Design of Scenario-specific Features for Voice Activity Detection and Evaluation for Different Speech Enhancement Applications

Dissertation

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
(Dr.-Ing.)
der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Simon Graf
aus

Aachen

Ulm 2022

Berichterstatter:	Prof. Dr.-Ing. Gerhard Schmidt Prof. Dr.-Ing. Tim Fingscheidt
Datum der Einreichung:	22.02.2022
Datum der mündlichen Prüfung:	21.09.2022

Danksagung / Acknowledgments

Diese Dissertation entstand während meiner Arbeit bei Nuance Communications Deutschland GmbH bzw. später Cerence GmbH in Ulm als externer Doktorand des Instituts für digitale Signalverarbeitung und Systemtheorie der Christian-Albrechts-Universität zu Kiel.

Danken möchte ich zuallererst Prof. Dr.-Ing. Gerhard Schmidt für die hervorragende Betreuung während meiner Promotion auf der einen Seite, aber auch viel grundsätzlicher für das Interesse und die Begeisterung, die er mir in motivierenden Vorlesungen und angeregten Diskussionen für das Thema Digitale (Sprach)Signalverarbeitung vermittelte und mir damit dieses spannende Forschungsfeld erst eröffnete.

Für die Übernahme des Korreferates und die hilfreichen Anmerkungen in seiner detailierten Beurteilung danke ich Prof. Dr.-Ing. Tim Fingscheidt. Auch den weiteren Mitgliedern der Prüfungskommision, Prof. Dr.-Ing. Michael Höft und Prof. Dr.-Ing. Hermann Kohlstedt, gilt mein herzlicher Dank.

Die stete Förderung meiner Promotion auf Firmenseite durch Dr.-Ing. Tim Haulick und Dr.-Ing. Markus Buck machte die Arbeit nicht nur möglich, sondern durch den direkten Praxisbezug auch in spezieller Weise interessant. Allen meinen Kollegen in Ulm und insbesondere Dr.-Ing. Tobias Herbig danke ich für die unermüdliche Unterstützung, prägnante Ratschläge (zu roten Fäden, der Relevanz von Einleitungen etc.) und amüsante Kaffeerunden.

Der Austausch mit meinen Doktorandenkollegen und Freunden Dipl.-Ing. Ingo Schalk-Schupp, Dr.-Ing. Naveen Kumar Desiraju und Dr.-Ing Jonas Jungclaussen weit über die fachlichen Belange unserer jeweiligen Forschung hinaus war mir eine wichtige Stütze und gab mehr als einmal den entscheidenden Impuls für die nächsten Schritte.

Einen besonderen Dank richte ich an meine Eltern Heike und Klaus-Martin Graf, die mir als Logopädin und Professor für Elektrotechnik das Thema dieser Arbeit sprichwörtlich schon in die Wiege legten. Nicht zuletzt danke ich Ramona Beck für ihren Rückhalt bei der Erstellung dieser Arbeit und für unseren Sohn Samuel, der es - nicht als einziger - offenbar nicht erwarten konnte, dass sie fertig würde.

Abstract

Many technical applications nowadays make use of human speech. In situations where controlling a device by hand is not possible or inconvenient, voice can be employed instead. Important use cases can be found in automotive environments where distractions of the driver have to be reduced. Speech-enabled applications allow for dictating messages, controlling devices by voice, or making phone calls out of the driving car via hands-free telephony. Even the communication between passengers inside the car can be facilitated using modern speech applications. In-car-communication (ICC) systems amplify the passenger's speech and allow for convenient conversations at high travel speeds. Also outside the car, mobile speech applications, such as smartphones, become more and more ubiquitous.

The desired speech signal that is recorded by microphones is inevitably superposed by background noise. In automotive scenarios, primarily stationary noise components are observable. In contrast, smartphones can be employed at almost every location resulting in a much higher variability of noise scenarios. Distinguishing the desired speech from background noise is an essential prerequisite for many algorithms that are incorporated in speech applications. When speech is present in the signal, capturing and preserving these desired components is targeted. Contrarily, the suppression of noise usually requires information on the background noise that can be gathered primarily during speech pauses.

Voice activity detection (VAD) aims at identifying presence of speech in a noisy signal. For this purpose, features are extracted: the signal is processed in such a way that certain distinctive properties of human speech are emphasized. Various features focusing on different speech properties have been introduced with the goal of telling apart speech and noise. A detector finally decides *whether* speech is present in the signal. Beyond this, automatic speech recognition systems may identify *what* is said usually incorporating a VAD.

In this thesis, many features for VAD are summarized and classified with respect to properties of human speech that are exploited. New features are introduced considering speech properties that are typically not taken into account. Since different features represent different aspects of human speech, a combination of multiple features is desirable. By considering advantages and drawbacks of each feature, the final detection result can be improved. Adequate feature combinations may increase the robustness against interferences.

In literature, the results of VAD algorithms are typically evaluated without considering a specific application. Different aspects of the detection are evaluated, however, they are not related to the final application's performance. The evaluations in this thesis are therefore dedicated to the requirements of the target application. Some important applications are analyzed with respect to their dependency on VAD results. The importance of accurate VAD results is exemplified for algorithms in an ICC system and for the suppression of babble noise. These applications cover important use cases of VAD with particularly challenging yet contrary conditions. Tried and tested for these rather extreme cases, the approaches discussed in this thesis are well suited also for other applications with less strict constraints.

Kurzzusammenfassung

Viele technische Geräte lassen sich heutzutage per Sprache kontrollieren, was die Bedienung auch in Situationen in denen haptische Eingaben nicht oder nur eingeschränkt möglich sind erleichtert. Anwendungsfälle finden sich insbesondere im Automobilbereich, mit dem Ziel, die Fahrsicherheit zu gewährleisten und Ablenkungen während der Fahrt zu reduzieren. Sprachdialogsysteme ermöglichen beispielsweise das Diktieren von Textnachrichten und die Bedienung vieler Fahrzeugfunktionen während Telefongespräche mittels Freisprecheinrichtungen geführt werden können. In den letzten Jahren gewannen Innenraumkommunikationssysteme zunehmende Popularität. Diese Systeme erleichtern Gespräche zwischen mehreren Personen im Fahrzeug, sogar bei höheren Fahrgeschwindigkeiten. Auch außerhalb der Fahrzeugumgebung sind mobile Sprachanwendungen, wie beispielsweise Smartphones, mittlerweile nahezu allgegenwärtig.

In allen diesen Anwendungen zeichnen Mikrofone neben dem gewünschten Sprachsignal unweigerlich auch Störgeräusche auf. Dabei dominieren im Automobilumfeld stationäre Fahrgeräusche wogegen für Smartphones, die in beinahe jeder denkbaren Umgebung Einsatz finden, vielfältige Störungen zu erwarten sind. Eine Grundvorraussetzung für viele Algorithmen die in Sprachanwendungen eingesetzt werden ist die Unterscheidung von Nutzsprache und Störungen. Wenn Sprachkomponenten im Signal vorliegen, müssen diese erhalten und vor Verzerrungen geschützt werden, dagegen erfordert die Unterdrückung von Störgeräuschen häufig Informationen über das Hintergrundgeräusch. Diese Informationen lassen sich am besten während Sprachpausen extrahieren.

Sprachaktivitätsdetektion dient dazu, Sprachpassagen zu identifizieren und sie von reinem Störgeräusch zu unterscheiden. Das Audiosignal wird dafür so prozessiert, dass charakteristische Sprachmerkmale besonders hervorgehoben werden. Basierend auf diesen Merkmalen entscheidet ein Detektor, ob gerade Sprache im Signal vorliegt oder nicht.

In dieser Arbeit wird eine Vielzahl von Merkmalen für Sprachaktivitätsdetektion zusammengefasst, geordnet nach den zugrundeliegenden Spracheigenschaften. Ergänzend werden neue Merkmale eingeführt, die in dieser Form bisher nicht betrachtet wurden. Generell bietet es sich an, Kombinationen aus mehreren Merkmalen für die Detektion heranzuziehen, da damit unterschiedliche Aspekte der menschlichen Spracherzeugung abgedeckt werden können. Wägt man Vor- und Nachteile der unterschiedlichen Merkmale für die Kombination ab, lässt sich das endgültige Detektionsergebnis und dessen Robustheit gegenüber Störungen optimieren.

Die Evaluierung von Sprachaktivitätsdetektoren betrachtet typischerweise verschiedene Aspekte der Detektion, bezieht sie aber selten auf die eigentliche Zielanwendung. In dieser Arbeit werden daher die speziellen Anforderungen unterschiedlicher Anwendungen explizit mit in die Evaluierung einbezogen. Die Wichtigkeit von Sprachaktivitätsdetektion wird für einige exemplarische Anwendungen – Algorithmen in einem Innenraumkommunikationssystem und einem System zur Verbesserung von Nutzsprache mit Stimmengewirr im Hintergrund – demonstriert. Diese Anwendungen weisen eigene, jeweils sehr spezielle, Anforderungen auf, die Lösungsansätze und Ergebnisse dieser Arbeit lassen sich aber ebenso auf andere Anwendungen mit weniger herausfordernden Randbedingungen übertragen.

Contents

Chapter 1

Introduction

Vocal language is one of the most important media for human communication. From childhood on, speech is utilized primarily in face-to-face communication. Addressing people by voice is a convenient way to express oneself, to share ideas, or to utter demands. For making use of this intuitive way of interaction, speech is integrated into diverse technical devices.

The telephone – invented in the middle of the 19^{th} century – first allowed transmitting speech by means of electrical signals. One century later, integrated circuits and microprocessors offered new opportunities to process speech represented by digital signals. Sophisticated digital signal processing methods allow for preparing speech signals for numerous applications [Hänsler and Schmidt, 2004].

Continuous improvements of speech applications over several decades provided convincing results in many situations. For low noise levels and stationary noise properties, an almost optimal performance can be expected. This experience led to an increasing user acceptance of speech applications. However, the demand for mobility confronts the algorithms with more and more challenging scenarios. Mobile speech applications, such as smartphones, are expected to work in almost any circumstances. Adverse noise conditions therefore have to be taken into account. The increasing requirements of new use cases particularly affect the voice activity detection (VAD) that is an important component of almost any speech application.

Telecommunication nowadays relies almost entirely on digital signal transmission. The speech is encoded into digital data streams that are transmitted, e.g., via internet. Different codecs exist that compress the data in order to reduce the load. Typically, more information has to be transmitted when speech is present in the signal compared to time intervals where speech is absent. To distinguish between both cases, most codecs incorporate speech detectors that identify presence of speech [Vary and Martin, 2006].

In automotive environments, distractions of the driver can be reduced by employing speech applications. For convenience and to satisfy legal restraints, hands-free telephony systems are offered for almost any new car. Microphones that record the speech signal are integrated, e.g., in the roof next to the driver. Due to the noisy environment, it is inevitable that interferences, such as the sound of passing cars, are recorded as well. Speech

enhancement techniques are applied to suppress the noise and to recover the clean speech signal. Speech detectors are employed to find speech activity in the noisy microphone signal. Relying on the detection results, properties of the noise are estimated primarily during speech pauses to prevent speech from being corrupted [Loizou, 2013]. Similar speech enhancement techniques are employed in modern hearing aids. Hearing impaired people particularly benefit from a selective amplification of speech. Again, speech detection is a key factor for separating the desired speech from interfering signal components.

Automatic speech recognition (ASR) systems that convert spoken language into written words allow users to dictate messages or to control devices by voice. End-pointing the relevant segments that contain speech is a typical preprocessing step for ASR. Recognition is performed only on segments where presence of speech was detected.

All these applications make use of speech detection for distinguishing speech and noise. Nevertheless, the intended purposes and thus also the exact requirements regarding the detection results may differ depending on the application. In this thesis, VAD approaches are hence introduced and discussed in the light of variable constraints imposed by applications. For evaluating the detection performance, measures are chosen that particularly reflect the different application requirements such as a quick detection of speech onsets.

1.1 Objectives of this thesis

Several VAD algorithms have been introduced in literature. These algorithms make use of different features that represent the diverse characteristics of speech. In this thesis, a comparative overview of VAD features is presented that associates the different features with corresponding characteristics of human speech. On that basis, VADs for two different applications are introduced and evaluated.

The following new contributions are discussed in this thesis:

- The harmonic structure of voiced speech is typically detected based on a high-resolution spectrum corresponding to a long window in time domain. To deal with shorter windows as required by some applications, a low-complexity feature for speech detection and pitch estimation is introduced.

- The sequential structure of speech is typically not employed for VAD. An algorithm is introduced that explicitly detects temporal alternations of voiced and unvoiced speech. A high robustness against various non-stationary interferences is achieved.

- Usually, the performance of VAD algorithms is expressed in terms of temporally averaged detection results that do not reflect the dynamic behavior of the algorithms. A new performance measure is introduced that explicitly addresses the transient behavior of VAD algorithms.

- Interfering speech from multiple background speakers is an exceptionally challenging issue for speech detection. A VAD algorithm is introduced that is robust against this babble noise.

1.2 Structure of this thesis

Starting with a description of typical VAD approaches, the effects of speech and noise properties on the robustness of detection and the final application's performance are discussed:

In *Chapter 2*, the basic concepts of VAD are outlined. The problem of speech detection based on a noisy signal is introduced along with important notations and naming conventions used throughout the thesis. The importance of VAD is highlighted by means of two exemplary applications: noise suppression and multiple algorithms in the context of an in-car-communication (ICC) system benefit from accurate VAD results. Constraints implied by the application frameworks, different use-case-dependent noise conditions, and algorithm requirements are discussed. At the end of the chapter, a general structure of VAD approaches is described. Most implementations can be subdivided according to this structure into three stages: feature extraction, basic detection, and post-processing.

Chapter 3 is dedicated to characteristic properties of human speech that can be associated with different features for VAD. First, the mechanisms of human speech production are briefly explained. It is further discussed how the manner of articulation is reflected by an audio signal. For feature extraction, the signal is processed in such a way that the resulting feature values indicate presence or absence of speech. Several features assorted from literature are summarized that target on different properties of speech. Additionally, new features are introduced dedicated to speech properties that are usually not taken into consideration. Finally, detectors are discussed that take a binary decision based on the soft feature value. Realizations of a complete VAD are exemplified by means of some standardized approaches.

The performance of VAD algorithms can be evaluated by comparing their detection results with a reference. In *Chapter 4*, signal and reference generation are discussed before different measures are summarized that may quantify the performance. A new measure reflecting the transient behavior of VAD algorithms is introduced. At the end of the chapter, a comparative evaluation of the features is presented that covers a variety of noise conditions.

In a final application, VAD acts as a control mechanism in conjunction with other components. In *Chapter 5*, this interaction is exemplified for the two applications mentioned in the beginning: a noise suppression system dedicated to babble noise as well as diverse algorithms in the context of an ICC system. Particularities of the algorithms regarding VAD performance are identified and appropriate feature combinations are found using the performance measures.

Chapter 2

Voice Activity Detection Fundamentals

Speech detection is an integral part of many different speech applications. By distinguishing time intervals of an audio signal that contain speech from time intervals where speech is absent, algorithms can be controlled and adjusted to the current situation.

In this chapter, the fundamentals of VAD are discussed. The problem and basic notations are introduced that define the scope of this thesis. By means of two exemplary applications, the need for accurate speech detection is demonstrated.

2.1 Description of the problem

Depending on the use-case, speech applications may be deployed in diverse surroundings such as automotive environments or public locations. As a consequence, microphones will not only record the desired speech but they will also capture noise components. VAD targets on identifying presence of desired speech and distinguishing it from interferences like the driving noise in a car or murmuring voices in crowded places. Particularly the latter condition – frequently referred to as babble noise – is challenging since both the desired as well as the interfering components contain speech. Later in this thesis, these two very different scenarios will be considered in more detail.

In general, the VAD problem can be formulated as follows: based on a microphone signal $x(n)$, the presence of speech should be detected. Like all time-domain signals throughout this thesis, the signal is expressed as a function of a discrete sample index n with a sampling frequency f_s. As stated above, two basic situations have to be distinguished: the microphone can either record a composition of speech $s(n)$ and noise components $b(n)$ or capture pure noise during absence of speech. As depicted in Figure 2.1, the microphone signal can hence be modeled by considering two cases

$$x(n) = \begin{cases} s(n) + b(n) & \text{for } VAD_\mathrm{model}(n) = 1 \\ b(n) & \text{for } VAD_\mathrm{model}(n) = 0 \end{cases} \tag{2.1}$$

where the presence or absence of speech is indicated by a function $VAD_{\text{model}}(n)$ that assumes binary values "1" or "0" for each sample n of the signal.

