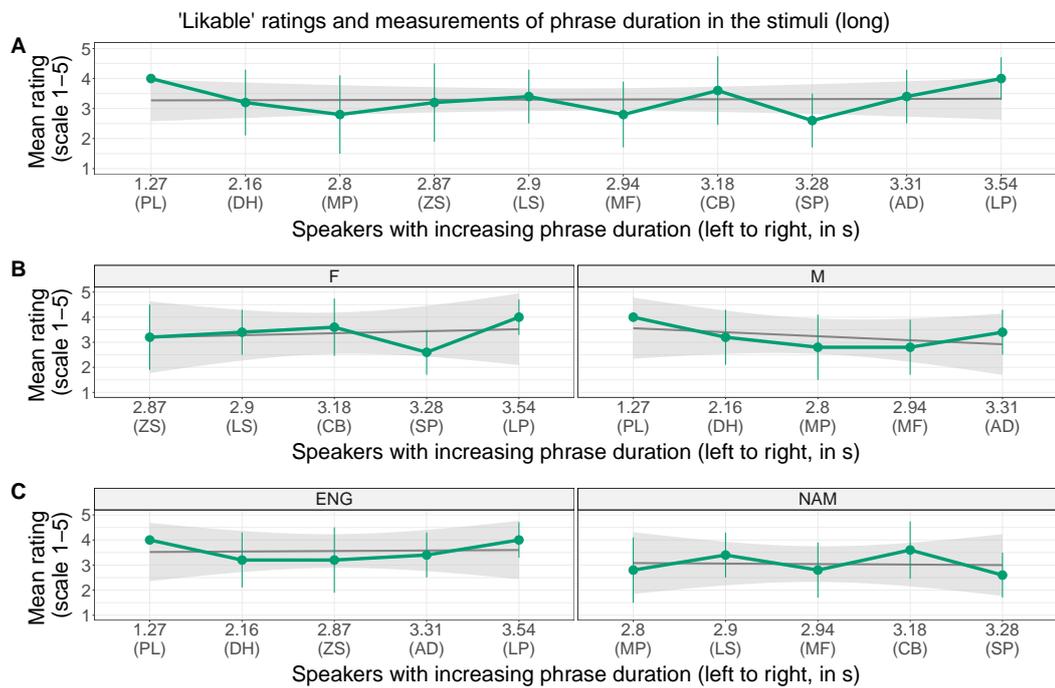


Figure K.2: The results of the correlation of *likable* ratings and the mean phrase duration of the long stimuli, split by A) gender and B) origin. The speakers and the respective mean value of the acoustic feature (here: phrase duration) of the stimuli are arranged in ascending order. The whiskers represent the standard deviation of the rating.