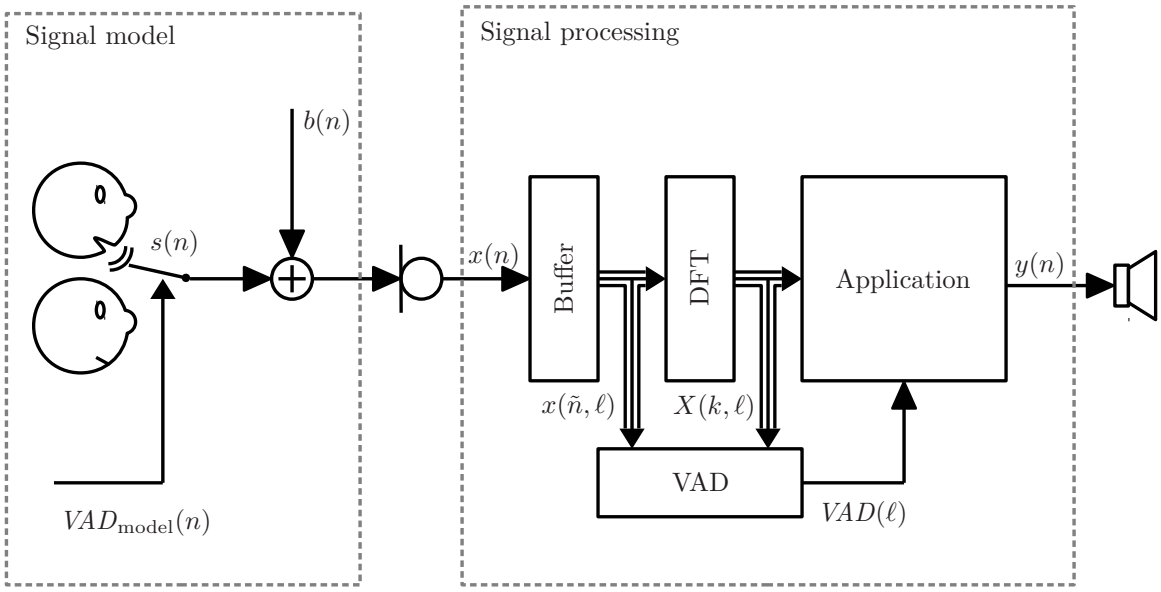


Figure 2.1: Signal model for speech detection and basic signal processing units: one microphone records a signal $x(n)$ in a noisy environment with a background noise component $b(n)$. A speech component $s(n)$ is either absent or present in the microphone signal, depending on the value of $VAD_{\text{model}}(n) \in \{0, 1\}$. The noisy microphone signal is typically processed in the frequency domain $X(k, \ell)$ for different applications. Many algorithms such as noise suppression may benefit from the information whether speech is absent or present in the signal, so a VAD is employed to detect presence of speech $VAD(\ell)$.

The goal of VAD algorithms is to reconstruct the value of $VAD_{\text{model}}(n)$ based on some observations

$$\cdots, x(n-2), x(n-1), x(n), \cdots, x(n + N_{\text{look-ahead}}) \tag{2.2}$$

of the noisy microphone signal. Usually, a number of samples from the past are considered but in principle, the temporal context may be extended even to $N_{\text{look-ahead}} > 0$ future samples after the examined time-instance. Realizing this look-ahead, however, introduces an additional latency in the signal path which is usually not compatible with conversational applications that require the processed signal to be immediately available. In this case, the decision has to rely on previous and the current samples whereas the non-causal part comprising future samples must be omitted.

Even though just a single value at time-instance n should be determined, considering this temporal context is essential: the audio signal is an oscillating waveform thus the instantaneous amplitude has little significance as it may assume any value within a temporal

envelope. Later, the signal's power – related to the temporal envelope – as well as the manner of oscillation will be considered as important cues for VAD. Both can be observed with a relatively short temporal context in the range of 50 milliseconds. Further extending the context may improve the detection accuracy at the expense of a reduced quickness as it will be discussed subsequently in this thesis.

Most speech enhancement algorithms already aggregate short sections of the signal that can be exploited as temporal context for VAD without additional efforts. This block-based signal processing operates on frames

$$x(\tilde{n}, \ell) = x(\ell R + \tilde{n} - N + 1) \qquad \text{with } \tilde{n} \in \{0, 1, \cdots, N-1\} \tag{2.3}$$

buffering blocks of N samples of the signal. The resulting frames are addressed by a frame index ℓ whereas the sample index within the frame is denoted as $\tilde{n}$. The shift between two succeeding frames is denoted as R corresponding to an overlap of $N - R$ samples. The resulting frame rate $f_\mathrm{r} = f_\mathrm{s}/R$ is lower than the original sample rate f_s of the audio signal.

The VAD usually refers to this reduced temporal resolution and generates a decision $VAD(\ell)$ for each frame. A higher resolution of detection is not needed as the target application treats the whole block as either speech or non-speech. Furthermore, the positions of begin and end of speech cannot be determined exactly on a sample level: even for clean speech signals, variations of up to 0.2 s between an expert's reference segmentation and manually as well as automatically generated labels were reported in [Kraljevski et al., 2015] hence detection results on a more granular level than frames would not be meaningful. For these reasons, a block-based decision per frame is considered throughout this thesis.

Frequency-selective operations, such as filtering, can be implemented very efficiently in the frequency domain. Therefore, many speech enhancement algorithms rely on a spectral representation of the audio signal. The buffered blocks of the signal are transferred into the frequency domain[1] by employing a short-time Fourier transform (STFT)

$$X(k, \ell) = \sum_{\tilde{n}=0}^{N-1} w_{\tilde{n}} \cdot x(\tilde{n}, \ell) \cdot e^{-j2\pi k\tilde{n}/N} \qquad \text{with } k \in \{0, 1, \cdots, N-1\}, \tag{2.4}$$

based on a discrete Fourier transform (DFT), where k denotes the frequency index and $w_{\tilde{n}}$ is a window function of length N. For real-valued signals in the time domain, the complex-valued DFT bins $X(k, \ell)$ are symmetric and equal the complex conjugated and flipped $X^*(N - k, \ell)$, so $K = N/2 + 1$ bins are sufficient to describe the full spectrum. Using fast Fourier transform (FFT) implementations, it is further possible to calculate the DFT very efficiently [Cooley and Tukey, 1965].

Such transformation of a time-domain signal into a vector of frequency-domain signals is often referred to as analysis filterbank [Vary and Martin, 2006]. The reverse operation called synthesis filterbank is omitted here as the resynthesized time-domain signal is not

[1] The uppercase letter $X(k, \ell)$ points to a variable in the frequency domain corresponding to $x(n)$ in the time domain.

relevant for VAD. The spectral representation in contrast holds valuable information for speech detection: many characteristics of human speech can be observed more easily in the frequency domain. So several approaches discussed in this thesis rely on this spectral information. In the end, however, the spectral values are always merged back to a single broadband VAD decision per frame that is no longer frequency-selective. Other approaches that exceed the scope of this thesis even perform a frequency-selective localization of speech portions to estimate, e.g, an ideal binary mask (IBM) [Wang and Brown, 2006].

The goal of VAD is to detect the presence of speech regardless of the speaker. Throughout the analyses in this thesis, hence multiple speakers male and female are considered to investigate the detection performance independent from the speaker. Further, VAD is not restricted to a specific language[2]. Whenever a distinct speech component $s(n)$ is present, the VAD should indicate presence of speech. Distinguishing between desired and interfering speech components in contrast is not targeted. As an exception, babble noise should not be confused with distinct speech: this mixture of multiple concurring background speech components should be attributed to the background noise $b(n)$ together with all other interferences.

Finally, the approaches considered in this thesis focus on the particularly challenging situation where only a single microphone is available. The VAD therefore has to rely solely on features that reflect temporal or spectral properties of speech. Beyond that, in devices that are equipped with more than one microphone, spatial information can be exploited for the detection and localization of speech activity [Matheja et al., 2013, Meier and Kellermann, 2016].

2.2 Applications

The analyses in this thesis are dedicated to two applications that both rely on VAD, however, with different purposes. The first application addresses the suppression of babble noise which is a typical use-case for mobile applications. On the other hand, multiple algorithms involved in an ICC system are investigated. Even though there are some particularly challenging constraints for both applications, the results of the analyses can be applied also to other applications.

Initially, in this section, both applications and their peculiarities are briefly summarized to highlight the variability of miscellaneous requirements on VAD. The same applications will be considered in more detail and evaluated at the end of this thesis.

2.2.1 Noise suppression

Noise suppression is used in many speech applications to enhance speech components in a noisy signal. Telephone calls for example are nowadays not limited to silent home or office environments. In a car, hands-free telephony is quite popular. Using mobile phones,

[2]However, since only English test sentences are considered in the analyses, the dependency on the language is not explicitly investigated.

calls can be even made from almost any place. As a consequence, an increased variety of noise conditions has to be taken into account. The conversational partner on the far-end receives not only the desired speech but also some background noise.

This effect can be reduced by employing noise suppression algorithms that reduce the noise and enhance the desired speech. Typically, the noise reduction is implemented in the frequency domain. Spectral weighting coefficients $0 \leq H(k,\ell) \leq 1$ are applied to the noisy spectrum $X(k,\ell) = S(k,\ell) + B(k,\ell)$ to generate an estimate

$$\hat{S}(k,\ell) = H(k,\ell) \cdot X(k,\ell) \tag{2.5}$$

of the original speech spectrum $S(k,\ell)$ and to reduce the background noise component $B(k,\ell)$ as illustrated in Figure 2.2.

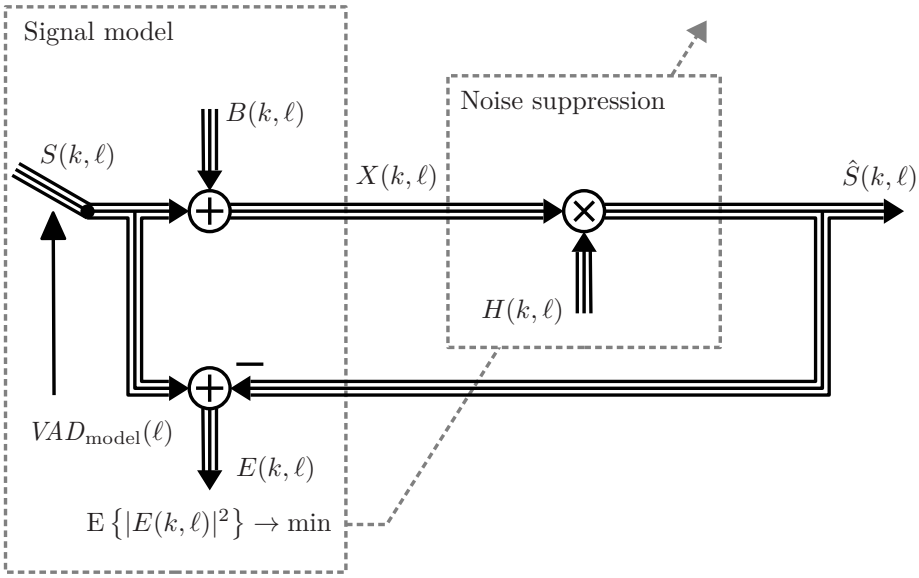


Figure 2.2: Signal model for noise suppression in the frequency domain and optimization criterion of the Wiener filter: spectral weighting coefficients $H(k,\ell)$ are chosen such that the noise component $B(k,\ell)$ is suppressed while the speech component $S(k,\ell)$ is preserved. Since only the noisy mixture $X(k,\ell)$ is accessible for the noise suppression, properties of the other components have to be estimated in practice.

For the following derivation, the signals are modeled by random processes: instead of considering just a specific recorded signal, this probabilistic approach more essentially describes the distribution of all possible signals. A priori knowledge about characteristics of certain signal components such as speech and noise can be integrated in this model which makes it a powerful tool for signal processing [Hänsler, 2001]: for the detection of speech, a recorded audio signal can be compared to the modeled distributions of speech and noise to find the more probable signal class. Similarly, modeling the noise component's distribution helps finding filter weights for noise suppression. By calculating the expected value $\mathrm{E}\{\square\}$ of the probabilistic model, a single representative signal can be determined.

This signal corresponds to the mean over all realizations of the random process. Frequently, the mean squared error (MSE) $\mathrm{E}\{|d-\hat{d}|^2\}$ between a desired signal d and an estimate $\hat{d}$ thereof is employed to deduce a stochastic error criterion that can be utilized for parameter optimization.

An optimal solution for noise suppression is given by the Wiener filter [e.g., Loizou, 2013]

$$H_{\mathrm{Wiener}}(k,\ell) = \underset{H}{\operatorname{argmin}} \left(\mathrm{E}\left\{ \left| S(k,\ell) - \hat{S}(k,\ell) \right|^2 \right\} \right) \tag{2.6}$$

that minimizes the MSE between the estimated and the original speech spectrum.

By inserting the estimate and separating the different components, the MSE

$$\begin{aligned} \mathrm{E}\left\{ \left| S(k,\ell) - \hat{S}(k,\ell) \right|^2 \right\} &= \mathrm{E}\left\{ |S(k,\ell) - H(k,\ell)\cdot X(k,\ell)|^2 \right\} && (2.7)\\ &= \mathrm{E}\left\{ |S(k,\ell)|^2 \right\} + |H(k,\ell)|^2 \cdot \mathrm{E}\left\{ |X(k,\ell)|^2 \right\} && (2.8)\\ &\quad - 2\cdot \mathrm{Re}(H(k,\ell)\cdot \mathrm{E}\{S(k,\ell)\cdot \underbrace{X^*(k,\ell)}_{S^*(k,\ell)+B^*(k,\ell)}\}) \end{aligned}$$

can be determined.

Assuming that speech and noise are orthogonal $\mathrm{E}\{S(k,\ell)\cdot B^*(k,\ell)\} = 0$, the expression can be simplified to

$$\begin{aligned} \mathrm{E}\left\{ \left| S(k,\ell) - \hat{S}(k,\ell) \right|^2 \right\} &= \Phi_{ss}(k,\ell) + |H(k,\ell)|^2 \cdot \Phi_{xx}(k,\ell) \\ &\quad - 2\cdot H(k,\ell)\cdot \Phi_{ss}(k,\ell), \end{aligned} \tag{2.9}$$

where the noisy signal's power spectral density (PSD)

$$\Phi_{xx}(k,\ell) = \mathrm{E}\left\{ |X(k,\ell)|^2 \right\} \tag{2.10}$$

may be expressed by a superposition of the noise PSD

$$\Phi_{bb}(k,\ell) = \mathrm{E}\left\{ |B(k,\ell)|^2 \right\} \tag{2.11}$$

and

$$\Phi_{ss}(k,\ell) = \mathrm{E}\left\{ |S(k,\ell)|^2 \right\} = \Phi_{xx}(k,\ell) - \Phi_{bb}(k,\ell) \tag{2.12}$$

addressing the speech component.

By differentiating Eq. (2.9) with respect to $H(k,\ell)$ and setting the derivative equal to zero, the Wiener filter solution

$$H_{\mathrm{Wiener}}(k,\ell) = \frac{\Phi_{ss}(k,\ell)}{\Phi_{xx}(k,\ell)} = 1 - \frac{\Phi_{bb}(k,\ell)}{\Phi_{xx}(k,\ell)} \tag{2.13}$$

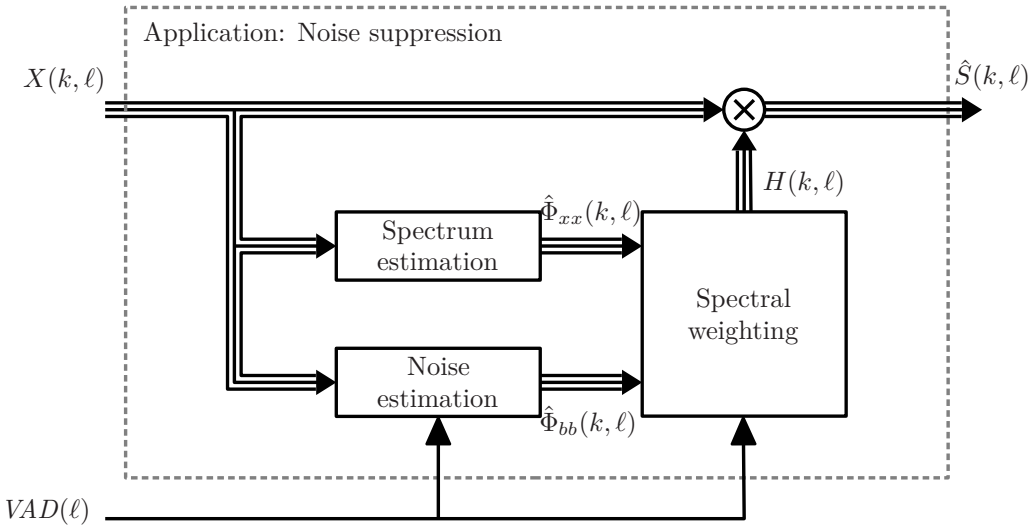


Figure 2.3: Realization of noise suppression using PSD estimates: based on the noisy mixture $X(k,\ell)$ the PSDs of the noise $\Phi_{bb}(k,\ell)$ as well as of the mixture itself $\Phi_{xx}(k,\ell)$ are estimated. Spectral weighting coefficients $H(k,\ell)$ are determined such that the noise component is suppressed. Noise estimation benefits from VAD by focusing on intervals of pure noise during speech pauses. Furthermore, the aggressiveness of the noise suppression can be controlled depending on the detected presence of speech. Speech distortions can be reduced by choosing less aggressive weighting coefficients during signal periods with speech activity.

for noise suppression can be found. Speech components $\Phi_{xx}(k,\ell) \approx \Phi_{ss}(k,\ell)$ that stick out of the noise can pass the filter whereas regions $\Phi_{xx}(k,\ell) \approx \Phi_{bb}(k,\ell)$ that are dominated by noise are attenuated.

In practice, the PSDs of the noisy signal and of the background noise have to be estimated to determine the spectral weights as shown in Figure 2.3. Since the noisy signal is directly accessible, estimation of $\Phi_{xx}(k,\ell)$ can easily be realized by replacing the expected value in Eq. (2.10) by a temporal average

$$\hat{\Phi}_{xx}(k,\ell) = \alpha_{\text{psd,x}} \cdot \hat{\Phi}_{xx}(k,\ell-1) + (1-\alpha_{\text{psd,x}}) \cdot |X(k,\ell)|^2 \tag{2.14}$$

using an infinite impulse response (IIR) filter with a smoothing constant $\alpha_{\text{psd,x}}$. In contrast, the estimation of the PSD of the background noise is much more challenging.

Several methods for estimating the noise PSD $\Phi_{bb}(k,\ell)$ have been introduced in literature. Some rely on a VAD to identify time intervals that do not contain speech [Marzinzik and Kollmeier, 2002]. The estimated noise PSD

$$\hat{\Phi}_{bb}(k,\ell) = \alpha_{\text{psd,b}}(\ell) \cdot \hat{\Phi}_{bb}(k,\ell-1) + (1-\alpha_{\text{psd,b}}(\ell)) \cdot |X(k,\ell)|^2 \tag{2.15}$$

is updated only during these intervals while the previous value is hold in presence of speech. Other algorithms rely on the sparsity of speech and assume that even during presence

of speech there are regions in the spectrum that are unaffected by speech. By tracking minima of the spectrum, these regions can be employed for noise estimation [Martin, 2001]. Combinations of both approaches are used, e.g., by minima controlled recursive averaging (MCRA) [Cohen and Berdugo, 2002] and the improved version IMCRA [Cohen, 2003].

Even though sophisticated methods for noise estimation exist, estimation errors can occur. These errors result in artifacts such as musical tones when applying the original Wiener filter for noise suppression. Modifications of the Wiener filter approach can be employed to reduce the impact of estimation errors on the output signal.

In particular, fast changes of the noise spectrum are typically not accurately tracked. E.g., when multiple persons talk to each other in the background, the resulting babble noise is quite fluctuating. In this case, a VAD can be employed to lower the filter weights during absence of desired speech attaining a more aggressive noise suppression. Corruption of desired speech can be prevented by reducing the aggressiveness of the noise suppression when desired foreground speech is detected. A complete system with a VAD-controlled noise overestimation is described and analyzed in the end of this thesis in Chapter 5.

2.2.2 Signal processing for in-car-communication systems

Speakers in a car are often confronted with high noise levels caused by the driving noise as well as other interfering noise components. This is a severe problem for calls via hands-free telephony out of the driving car but it also complicates conversations between passengers in the cabin. To outweigh the noise, the speakers have to raise their voices inconveniently. Turning their heads to face each other improves the communication, however, doing so is uncomfortable and poses safety issues when the driver loses focus on the traffic.

ICC systems as illustrated in Figure 2.4 may facilitate communication between occupants in the car. Microphones are placed in front of the different speakers in order to record primarily their local speech. These signals are processed with low delay and are played back via loudspeakers close to the listeners. In order to recover the clean speech from the noisy microphone signal, different signal processing techniques are applied: noise reduction attenuates the background noise [Lüke et al., 2011], echo cancellation may remove echo components caused by entertainment playback, e.g., radio, music, or navigation prompts [Franzen et al., 2018], whereas feedback cancellation and suppression target on the ICC's output signal that should not be processed again as ICC input [Schmidt and Haulick, 2006].

Particular challenges arise from the compact arrangement of speakers, microphones, loudspeakers, and listeners within the same environment: the original speech and the reproduced signal superpose in the cabin so that the components are only jointly accessible. This affects both the human perception as well as the system's stability:

- Ideally, the ICC system reinforces the speaker's voice without compromising the naturalness of communication. As one important aspect for achieving an optimal hearing impression, both the original and the reproduced components should reach the listener's ears almost synchronously with only little delay. According to ITU-T

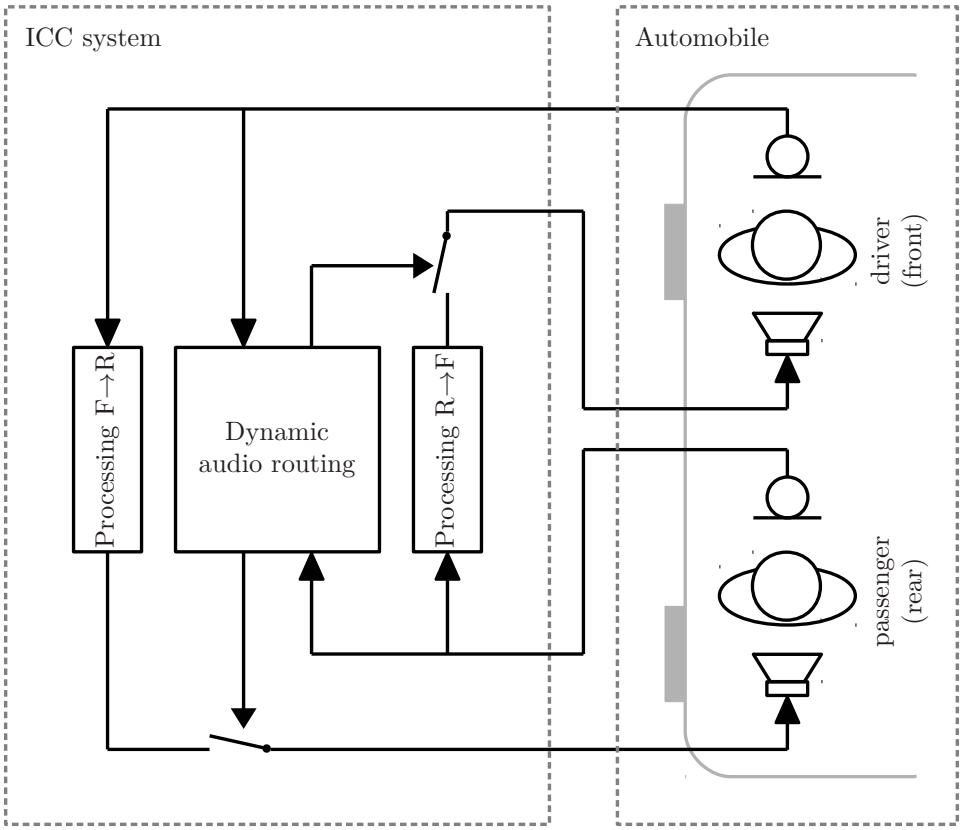


Figure 2.4: ICC systems facilitate conversations between driver and passengers in an automobile: the speech is recorded with microphones, processed by the ICC system and played back via loudspeakers close to the listener. When both directions from front to rear (F→R) and vice versa (R→F) are supported, conflicting situations have to be prevented where signal processing of one direction is affected by the loudspeaker signal of the other one. Dynamic audio routing can be employed to enable only the current speaker's path.

Recommendation P.1150 [ITU, 2020], the delay shall be less than 15 ms and shall be minimized as far as possible in the implementation. Otherwise, human listeners may perceive the mixture as annoying. Furthermore, speakers should not be distracted by the reinforcement of their own voice. The overall system delay including the ICC's processing delay, audio buffers, as well as the acoustic paths in the cabin should hence be as low as possible. This requirement necessitates fast and efficient signal processing algorithms.

- The system's stability on the other hand suffers from a very low delay: the microphones capture feedback components stemming from the loudspeakers in addition to the local speech. Both components are strongly correlated and hence can hardly be separated. Depending on the system's gain, repeated processing in a closed loop may

result in reverberation or even howling artifacts. Introducing an additional delay in the processing path may decorrelate both signal components, however, this is not desirable in an ICC application as the impression of an instantaneous reinforcement has to be preserved. In order to stabilize the system, other techniques such as feedback cancellation or suppression can be applied [Bulling, 2018].

Similar problems are encountered for other applications where microphones and loudspeakers are barely spatially separated such as hearing aids [Strasser and Puder, 2015] or communication systems in full-face firefighter masks [Brodersen et al., 2019]. Most notably the feedback problem is a severe challenge that was investigated in numerous publications [Van Waterschoot and Moonen, 2010]. VAD may help reducing this issue to some extent, e.g., by a VAD-based noise injection that dynamically adds artificial noise components during speech pauses which supports the adaptation of feedback cancellation filters [Mishra et al., 2018].

For several other algorithms in an ICC system, however, VAD has as a more central importance. The following discussions hence will address three different algorithms that all rely on VAD:

- *Noise power estimation* can be limited to intervals that do not contain speech. Using, e.g., a Wiener-filter, the noise estimate can be employed for noise suppression. As discussed in the previous section, both noise estimation and suppression may benefit from VAD.

- Using *automatic gain control (AGC)*, the ICC gain is adjusted dynamically. A constant audio impression over a wide range of noise conditions and speakers can be achieved by compensating level differences relative to a target speech level: louder environments and weaker voices both necessitate higher gains. Estimation of noise and speech levels can be controlled by means of a VAD.

- When multiple processing directions should be supported by the ICC system, e.g., driver to passenger and vise versa, *dynamic audio routing* can be employed to enable only the relevant speech path depending on who is currently speaking. In this way, cross-talk between opposite processing directions can be avoided. The respective speakers can be identified using a VAD-based speaker activity detection.

The strict low-latency requirement of ICC applications impacts the VAD as well as most other algorithms involved [Schmidt and Haulick, 2006]. Short frame lengths are applied for analysis and synthesis of the frequency-domain representation. On the one hand, this results in a high temporal resolution with low processing delay but on the other hand, the spectral resolution is reduced. The VAD therefore has to cope with limited resources. Particularly algorithms in the frequency domain may not rely on a sufficiently high spectral resolution. The spectral fine-structure of speech is usually inaccessible. Furthermore, efficient algorithms are needed as the high frame rate goes along with an increased CPU consumption.

Speech detection approaches that specifically consider these requirements are introduced in Chapter 5. Different criteria on the VAD performance such as a low false-alarm rate, a high speech detection rate, or a swift reaction after speech onsets are exemplified for the three algorithms mentioned above.

2.3 Basic structure of VAD algorithms

VAD can typically be divided into multiple stages [Ramírez et al., 2007] as illustrated in Figure 2.5: first, one ore more features are extracted based on the noisy signal. Some features rely on the time-domain representation $x(\tilde{n}, \ell)$ whereas others employ the representation $X(k, \ell)$ in the frequency domain.

Throughout this thesis, features are denoted as f (or $\boldsymbol{f}$ for multi-dimensional features) followed by an index that addresses the underlying algorithm. These features target on characteristic properties of speech, such as a high signal power or a distinct harmonic structure in case of voiced speech. Based on these features, a distinction between speech and pure noise is desired. It is therefore important to have a low overlap of the distribution of feature values for speech and noise.

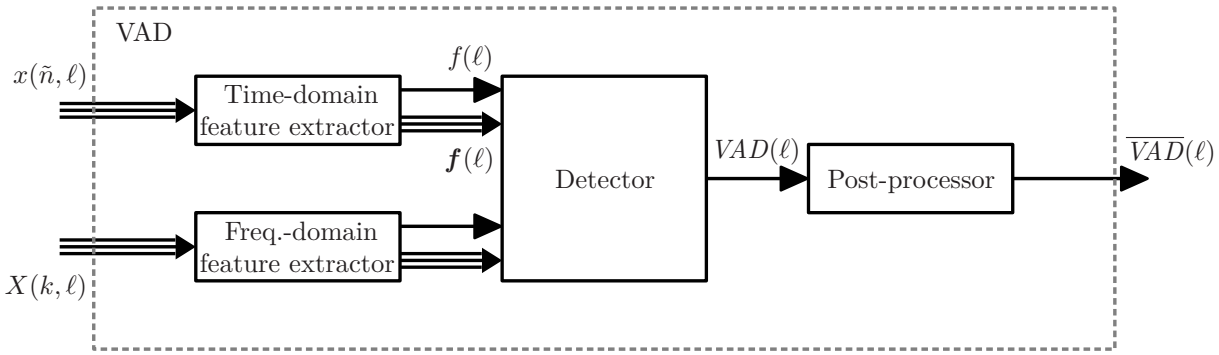


Figure 2.5: Basic structure of VAD algorithms: features $f(\ell)$ (or $\boldsymbol{f}(\ell)$ for non-scalar features) in time or frequency domain are extracted. A speech detector is applied to achieve a preliminary detection result $VAD(\ell)$ that is often post-processed for the final result $\overline{VAD}(\ell)$.

As an example, the signal's short-term power

$$f_{\mathrm{STP}}(\ell) = \sigma_x^2(\ell) = \frac{1}{N} \sum_{\tilde{n}=0}^{N-1} x^2(\tilde{n}, \ell) \tag{2.16}$$

can be employed as a basic feature for speech detection.

As illustrated in Figure 2.6, the feature assumes high values during presence of speech whereas values close to zero are observed for noise. This simple feature is not normalized at all and hence the exact values strongly depend on the recording setup. Even simple

deviations such as a different microphone gain have to be considered later in the detector. Normalization of features as described in Chapter 3 is desirable as it makes VAD more robust against changing conditions.

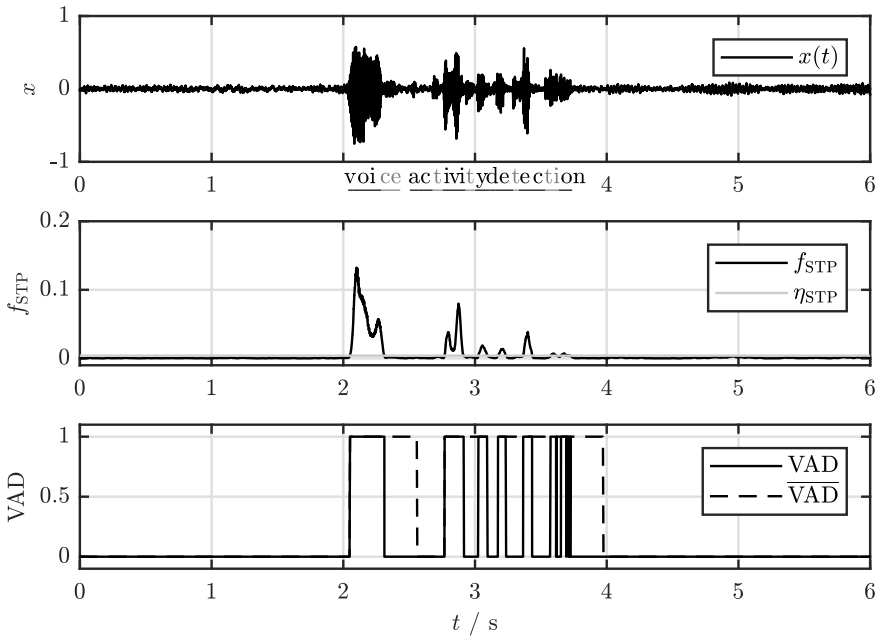


Figure 2.6: Sample recording and basic feature for VAD: in this signal example, the utterance *Voice Activity Detection* is spoken between 2 s and 3.7 s with a moderate level of background noise. This recording will be used later on in Chapter 3 to illustrate the different peculiarities of all the features discussed. In the temporally aligned transcription, voiced speech components are indicated by black letters whereas gray letters correspond to unvoiced components. The different stages of a VAD are illustrated for a basic feature: the signal's short-term power f_{STP} is calculated, a threshold is applied followed by a post-processing for the final VAD.

In a detector, a decision on presence or absence of speech is taken. A basic detector can be realized by comparing the feature values to a threshold

$$VAD(\ell) = \begin{cases} 1 & \text{if } f(\ell) > \eta \\ 0 & \text{else} \end{cases} . \tag{2.17}$$

When the feature exceeds the threshold η, speech is detected, otherwise the detector decides for absence of speech. By adjusting the threshold, the detection results can be controlled. Using a low threshold, much speech is captured, however, it is also more likely that the detector is triggered by noise. On the other hand, a higher threshold increases the robustness against noise but also missed speech has to be expected. Measures as described in Chapter 4 may help finding a reasonable tradeoff between both situations.

Typically, a post-processing is applied in order to correct errors of the detector. The number of missed speech frames during presence of speech can be reduced by introducing a hangover [Vlaj et al., 2016]. When speech was detected in a frame, the decision is hold for L_{ho} subsequent frames

$$\overline{VAD}(\ell) = \begin{cases} 1 & \text{if } \sum_{\tilde{\ell}=0}^{L_{\text{ho}}-1} VAD(\ell - \tilde{\ell}) > 0 \\ 0 & \text{else} \end{cases} \tag{2.18}$$

even though the detector no longer indicates speech.

Chapter 3

Speech Characteristics and Features for Speech Detection

Speech detection starts with the extraction of features that represent characteristic properties of speech. The microphone signal is processed such that conclusions on presence of speech can be made based on the resulting feature values.

As human speech production is a complex process that involves different mechanisms, a single feature usually does not capture the variety of speech adequately [Espi et al., 2010]. Therefore, in the following, properties of human speech are described and multiple features associated with the different characteristics are summarized. Later in this work, the features will be systematically investigated in a comparative analysis similar to the author's article in [Graf et al., 2015a].

3.1 Human speech

A speaking person emits a sound wave to transmit a message via human speech to a listening person or to a speech-driven technical device. To utter the message, the speaker actively shapes the sound wave. Different characteristics of this sound wave are perceived and interpreted by the listener to retrieve the originally spoken message [O'Shaughnessy, 2000].

The different aspects of this articulation are explained in the following. It is further discussed, how the manner of articulation is reflected by a recorded audio signal. Features are summarized that indicate the presence of speech based on the recorded signal.

In this thesis, the characteristics are grouped into two categories: first, segmental properties are discussed that address the articulation of single phones. The respective features focus on instantaneous cues of speech, such as a high power or harmonic signal components. On the other hand, suprasegmental properties are considered that correspond to sequences of phones. To make use of these properties, the signal's evolution is investigated over a longer time range. By doing so, speech can be distinguished more robustly from noise that is either very stationary over time or varies in a different manner than speech.

3.2 Segmental properties: Phones

Phones in linguistics constitute elementary acoustic units of spoken human speech. The speaker converts a message into a spoken utterance by realizing different phonemes[1]. Speech propagates through the air via sound waves and is perceived by the listeners who recognize the original message.

Audio signal processing and particularly VAD in contrast treat the speech signal as a sequence of sound events. The role of those events for the meaning of utterances is usually not taken into account. In this context, most phones can be considered as small sections of the audio signal where the spectral distribution of power stays almost constant. All vowels and all consonants except plosives exhibit a quasi stationarity for about a hundredth of a second that can be exploited for spectral analyses [O'Shaughnessy, 2000].

In this section, the different types of phones are discussed. The underlying mechanism of speech production and the resulting properties of the sound wave are summarized. Features are derived that indicate the presence of speech by reflecting the wave's properties.

3.2.1 Source-filter model

The audio wave that is emitted by the speaker has a variable spectral distribution. Respiration generates an air stream in the lungs that is shaped later to the different phones. By shaping the spectrum, the speaker transforms the message into a spoken utterance.

The source-filter model [Fant, 1970] is frequently employed to describe the different mechanisms of speech production. As illustrated in Figure 3.1, the model divides the speech production process into two stages: excitation and filtering.

The air that streams out of the lungs passes the vocal cords which causes a characteristic excitation structure. Two modes of the vocal cords have to be distinguished: when the vocal cords are open, the air stream passes them without being changed significantly. In this case, a noise-like excitation can be observed that is relevant for unvoiced phones. On the other hand, when the vocal cords are closing, they vibrate and produce periodically occurring impulses. This harmonically rich signal is the second excitation structure that forms the basis of voiced phones.

In the source-filter model, these two types of excitation are represented by two different signal generators. The noise-like excitation of unvoiced speech is modeled by a noise generator whereas the harmonic excitation of voiced speech is modeled by an impulse sequence as illustrated in Figure 3.2. For the latter model, the distance between two consecutive impulses corresponds to the fundamental frequency also often referred to as pitch.

After passing the vocal cords, the sound wave gets further shaped in the vocal tract. Speech organs including tongue, lips, and palate dynamically configure the nasal and the

[1] Linguists distinguish between *phonemes* being meaningful elementary units that are needed to distinguish one word from another in a certain language and their realization through acoustic events, called *phones* [Dresher, 2011]. For the subject matter of this thesis, this distinction, however, appears to be of less importance.

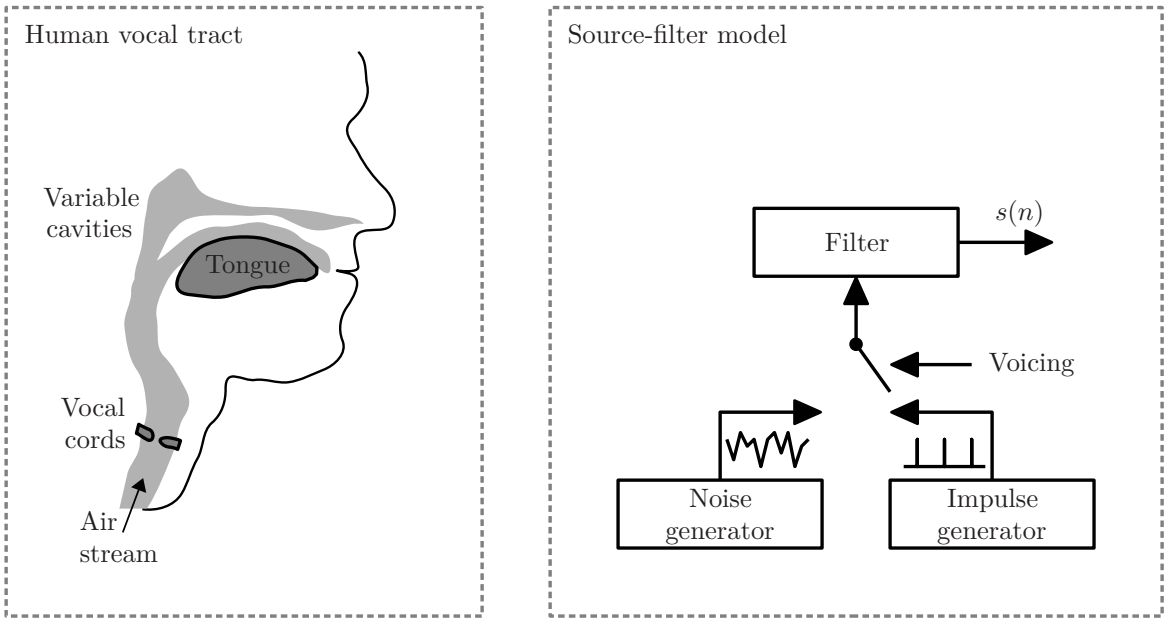


Figure 3.1: Human vocal tract and corresponding source-filter model: the air stream from the lungs passes the vocal cords resulting into a voiced or an unvoiced excitation. In the model, this excitation is represented by two different signal generators for impulse sequences or noise that are selected depending on the degree of voicing. This excitation is spectrally shaped by variable cavities in the vocal tract that are controlled, i.a., by the tongue position. Resonances that can be modeled by a filter emphasize varying frequency regions, which is crucial for distinguishing different phones in the speech signal $s(n)$.

oral cavity. Depending on these settings, variable resonances emphasize some characteristic frequency regions. These formant frequencies are essential for speech recognition. They are associated with certain phones and hence are distinctive properties of speech.

According to the source-filter model, this second stage of speech production can be described by an envelope filter $h_{\mathrm{env},n}(\tilde{n})$ of length N_{env} that reshapes the excitation signal $x_{\mathrm{excite}}(n)$. In Figure 3.3, the envelope filter is illustrated in time and in frequency domain. The formant frequencies are emphasized in the resulting signal

$$s(n) = \sum_{\tilde{n}=0}^{N_{\mathrm{env}}-1} h_{\mathrm{env},n}(\tilde{n}) \cdot x_{\mathrm{excite}}(n-\tilde{n}) \tag{3.1}$$

by choosing appropriate filter parameters[2].

In the following sections, the relevance of the voicing properties as well as of the formant structure of speech for the purpose of speech detection are discussed. Established

[2]The finite impulse response (FIR) filter in Eq. (3.1) with an almost infinite temporal extent N_{env} illustrates the convolutive mixture between excitation and envelope. However, for practical applications usually an implementation based on an IIR filter is preferred that is capable of modeling spectral peaks around the formant frequencies using less parameters as it will be discussed in Section 3.2.5.

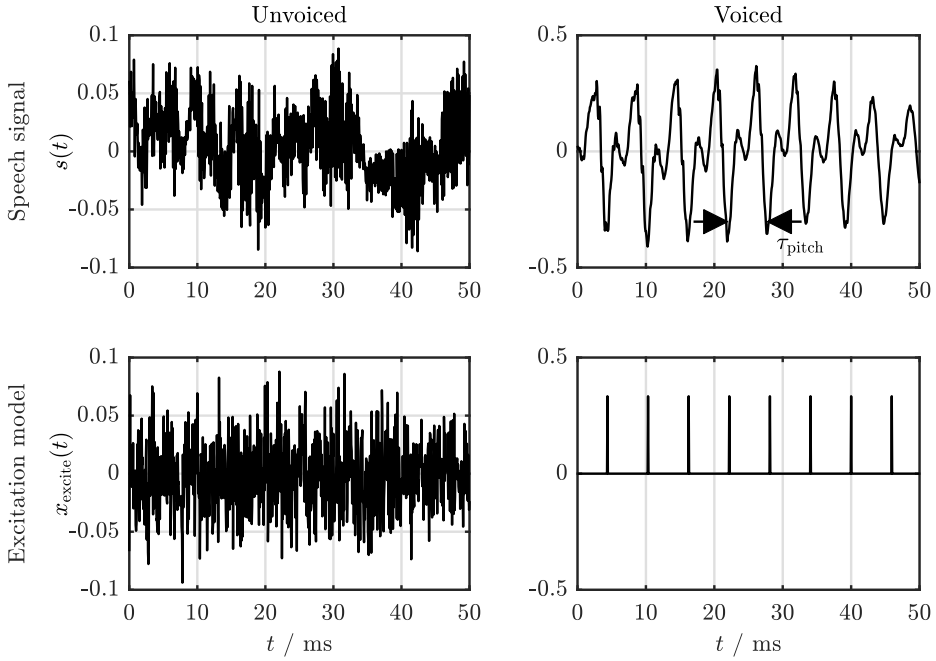


Figure 3.2: Unvoiced vs. voiced excitation: the excitation for unvoiced speech resembles a random noise signal. In contrast for voiced speech, repetitive peaks are observable that can be modeled by a sequence of impulses spaced at intervals corresponding to the pitch period τ_{pitch}.

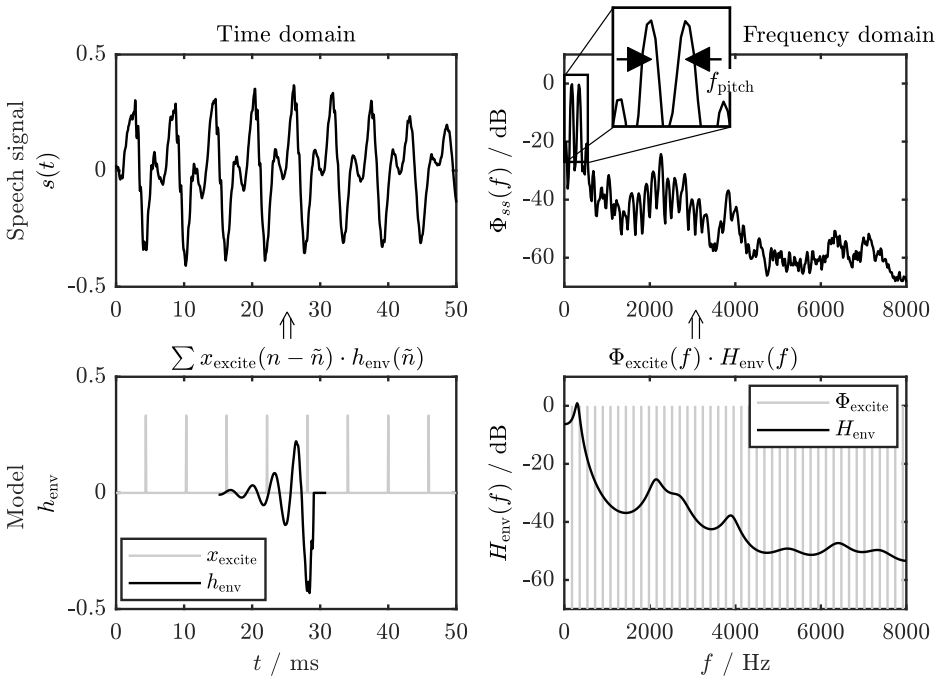


Figure 3.3: Envelope filter in time and frequency domain: the impulse sequence in time domain is convolved with the envelope filter whereas in frequency domain, the harmonic excitation spectrum for pitch frequency f_{pitch} is weighted with the spectral envelope.

features that represent excitation and filtering of speech production are summarized and are complemented by new features.

3.2.2 Power and SNR

The sound wave's power can be seen as a first indicator for the presence of speech. Since power is essential for the transmission of speech, many features for speech detection were derived that make use of this basic property. However, since any interfering noise also contributes power to the sound wave, power alone is not a very discriminative property.

In a silent environment, the audio signal's short-term power $f_{\mathrm{STP}}(\ell) = \sigma_x^2(\ell)$ as defined in Eq. (2.16) can be employed for speech detection. During silence, the feature approaches zero whereas high values can be expected in presence of speech. Implementation of the feature is quite simple, however, there is a major drawback concerning the decision threshold: the feature is not normalized, hence it is not capable of adapting to varying situations. When the background noise level changes or when the signal is rescaled, the decision threshold has to be adjusted accordingly.

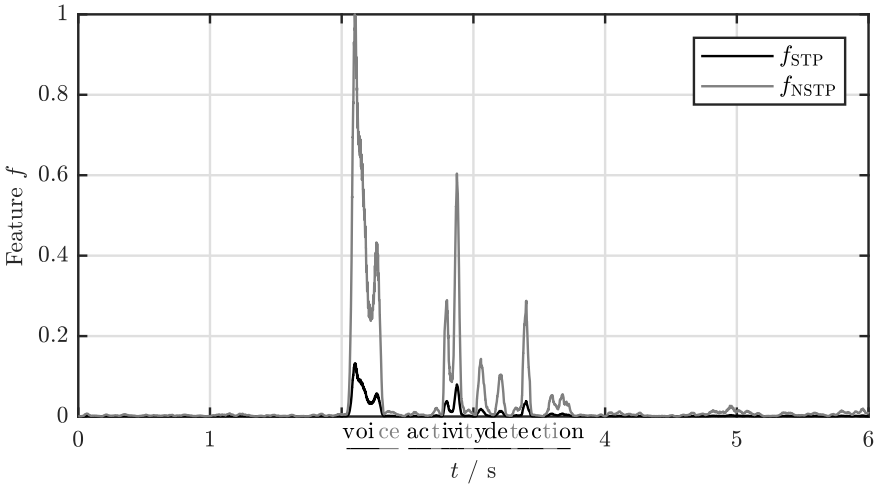


Figure 3.4: Short-term power and normalized feature: without normalization, the range of feature values is arbitrary and strongly depends on the recording conditions. Normalization forces the feature values to the range $0 \leq f_{\mathrm{NSTP}} \leq 1$ such that a fixed threshold can be applied irrespective of the original signal's scaling.

Normalization of the power increases the separability between speech and background noise components. Slow variations of the background noise can be taken into account by tracking changes with time. In contrast, non-stationary interferences are likely to falsely trigger power-based speech detectors.

When both peak power $\sigma_{x\mathrm{max}}^2$ and silence power $\sigma_{x\mathrm{min}}^2$ of the entire signal are known in advance, a static normalization as shown in Figure 3.4 can be applied in order to reduce

the dependency between the feature values

$$f_{\mathrm{NSTP}}(\ell) = \frac{\sigma_x^2(\ell) - \sigma_{x\mathrm{min}}^2}{\sigma_{x\mathrm{max}}^2 - \sigma_{x\mathrm{min}}^2} \tag{3.2}$$

and the recording conditions. In their early algorithm for speech endpointing, Rabiner and Sambur [1975] used this normalization but replaced the power by a mean of magnitude values. This modification simplifies computation in integer arithmetic and additionally reduces the impact of outliers.

In conversational applications, the output signal needs to be calculated continuously without introducing a significant delay compared to the microphone input signal. Instead of the full recording, only short blocks of the signal can be accessed in this case. Hence, speech and noise level have to be tracked dynamically.

Marzinzik and Kollmeier [2002] introduced an approach that determines the power envelope for three different frequency regions. Minimum $\sigma_{x\mathrm{min}}^2(\ell)$ and maximum $\sigma_{x\mathrm{max}}^2(\ell)$ values of the power are tracked dynamically for the full-band signal as well as a low-pass (LP) and a high-pass (HP) filtered version of the signal. The six-dimensional feature vector

$$\boldsymbol{f}_{\mathrm{PED}}(\ell) = [\Gamma(\ell), \Delta(\ell), \Gamma_{\mathrm{LP}}(\ell), \Delta_{\mathrm{LP}}(\ell), \Gamma_{\mathrm{HP}}(\ell), \Delta_{\mathrm{HP}}(\ell)]^T \tag{3.3}$$

is based on the dynamic range

$$\Delta(\ell) = 10 \log_{10} \left(\frac{\sigma_{x\mathrm{max}}^2(\ell)}{\sigma_{x\mathrm{min}}^2(\ell)} \right) \tag{3.4}$$

and the logarithmic power normalized relative to the dynamic range

$$\Gamma(\ell) = \frac{10 \log_{10} \left(\frac{\sigma_x^2(\ell)}{\sigma_{x\mathrm{min}}^2(\ell)} \right)}{\Delta(\ell)} \tag{3.5}$$

for the three frequency regions as visualized in Figure 3.5.

Originally, the detection of speech was based on a set of heuristically specified rules that were applied to the feature vector. In contrast, in this thesis, a neural network will be applied to the feature vector as described in Section 3.4.

The features discussed so far make use of the peak power for normalization. The typical power of speech hence has to be known in advance or has to be estimated during runtime.

In presence of noise, the Lombard reflex [Junqua, 1996] lets most speakers raise their voices to predominate the signal. Searching for signal components that stick out of the background noise is therefore often sufficient. When only the noise power is tracked, the signal-to-noise ratio

$$\delta_{\mathrm{SNR1}}(\ell) = 10 \log_{10} \left(\frac{\sigma_x^2(\ell)}{\hat{\sigma}_b^2(\ell - 1)} \right) = 10 \log_{10} \sigma_x^2(\ell) - 10 \log_{10} \hat{\sigma}_b^2(\ell - 1) \tag{3.6}$$

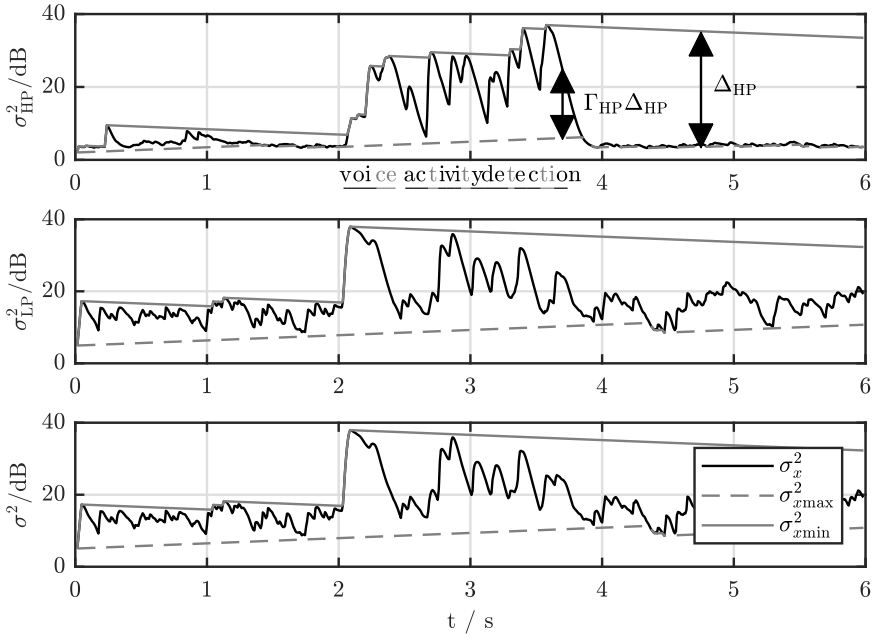


Figure 3.5: Power envelope dynamics feature: for three different frequency regions, maximum and minimum power are tracked. The six-dimensional feature $\boldsymbol{f}_{\mathrm{PED}}$ relies on the dynamic ranges $\Delta(\ell)$ as well as normalized power estimates $\Gamma(\ell)$ relative to the dynamic ranges.

can be employed for normalization. During absence of speech, the SNR[3] fluctuates around 0 dB whereas higher values are assumed for speech. Estimating the noise power is a crucial aspect here. Given that a VAD result is already available, e.g., from the previous frame, the noise estimate

$$\hat{\sigma}_b^2(\ell) = \alpha_{\mathrm{SNR1,b}} \cdot \hat{\sigma}_b^2(\ell-1) + (1-\alpha_{\mathrm{SNR1,b}}) \cdot \sigma_x^2(\ell) \tag{3.7}$$

can be kept constant ($\alpha_{\mathrm{SNR1,b}} = 1$) during presence of speech and can be updated ($\alpha_{\mathrm{SNR1,b}} < 1$) otherwise. In order to improve the robustness against fluctuating noise, Van Gerven and Xie [1997] proposed a feature

$$f_{\mathrm{SNR1}}(\ell) = \frac{\delta_{\mathrm{SNR1}}(\ell)}{\sqrt{\overline{\delta^2_{\mathrm{SNR1}}}(\ell)}} \tag{3.8}$$

based on an additional normalization of the signal-to-noise ratio (SNR) with respect to the SNR's variance

$$\overline{\delta^2_{\mathrm{SNR1}}}(\ell) = \alpha_{\mathrm{SNR1},\delta} \cdot \overline{\delta^2_{\mathrm{SNR1}}}(\ell-1) + (1-\alpha_{\mathrm{SNR1},\delta}) \cdot \delta^2_{\mathrm{SNR1}}(\ell) \tag{3.9}$$

[3]The definition employed here is sometimes referred to as *a posteriori* SNR [Tan and Lindberg, 2009] or input-to-noise ration (INR) since the numerator addresses the input signal after the mixing of speech and noise components.

during presence of noise. Again, the smoothing parameter $\alpha_{\mathrm{SNR1},\delta}$ is controlled by a preliminary VAD: a threshold is applied to the feature $f_{\mathrm{SNR1}}(\ell) > \eta_{\mathrm{SNR1,noise}}$ as illustrated in Figure 3.6 to limit the noise estimation to intervals of absence of speech.

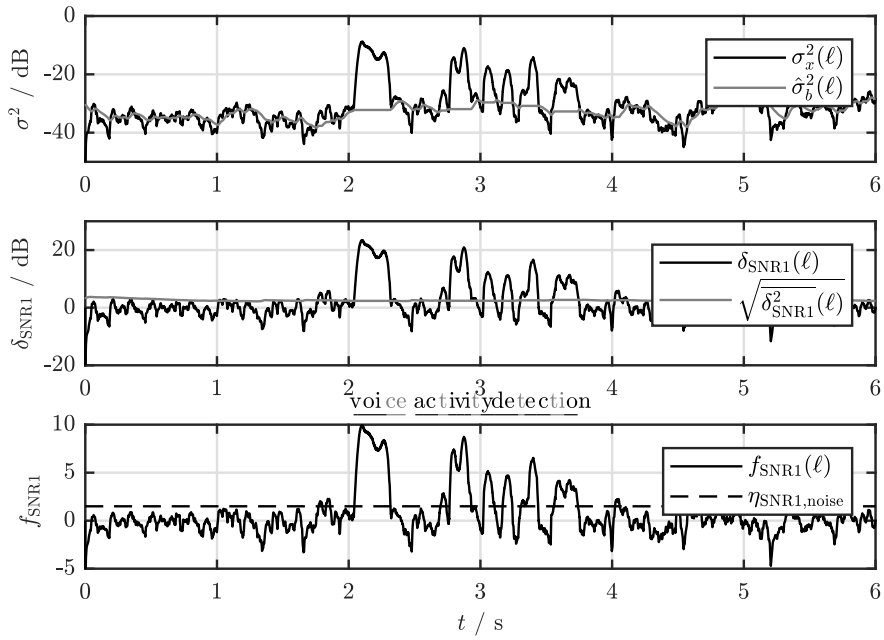


Figure 3.6: SNR feature: the background noise power $\hat{\sigma}_b^2(\ell)$ is tracked to determine the SNR $\delta_{\mathrm{SNR1}}(\ell)$ as well as the SNR's standard deviation $\sqrt{\overline{\delta^2_{\mathrm{SNR1}}}(\ell)}$ for normalization. When the feature $f_{\mathrm{SNR1}}(\ell)$ defined as the ratio of current SNR and the standard deviation exceeds a threshold $\eta_{\mathrm{SNR1,noise}}$, the noise estimator is kept constant. For the actual speech detection, a second (different) threshold can be applied to the feature.

Pencak and Nelson [1995] introduced an SNR-based feature that estimates both signal and noise power based on a single frame. Assuming that speech has a sparse frequency distribution where only few spectral bins are excited, the lowest k_{low} bins can be attributed to the noise power

$$\sigma^2_{b,\mathrm{spec}}(\ell) = \frac{1}{k_{\mathrm{low}}} \sum_{k=0}^{k_{\mathrm{low}}-1} \tilde{\Phi}_{xx,\mathrm{sorted}}(k,\ell), \tag{3.10}$$

whereas the highest $k_{\mathrm{high}}(\ell)$ bins that contribute 40% of the total power are accumulated

$$\sigma^2_{x,\mathrm{spec}}(\ell) = \frac{1}{k_{\mathrm{high}}(\ell)} \sum_{k=K-k_{\mathrm{high}}(\ell)}^{K-1} \tilde{\Phi}_{xx,\mathrm{sorted}}(k,\ell) \tag{3.11}$$

for the signal power. Both powers rely on a spectrum with ascendingly sorted values: $\tilde{\Phi}_{xx,\mathrm{sorted}}(0,\ell) \leq \tilde{\Phi}_{xx,\mathrm{sorted}}(1,\ell) \leq \cdots \leq \tilde{\Phi}_{xx,\mathrm{sorted}}(K-1,\ell)$. Again, the ratio of signal and

noise power

$$f_{\mathrm{SNR2}}(\ell) = \frac{\sigma^2_{x,\mathrm{spec}}(\ell)}{\hat{\sigma}^2_{b,\mathrm{spec}}(\ell)} \tag{3.12}$$

is employed as a feature. When using this approach, a flat spectrum for pure noise is essential. Otherwise, for an unequal distribution of noise power, spectral bins exhibiting comparatively high values might accidentally be attributed to speech power. To overcome this problem, the spectrum's long-term average is equalized before the sorting is applied: the instantaneous spectrum $\hat{\Phi}_{xx}(k,\ell)$ is normalized with respect to a long-term average $\overline{\Phi}_{xx}(k,\ell)$ such that the resulting spectrum $\tilde{\Phi}_{xx}(k,\ell)$ is flattened during absence of speech.

The short-term power only considers short intervals for the detection of speech while disregarding the long-term temporal context. This context, however, contains valuable information on presence of speech. Ramírez et al. [2004a] hence extended the time range that is taken into account for an SNR-based speech detection. Instead of the short-term power, their approach relies on a long-term envelope

$$\Phi_{\mathrm{LTSE}}(k,\ell) = \max_{-L_{\mathrm{LTSE}} \leq \tilde{\ell} \leq L_{\mathrm{LTSE}}} \left(\hat{\Phi}_{xx}(k, \ell + \tilde{\ell}) \right) \tag{3.13}$$

for each spectral bin. The approach incorporates a look-ahead by L_{LTSE} frames and hence introduces a corresponding algorithmic delay during online processing. The final long-term spectral divergence (LTSD) feature

$$f_{\mathrm{LTSD}}(\ell) = 10 \log_{10} \left(\frac{1}{K} \sum_{k=0}^{K-1} \frac{\Phi_{\mathrm{LTSE}}(k,\ell)}{\hat{\Phi}_{bb}(k,\ell)} \right) \tag{3.14}$$

relies on an average of SNR values over frequency. Again, the noise estimation is controlled based on the VAD result: $\hat{\Phi}_{bb}(k,\ell)$ according to Eq. (2.15) is updated only during absence of speech. In [Ramírez et al., 2004b], the same authors introduced an extended mechanism based on a multi-band quantile SNR estimation: Instead of the maximum operator in Eq. (3.13), this approach relies on the 90% percentile for estimating the envelope whereas the median is applied for the noise estimate.

High power and SNR both are necessary conditions that enable accurate detection of speech. However, as discussed, noise components cannot sufficiently be rejected relying solely on these basic features. Considering more characteristic properties of speech, e.g., the voicing is therefore advisable.

3.2.3 Voicing

The degree of voicing is an essential factor for speech detection: a majority of phones is voiced with a harmonically rich excitation structure. All vowels but also many consonants belong to this group. On the other hand, there are some unvoiced phones, such as many fricatives. For this group, the excitation is rather noisy and does not show a harmonic

structure [Hu and Wang, 2008]. When both situations should be covered for the detection of speech, a combination of multiple features seems promising.

In this section, features are discussed that represent the excitation structures of speech. Due to the completely different characteristics of voiced and unvoiced excitation, the features are usually specialized to one of both. While most of the features in this section target at the distinct harmonic structure of voiced speech, the first two features discussed are dedicated to unvoiced speech.

Unvoiced speech

The noise-like excitation of unvoiced speech can be modeled by a white noise random process: samples at two discrete time instances are uncorrelated for this process irrespective of the spacing between both. A realization of this random process is hence characterized by rapidly time-varying values that go along with frequent changes of the signal amplitude's sign. This property is even emphasized when the signal is shaped with a high-pass envelope as it is usually the case for unvoiced phones.

To quantify these changes, a zero-crossing rate (ZCR) [Rabiner and Sambur, 1975]

$$f_{\mathrm{ZCR}}(\ell) = \frac{1}{N-1} \sum_{n=\ell R-N+2}^{\ell R} 0.5 \cdot |\operatorname{sign}(x(n)) - \operatorname{sign}(x(n-1))| \tag{3.15}$$

can be calculated that relates the number of changes of the sign

$$\operatorname{sign}(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x = 0 \\ -1 & \text{for } x < 0 \end{cases} \tag{3.16}$$

within a certain time interval to the respective interval's duration. Inherently, the feature is limited to values in the range $0 \leq f_{\mathrm{ZCR}}(\ell) \leq 1$. As illustrated in Figure 3.7, unvoiced speech is indicated by high feature values whereas the feature assumes lower values for voiced speech and noise.

While the zero-crossing rate is a basic feature in the time domain, the spectral entropy measure

$$f_{\mathrm{SE}}(\ell) = -\frac{1}{\log K} \sum_{k=0}^{K-1} \tilde{\Phi}_{\mathrm{SE}}(k,\ell) \cdot \log\left(\tilde{\Phi}_{\mathrm{SE}}(k,\ell) + \epsilon\right) \tag{3.17}$$

can be used to assess the whiteness of a signal in the frequency domain [Kristjansson et al., 2005]. As basis for this entropy-like feature, the spectrum is handled as a probability distribution: it is normalized such that accumulation of the normalized spectrum

$$\tilde{\Phi}_{\mathrm{SE}}(k,\ell) = \frac{\hat{\Phi}_{xx}(k,\ell)}{\sum_{k=0}^{K-1} \hat{\Phi}_{xx}(k,\ell)} \tag{3.18}$$

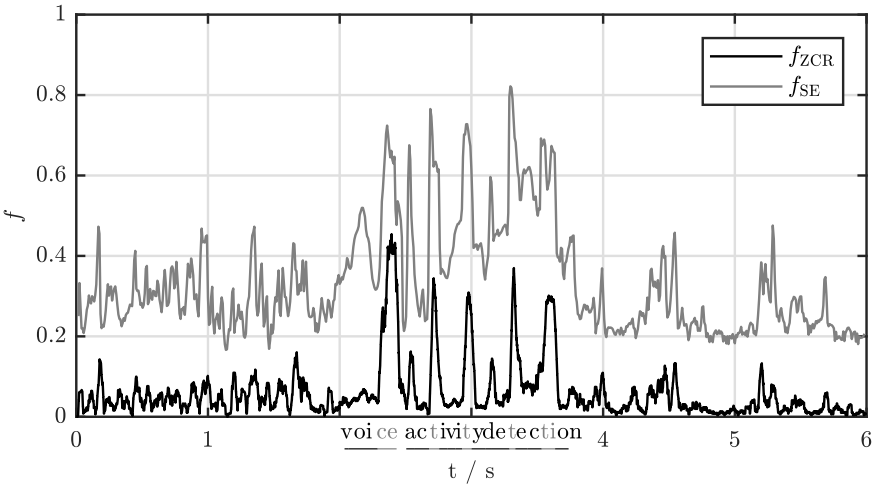


Figure 3.7: Zero-crossing rate and spectral entropy: both features indicate presence of unvoiced speech by high values. The corresponding phones can hence be detected based on a time-domain (f_{ZCR}) or on a frequency-domain (f_{SE}) feature.

over frequency leads to one.

For a white signal characterized by an identical excitation of all frequencies, the feature assumes its maximum value one. Otherwise, the value is lower with $0 \leq f_{\mathrm{SE}}(\ell) \leq 1$. The feature is exemplified along with the ZCR in Figure 3.7.

Another feature closely related to spectral entropy is given by the spectral flatness measure [Madhu, 2009]. This feature is based on the ratio between geometric mean and arithmetic mean of the spectrum for quantifying the whiteness of a signal. The geometric mean is always lower than the arithmetic mean or equal for a flat spectrum, hence the feature assumes values between zero and one with one indicating a perfectly white spectrum.

This almost equal excitation in the upper part of the spectrum above 2 kHz is characteristic for unvoiced speech. In contrast, in the second part of this section the attention will shift towards the lower frequencies where a harmonic excitation structure is observable for voiced speech.

Voiced speech

Voiced speech usually exhibits a distinct harmonic excitation that is quite characteristic for human speech. This property can be exploited by features that are extracted in time domain but also in frequency domain. To capture the harmonic structure, periodic signal components in a certain frequency range have to be detected.

According to the source-filter model, the excitation can be expressed in time domain by a sequence of impulses that is convolved later with the vocal tract filter. The signal hence has a periodic structure with similarly shaped and recurring peaks. The distance

between two adjacent peaks corresponds to the pitch period.

The goal of voiced speech detection is to detect whether such a periodic structure is present in the signal. Estimating the actual pitch period is not necessarily requested, however, it is provided by most algorithms as a side-effect.

A repetitive signal modeled by a time-domain random process can be detected using the normalized auto-correlation function (ACF)

$$ACF(\tau, \ell) = \frac{\mathrm{E}\left\{x(\ell R) \cdot x(\ell R - \tau)\right\}}{\mathrm{E}\left\{x^2(\ell R)\right\}} \tag{3.19}$$

that reaches values in the range $-1 \leq ACF(\tau, \ell) \leq 1$ depending on the degree of periodicity. The measure is maximized to one for perfectly periodic signals when the correlation lag τ matches the signal's periodic time τ_p.

The pitch frequency of human speech usually is in the range between 50 Hz and 250 Hz [Nelson and Pencak, 1995]. Pitch periods in the corresponding range have to be considered for the detection of voiced speech. The ACF's maximum

$$f_\mathrm{ACF}(\ell) = \max_{\tau \in \mathbb{T}_\mathrm{ACF}} \left(ACF(\tau, \ell)\right) \tag{3.20}$$

in the relevant periodic time range $\mathbb{T}_\mathrm{ACF} \hat{=} [4\,\mathrm{ms}, 20\,\mathrm{ms}]$ can be employed as a feature that represents the degree of voicing. High values can be expected for voiced speech, whereas for unvoiced speech and noise the feature values are low as illustrated in Figure 3.8.

The auto-correlation in Eq. (3.19) can be estimated by an inverse DFT

$$\widehat{ACF}(\tau, \ell) = \frac{\sum_{k=0}^{N-1} \hat{\Phi}_{xx}(k, \ell) \cdot e^{2\pi j k \tau / N}}{\sum_{k=0}^{N-1} \hat{\Phi}_{xx}(k, \ell)} \tag{3.21}$$

of the signal's PSD estimate. Compared to a direct calculation in time domain, this approach is usually more efficient and allows for interpolation between samples. Nevertheless, other measures were introduced, e.g., in [Tucker, 1992] or in [Orlandi et al., 2003] that directly operate on the time-domain signal. Ghaemmaghami et al. [2010] proposed evaluating the zero-crossings of the ACF in addition to the maximum search in Eq. (3.20) to improve the robustness against noise. To suppress non-repetitive components, Kurth and Cornaggia-Urrigshardt [2014] introduced a shift-ACF that takes into account more than one repetition by concatenating shift-product and shift-minimum operations. Their approach is a generalization of the standard ACF.

The ACF depends on both the harmonic excitation as well as on properties of the vocal tract. To separate both effects, the cepstrum

$$CEP(\tau, \ell) = \sum_{k=0}^{N-1} \log\left(\hat{\Phi}_{xx}(k, \ell)\right) \cdot \cos\left(\frac{\tau(k + 0.5)\pi}{N}\right) \tag{3.22}$$

can be employed instead [Kristjansson et al., 2005]. For transformation into the cepstral domain, usually a discrete cosine transform (DCT) is chosen. In contrast to the ACF,

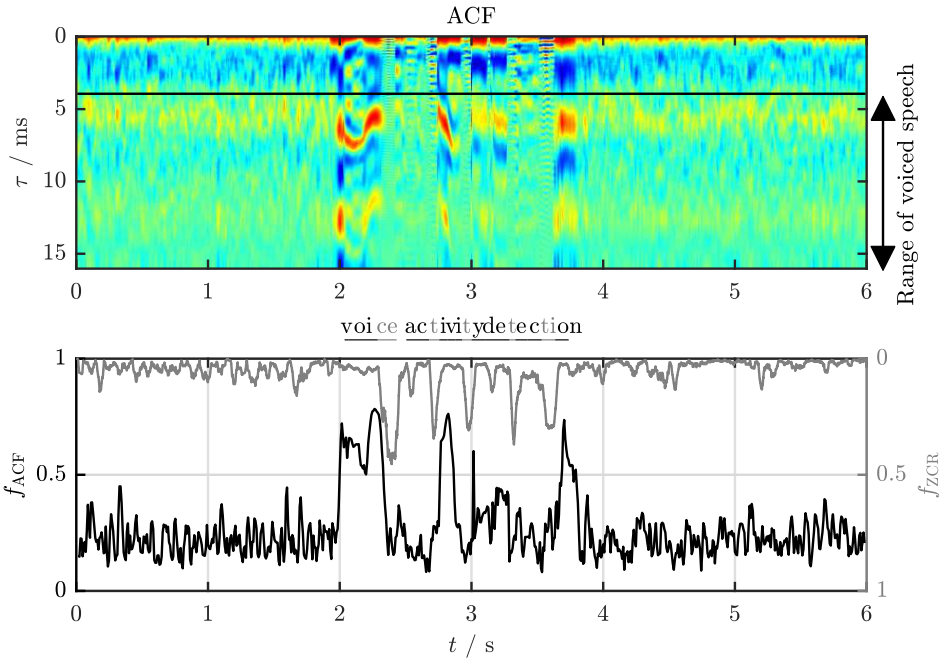


Figure 3.8: Auto-correlation function: the feature f_{ACF} indicates voiced speech by high values. For comparison, f_{ZCR} is depicted again (vertically mirrored for better visibility). The peaks of both features appear interleaved since voiced and unvoiced phones occur alternately in the utterance. This finding will be exploited later in this thesis by a dedicated feature.

a logarithm is applied to the power spectrum before the transformation. This logarithm converts the convolutive mixture in time domain of the excitation signal and the vocal tract filter in Eq. (3.1) into a sum $CEP = CEP_{\mathrm{env}} + CEP_{\mathrm{excite}}$ of two cepstral components that relate to separated cepstral regions. Harmonic components are represented by a cepstral peak

$$f_{\mathrm{CEP1}}(\ell) = \max_{\tau \in \mathbb{T}_{\mathrm{CEP}}} \left(CEP(\tau, \ell)\right) - \min_{\tau}(CEP(\tau, \ell)) \tag{3.23}$$

for the higher order cepstral bins corresponding to the cepstral range $\mathbb{T}_{\mathrm{CEP}} \hat{=} [4\,\mathrm{ms}, 20\,\mathrm{ms}]$ as shown in Figure 3.9. On the other hand, the lower order cepstral bins characterize the vocal tract as it will be discussed in Section 3.2.5. A static harmonic feature based on the cepstrum was proposed in [Fukuda et al., 2010].

The features discussed so far transform the spectral representation back into the time domain or into the cepstral domain before searching for the harmonic structure. In contrast, using the harmonic product spectrum (HPS)

$$HPS(k, \ell) = \sum_{h=1}^{H} \log\left(\hat{\Phi}_{xx}(h \cdot k, \ell)\right) = \log\left(\prod_{h=1}^{H} \hat{\Phi}_{xx}(h \cdot k, \ell)\right), \tag{3.24}$$

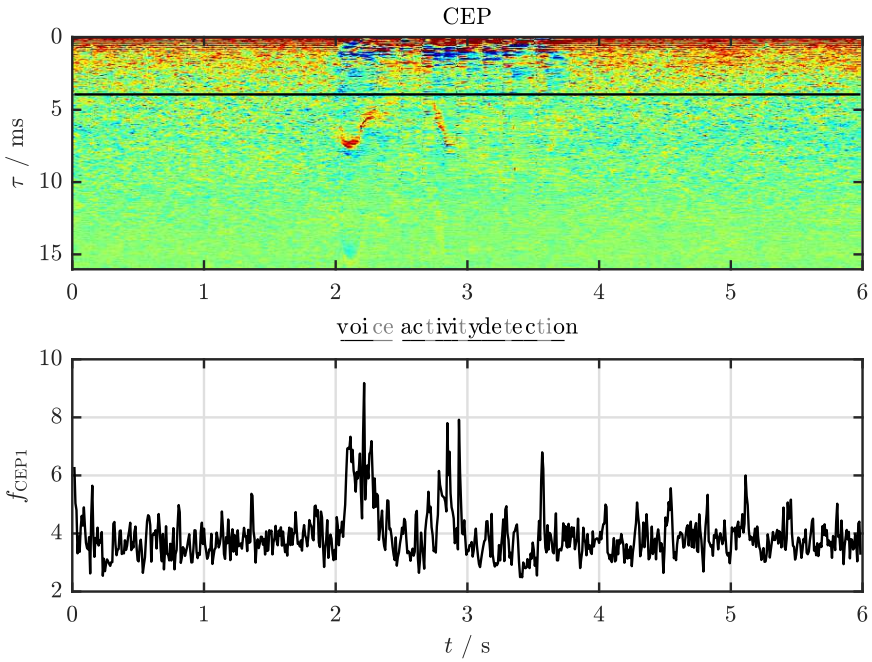


Figure 3.9: The cepstrum reflects the presence of voiced speech by a peak similar to ACF.

the presence of harmonic components can directly be detected in frequency domain [Sadjadi and Hansen, 2013]. For this, a spectral bin k that corresponds to a potential pitch frequency and $H-1$ harmonics are aggregated where the product over the respective frequency bins emphasizes a harmonic structure. The measure is maximized when all harmonics are excited in a signal, so the maximum

$$f_{\mathrm{HPS}}(\ell) = \max_k(HPS(k,\ell)) - HPS(1,\ell) \tag{3.25}$$

can be employed as a feature for voiced speech detection as shown in Figure 3.10. Normalization with respect to aperiodic components increases the robustness as described in [Nelson and Pencak, 1995] and similarly in [Ishizuka and Nakatani, 2006].

Even though a detection of harmonic components is possible based on the spectrum, estimating the exact pitch frequency is more difficult compared to the ACF-based methods. As the spectral resolution $\frac{N}{f_\mathrm{s}}$ (the number of bins per kHz) is usually not sufficiently high, the pitch, i.e., the distance between harmonics cannot be derived accurately. This effect even gets worse when the application necessitates short frame lengths as discussed in the next section.

3.2.4 Detection of voiced speech for short frame lengths

Multiple periods of the repetitive excitation have to be captured in order to resolve the pitch. Detection of voiced speech therefore typically requires a long frame length that

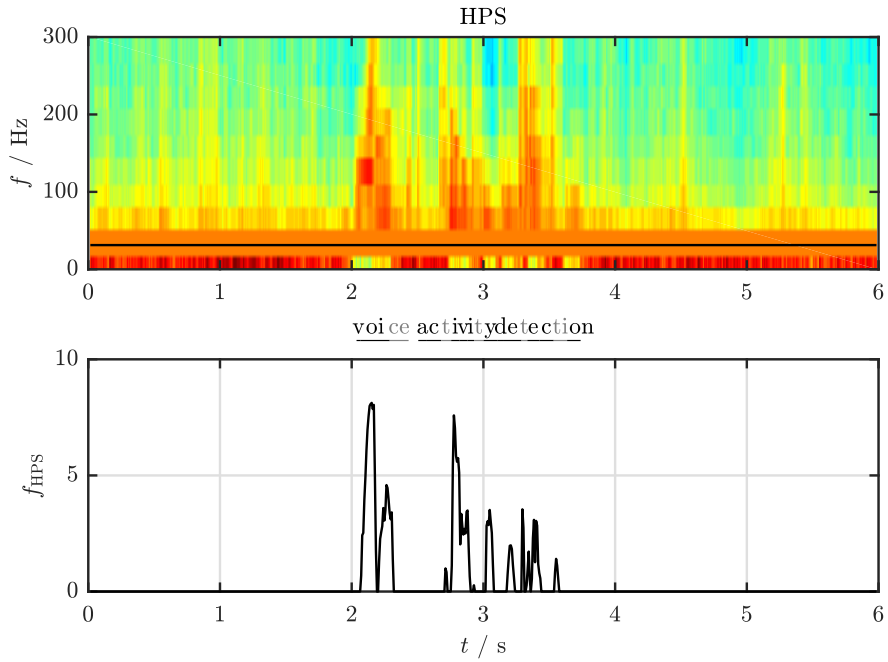


Figure 3.10: Harmonic product spectrum: presence of voiced speech can be detected also in the frequency domain. However, estimating the exact pitch frequency is difficult due to the low spectral resolution.

significantly exceeds the pitch period of speech. To capture a very low pitch of 50 Hz, the frame must be chosen longer than 20 ms. Some applications, such as ICC systems, however, operate on much shorter frames in order to keep the latency and the computational complexity low. In this case, a single frame of length N' is not sufficient for the detection. To overcome this limitation, the temporal context can be extended by jointly considering multiple frames as illustrated in Figure 3.11.

Spectral refinement

One method for pitch estimation that extends the effective frame length in frequency domain was introduced by Krini and Schmidt [2007]. Using this technique, an efficient computation of a high-resolution spectrum is possible provided that multiple low-resolution spectra are already available. Based on this refined spectrum, detection of voiced speech can be implemented, e.g, using the ACF according to Eq. (3.20) and (3.21) with (2.14).

In the following, the derivation of spectral refinement is briefly summarized to outline the basic idea. For details, please consult the original publications [Krini and Schmidt, 2007, 2012].

Spectral refinement corresponds to an extension of the temporal context in time domain:

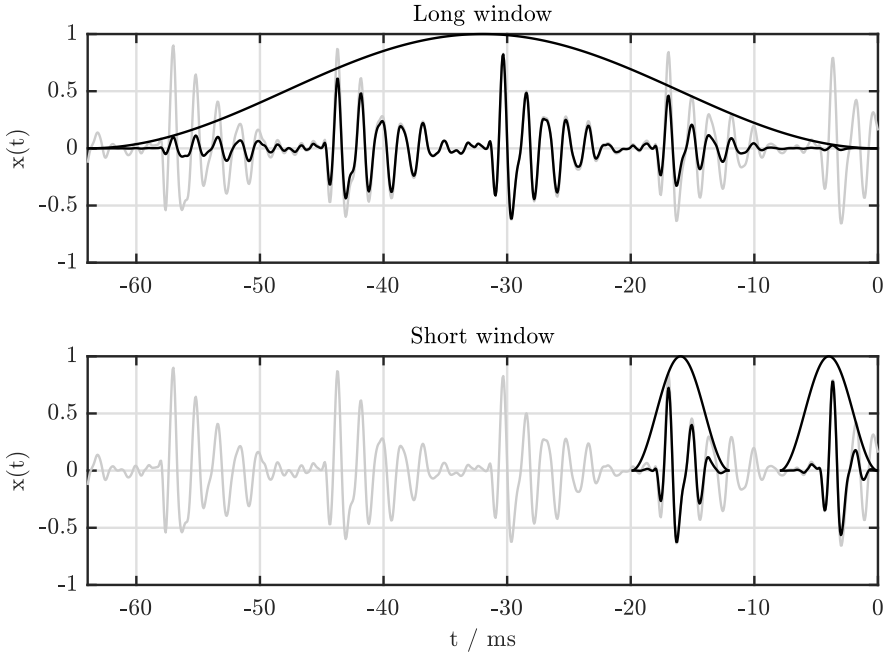


Figure 3.11: Long vs. short window lengths: the long frame captures multiple repetitions of the voiced speech's excitation which allows for pitch estimation and detection. The short frames in contrast only capture a single peak each. To gather the periodic structure using short frames, multiple frames have to be considered jointly.

as depicted in Figure 3.12, M short frames[4] $\boldsymbol{x}'(\ell) = [w'_0 \cdot x(\ell R - N' + 1), \cdots, w'_{N'-1} \cdot x(\ell R)]^T$ of length N' each are stacked successively in a supervector. This vector is processed with a combination matrix $\boldsymbol{s} \in \mathbb{R}^{N \times MN'}$ to achieve a longer frame

$$\hat{\boldsymbol{x}}(\ell) = \boldsymbol{s} \cdot \begin{bmatrix} \boldsymbol{x}'(\ell) \\ \boldsymbol{x}'(\ell - 1) \\ \vdots \\ \boldsymbol{x}'(\ell - (M-1)) \end{bmatrix} \tag{3.26}$$

with an increased effective frame length of $N = N' + (M-1) \cdot R$. The goal is to find a matrix $\boldsymbol{s}$ such that a long frame $\boldsymbol{x}(\ell) = [w_0 \cdot x(\ell R - N + 1), \cdots, w_{N-1} \cdot x(\ell R)]^T$ is approximated by the combination of multiple overlapping short frames.

Both the long frame and the short frames are excerpts from the same audio signal $x(n)$, however, they address different time intervals and are weighted with different window functions. To approximate the long window w_n by a weighted sum $\hat{w}_n$ of overlapping short windows, the combination matrix therefore has to temporally align the short windows w'_n and scale them accordingly as illustrated in Figure 3.13.

[4]Please note that in the following short frames and low-resolution spectra as well as the short correlation's lags are marked with $\square'$. For the frame index ℓ no such distinction is made as for both long and short frames a higher frame rate is considered now.

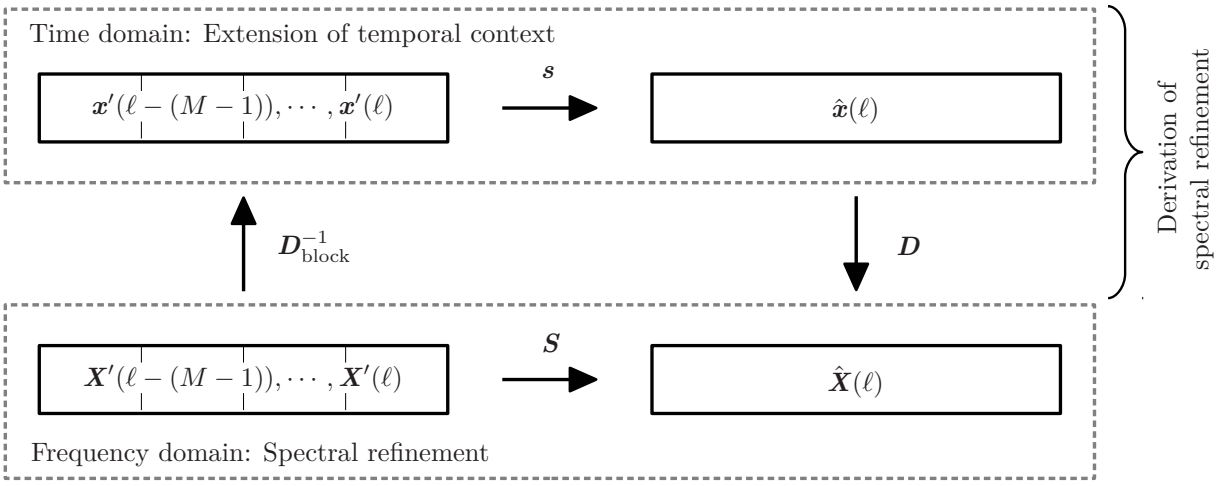


Figure 3.12: Spectral refinement and corresponding extension of temporal context in the time domain: multiple short frames $\boldsymbol{x}'$ (or low-resolution spectra $\boldsymbol{X}'$) are stacked and processed with a matrix $\boldsymbol{s}$ (or $\boldsymbol{S}$) in time (or frequency) domain to achieve a longer frame $\hat{\boldsymbol{x}}$ (or high-resolution spectrum $\hat{\boldsymbol{X}}$).

The temporal alignment can be realized by arranging M short window functions as rows in a matrix

$$\boldsymbol{w}' = \begin{bmatrix} w'_0 & w'_1 & \cdots & w'_{N'-1} & \mathbf{0}^{1\times(M-1)R} & \\ \mathbf{0}^{1\times R} & w'_0 & w'_1 & \cdots & w'_{N'-1} & \mathbf{0}^{1\times(M-2)R} \\ \vdots & & \ddots & & & \ddots \\ \mathbf{0}^{1\times(M-1)R} & & & w'_0 & w'_1 & \cdots & w'_{N'-1} \end{bmatrix} \tag{3.27}$$

where the frame shift between successive frames is considered by applying zero-padding before and after the sliding windows. The corresponding scaling factors $s_0,\ldots,s_{M-1}$ can now be determined by relating this matrix to the desired long window function

$$\begin{bmatrix} w_0, w_1, \cdots, w_{N-1} \end{bmatrix} = \begin{bmatrix} s_0, s_1, \cdots, s_{M-1} \end{bmatrix} \cdot \boldsymbol{w}' \tag{3.28}$$

that shall be reproduced by a weighted sum over the matrix's rows. The least square solution for this equation system can be found by multiplying both sides of Eq. (3.28) with the matrix's pseudo inverse [Krini and Schmidt, 2012]

$$\boldsymbol{w}'^{+} = \boldsymbol{w}'^{T}(\boldsymbol{w}'\boldsymbol{w}'^{T})^{-1}. \tag{3.29}$$

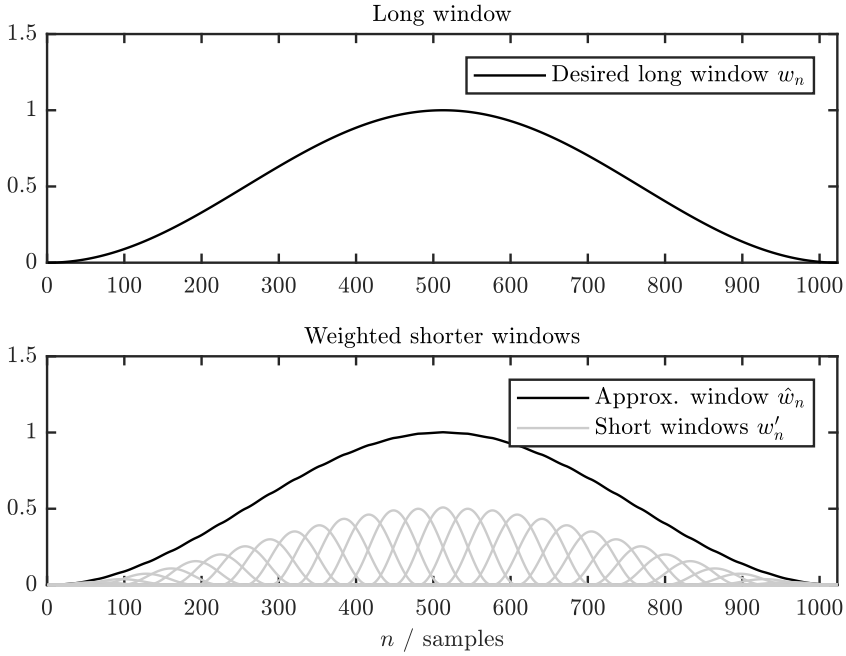


Figure 3.13: Approximation of a long window: multiple short windows are temporally aligned, rescaled, and accumulated such that the weighted sum approximates the desired long window.

With this solution, the combination matrix in time domain

$$\boldsymbol{s} = \begin{bmatrix} s_0 & 0 & 0 & \boldsymbol{0}^{R\times N'} & & \cdots & \boldsymbol{0}^{(M-1)R\times N'} \\ 0 & \ddots & 0 & s_1 & 0 & 0 & \\ 0 & 0 & s_0 & 0 & \ddots & 0 & \\ & & & 0 & 0 & s_1 & \ddots \; s_{M-1} \; 0 \quad 0 \\ & & & & & & 0 \quad \ddots \quad 0 \\ \boldsymbol{0}^{(M-1)R\times N'} & & & \boldsymbol{0}^{(M-2)R\times N'} & & & 0 \quad 0 \quad s_{M-1} \end{bmatrix} \tag{3.30}$$

can be expressed by rearranging the scaling factors such that the individual short frames in Eq. (3.26) are temporally aligned and scaled for an extended temporal context.

For the derivation of spectral refinement, this time-domain solution can be transferred into the frequency domain by applying a DFT matrix: $\hat{\boldsymbol{X}}(\ell) = \boldsymbol{D} \cdot \hat{\boldsymbol{x}}(\ell)$. Inserting the solution from Eq. (3.26)

$$\hat{\boldsymbol{X}}(\ell) = \boldsymbol{D} \cdot \boldsymbol{s} \cdot \begin{bmatrix} \boldsymbol{x}'(\ell) \\ \boldsymbol{x}'(\ell-1) \\ \vdots \\ \boldsymbol{x}'(\ell-(M-1)) \end{bmatrix} \tag{3.31}$$

and replacing the vector of stacked short frames by a vector of stacked low-resolution spectra that are transferred into the time domain by applying an inverse block-DFT matrix $\boldsymbol{D}_{\text{block}}^{-1}$ finally yields an expression

$$= \underbrace{\boldsymbol{D} \cdot \boldsymbol{s} \cdot \boldsymbol{D}_{\text{block}}^{-1}}_{\boldsymbol{S}} \cdot \begin{bmatrix} \boldsymbol{X}'(\ell) \\ \boldsymbol{X}'(\ell-1) \\ \vdots \\ \boldsymbol{X}'(\ell-(M-1)) \end{bmatrix} \tag{3.32}$$

that resembles the extension of the temporal context in the time domain: multiple low-resolution spectra $\boldsymbol{X}'(\ell)$ obtained from M successive short frames are stacked and processed with a spectral refinement matrix $\boldsymbol{S} \in \mathbb{C}^{N \times MN'}$ targeting on a refined spectrum $\hat{\boldsymbol{X}}(\ell)$ that approximates a high-resolution spectrum $\boldsymbol{X}(\ell)$ gained with a longer frame of length N.

The resulting matrix $\boldsymbol{S}$ subsuming the three matrix operations is sparse with most elements being close or equal to zero. Therefore, spectral refinement can be implemented very efficiently by means of subband filters as described in [Krini and Schmidt, 2007].

Extended auto-correlation function

In the same publication, Krini and Schmidt [2007] introduced an extended ACF (EACF) approach that further increases the effective frame length. Using this method, the period range that can be taken into account for pitch detection is not limited to the single frame's ACF. Instead, the range is extended by additionally considering (normalized) cross-correlation functions (CCFs)

$$CCF(\tau', \ell, \Delta\ell) = \frac{\sum_{k=0}^{N'-1} X'^{*}(k,\ell) \cdot X'(k, \ell-\Delta\ell) \cdot e^{2\pi j k \tau'/N'}}{\sqrt{\sum_{k=0}^{N'-1} \hat{\Phi}_{xx}(k,\ell) \cdot \sum_{k=0}^{N'-1} \hat{\Phi}_{xx}(k, \ell-\Delta\ell)}} \tag{3.33}$$

between the current frame ℓ and preceding frames $\ell - \Delta\ell$. Repetitive structures in a signal with period lengths that exceed the frame length are again indicated by a peak. The short ACF (expressed here by a CCF with $\Delta\ell = 0$) and multiple CCFs are accumulated to

$$EACF(\tau, \ell) = b_\tau \cdot \sum_{\Delta\ell=0}^{(M-1)/2} b'_{\tau-\Delta\ell \cdot R} \cdot CCF(\tau - \Delta\ell \cdot R, \ell, \Delta\ell) \tag{3.34}$$

which covers an extended range $0 \leq \tau \leq \frac{N}{2}$. The short frames are temporally aligned by mapping the EACF's τ to $\tau' = \tau - \Delta\ell \cdot R$ relative to the respective time intervals covered by the short CCFs. The CCF is weighted with $b'_{\tau'}$ in the relevant region $-\frac{N'}{2}+1 \leq \tau' \leq \frac{N'}{2}$. For other arguments, its contribution is set to zero.

In [Krini and Schmidt, 2007], the EACF was calculated based on a refined spectrum to capture very long pitch periods that were not resolved using pure spectral refinement. A linear transition between ACF and a single CCF was proposed for the weighting coefficients $b'_{\tau'}$.

The EACF approach, however, can also be applied directly to the short frames as described in [Graf et al., 2017b]. In this case, an advanced selection of weighting coefficients $b'_{\tau'}$ and b_τ is needed to shape the transitions between multiple CCFs. Estimation of the CCF involves a product operation in Eq. (3.33) that introduces a non-linear dependency on the windowed time-domain signal. Consequently, the effect of the window function cannot be analyzed independently of the signal which contrasts to the spectral refinement approach. Only the envelopes of ACF and CCF for a long and a short window can be determined assuming a constant excitation $x(n) = 1$. The long envelope then can be expressed as a cyclic auto-correlation

$$\tilde{w}_\tau = \sum_{\tilde{n}=0}^{N-1} w_{\tilde{n}} \cdot w_{(\tau+\tilde{n}) \bmod N} \tag{3.35}$$

of the original long window. Analogously, the short window's envelope $\tilde{w}'_{\tau'}$ can be determined. With the EACF according to Eq. (3.34), the long envelope should be reproduced when inserting the short envelopes. As a first step, a vector of weighting coefficients $\boldsymbol{b}' = [b'_{-\frac{N'}{2}+1}, b'_{-\frac{N'}{2}+2}, \cdots, b'_{\frac{N'}{2}}]$ is determined that compensates the short envelopes in such a way that a flat envelope equal to one is achieved

$$\boldsymbol{1}^{1\times R} = \boldsymbol{b}' \begin{bmatrix} \tilde{w}'_{-\frac{N'}{2}+1} & 0 & 0 & 0 \\ 0 & \tilde{w}'_{-\frac{N'}{2}+2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \tilde{w}'_{-\frac{N'}{2}+R} \\ \tilde{w}'_{-\frac{N'}{2}+R+1} & 0 & 0 & 0 \\ 0 & \tilde{w}'_{-\frac{N'}{2}+R+2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \tilde{w}'_{-\frac{N'}{2}+2R} \\ & & \vdots & \\ \tilde{w}'_{\frac{N'}{2}-R+1} & 0 & 0 & 0 \\ 0 & \tilde{w}'_{\frac{N'}{2}-R+2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \tilde{w}'_{\frac{N'}{2}} \end{bmatrix} \tag{3.36}$$

by the sum over the weighted overlapping short CCFs. A similar approach was discussed in [Withopf et al., 2012] for the design of synthesis filterbank windows with perfect reconstruction. The equation system in Eq. (3.36) is underdetermined, so there is no unique solution. As the columns are orthogonal, a simple solution

$$b'_{\tau'} = \frac{\tilde{w}'_{\tau'}}{\sum_{\Delta\ell=0}^{M-1} (\tilde{w}'_{(\tau'+\Delta\ell\cdot R) \bmod N'})^2} \tag{3.37}$$

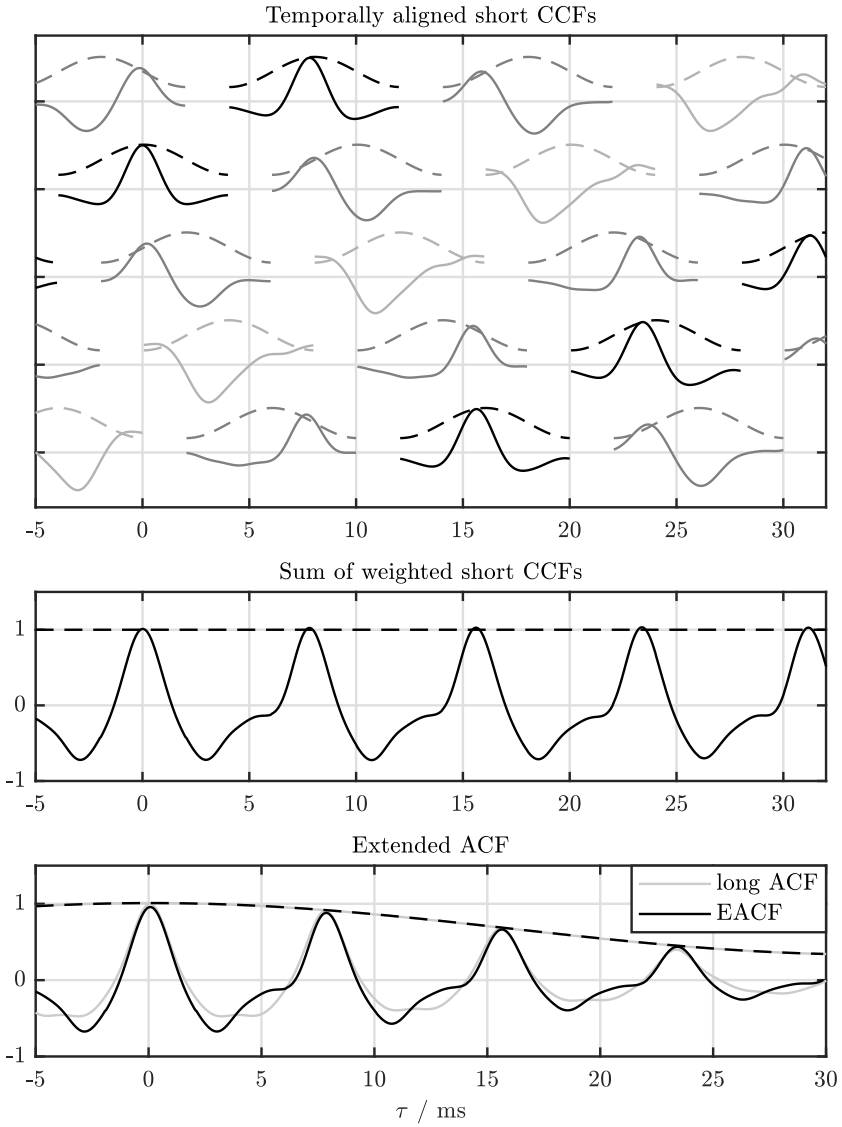


Figure 3.14: Approximation of a long ACF by a sum of temporally aligned and weighted short CCFs: CCFs between the current frame and multiple past frames are calculated. The envelopes (dashed lines) are reshaped such that the envelope after summation is flat. After applying the long envelope, the final EACF approximates the corresponding long ACF.

can be determined by multiplying both sides with the matrix's pseudo-inverse.

Based on this flat envelope, the desired long envelope corresponding to an ACF with length N can easily be approximated: since all flat elements are one, the long envelope $\tilde{w}_\tau$ can directly be applied as weighting coefficients b_τ as illustrated in Figure 3.14.

Even though it is possible to explore an extended range of pitch periods, EACF always takes the current short frame as a reference that is correlated with one or multiple previous frames. For very long pitch periods and short frames, the current frame may lie between two peaks of the harmonic excitation without capturing any of them. In this case, the instantaneous approximation misses the voiced speech. Temporal smoothing

$$\overline{EACF}(\tau, \ell) = \frac{1}{L_{\mathrm{EACF}}} \sum_{\tilde{\ell}=0}^{L_{\mathrm{EACF}}-1} EACF(\tau, \ell - \tilde{\ell}) \tag{3.38}$$

helps to bridge these dropouts. Choosing $L_{\mathrm{EACF}} = \frac{M}{2}$, almost the same temporal context is considered compared to the standard ACF with a longer window. Alternatively, the smoothing can be implemented more efficiently using an IIR filter.

A comparison of spectral refinement and extended ACF as depicted in Figure 3.15 shows that both methods are capable of approximating the long ACF with almost identical results. For the simulation, a sample rate $f_\mathrm{s} = 16\,\mathrm{kHz}$ with short frames of 128 samples and 75% overlap are chosen. A long frame length of 1024 samples is targeted. Both the short and the long frames are weighted with Hann windows. The corresponding VAD features as well as the pitch estimates are depicted in Figure 3.16.

The computational complexity of both approaches is on a similar level: for spectral refinement, $M \cdot N'/2 + M \cdot N'$ operations are required for the refinement [Krini and Schmidt, 2014] followed by an IFFT of order $N\,\mathrm{ld}(N)$ for calculating the long ACF based on the refined power spectrum. The effort for EACF in contrast is dominated by the individual short CCFs that require $M/2$ IFFTs of order $N'\,\mathrm{ld}(N')$.

Low-complexity algorithm

Motivated by the observation that a long ACF can sufficiently be approximated by means of multiple shorter CCFs, a novel approach for voiced speech detection was introduced in [Graf et al., 2017a]. Unlike the methods described before, the algorithm is capable of detecting harmonic components without explicitly searching for the pitch period. Contrariwise, the pitch period can optionally be obtained upon making the decision on voiced speech. The method is implemented completely in the frequency domain, such that the inverse DFTs in Eq. (3.33) can be omitted.

As discussed before, the content of two frames can be compared by means of a CCF. A peak of this function indicates a repetitive signal portion that occurs in both frames. For EACF, multiple CCFs are combined such that the range of pitch periods that can be taken into account for detection of voiced speech is extended. Subsequently, the most prominent peak can be found in the extended range analogously to a standard ACF with a long window. In contrast, the method described here first detects a peak for each pair of frames separately and combines the results afterward.

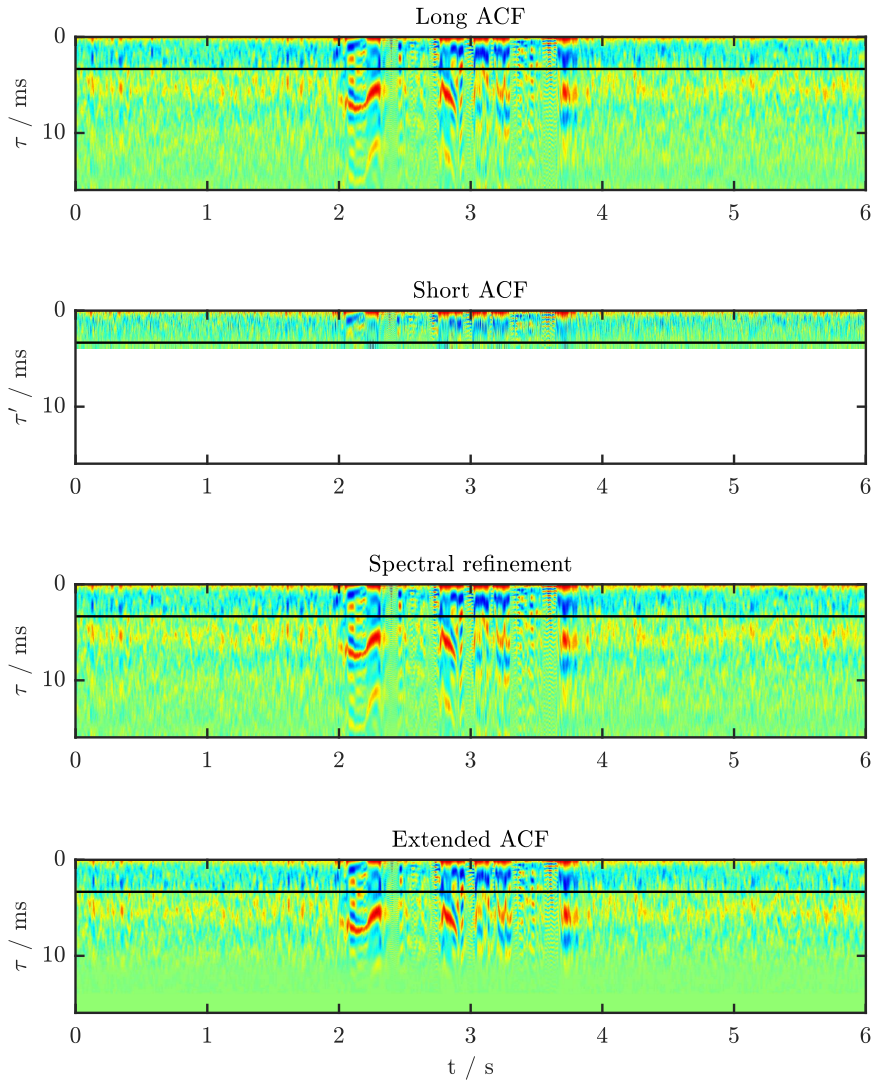


Figure 3.15: Spectral refinement vs. extended ACF: the long ACF reflects voiced speech that cannot be captured by a single short ACF. Both extension methods, spectral refinement and EACF are capable of approximating the ACF based on multiple short frames.

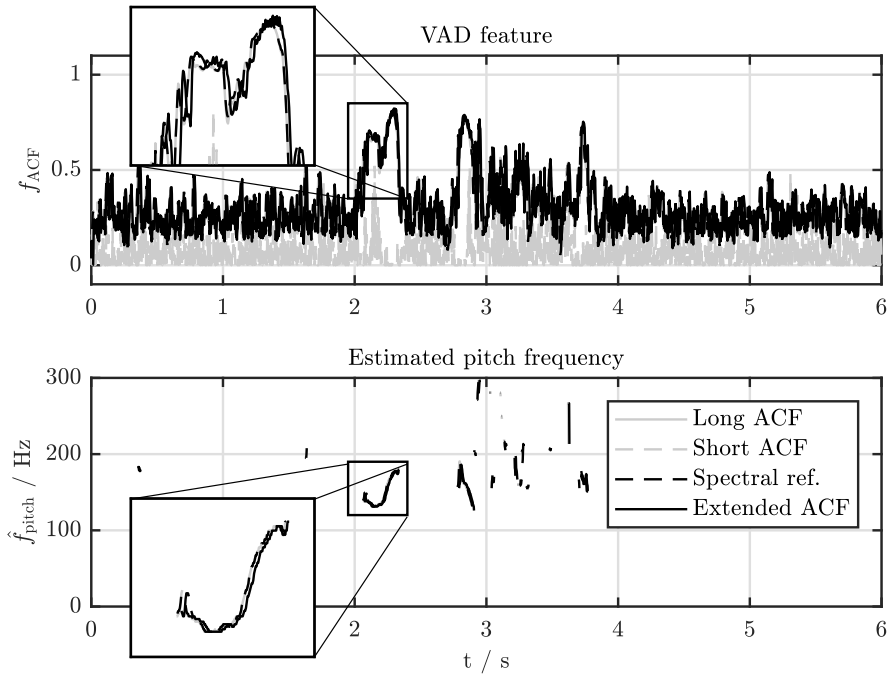


Figure 3.16: VAD feature and pitch estimated based on different ACF estimates: both spectral refinement and EACF approximate the long ACF well. Using these techniques, voiced speech can be detected that is inaccessible with the short frame's ACF.

Using a standard CCF, a peak indicates presence of a repetitive structure, however, its distinctiveness also depends on the signal's shape. The CCF does not account for the separation of speech according to the source-filter model into an impulse sequence as excitation and a filter that shapes the envelope. For the detection of periodic structures, only the period time of the impulse sequence is of interest but not the shape. Therefore, a generalized cross-correlation function (GCC) [Knapp and Carter, 1976]

$$GCC(\tau', \ell, \Delta\ell) = \frac{1}{N'} \sum_{k=0}^{N'-1} \underbrace{\frac{X'^{*}(k,\ell) \cdot X'(k,\ell-\Delta\ell)}{|X'^{*}(k,\ell) \cdot X'(k,\ell-\Delta\ell)|}}_{GCS(k,\ell,\Delta\ell)=e^{j\varphi(k,\ell,\Delta\ell)}} \cdot e^{2\pi jk\tau'/N'} \tag{3.39}$$

is more suitable for the detection as the peak get emphasized irrespectively of the shape as illustrated in Figure 3.17.

The measure relies only on the phase information $\varphi(k, \ell, \Delta\ell)$ of the cross-spectrum but removes the magnitude information by spectral whitening. A phase that is perfectly linear over frequency maximizes the GCC according to

$$\frac{1}{N'} \sum_{k=-N'/2+1}^{N'/2-1} e^{-2\pi jk\tau'_{\mathrm{p}}/N'} \cdot e^{2\pi jk\tau'/N'} \approx \mathrm{sinc}(\tau' - \tau'_{\mathrm{p}}) \tag{3.40}$$

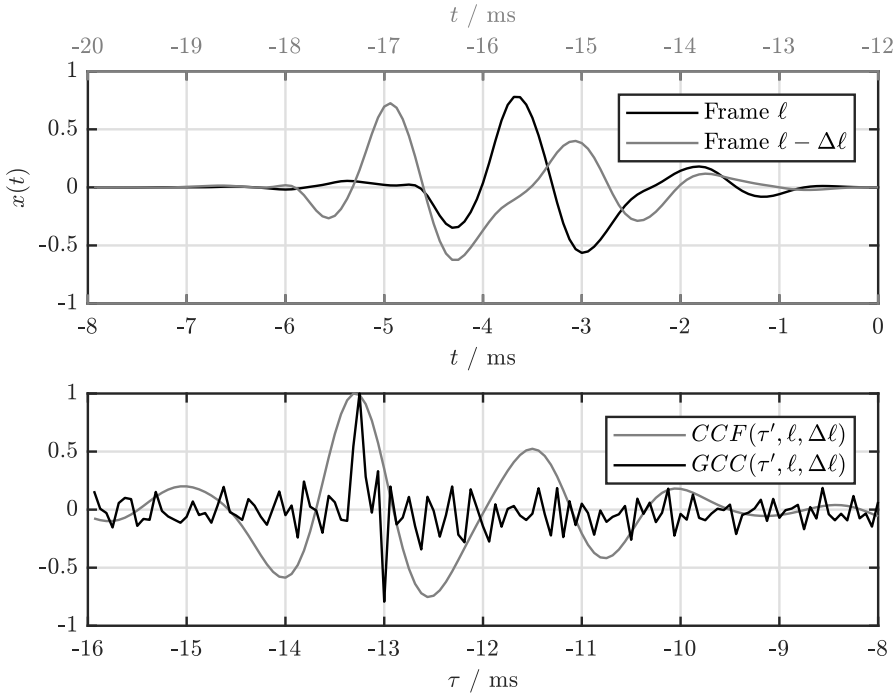


Figure 3.17: Cross-correlation vs. generalized cross-correlation: the voiced speech content captured by two short frames is shaped similarly but temporally shifted. Both CCF and GCC reflect this shift in their maximum positions, however, CCF still depends on the shape. In contrast, GCC emphasizes the peak irrespective of the original shape.

where a small leakage of sinc functions shifted by integer multiples of N' is omitted for brevity. For $\tau' = \tau'_{\mathrm{p}}$, the sinc function

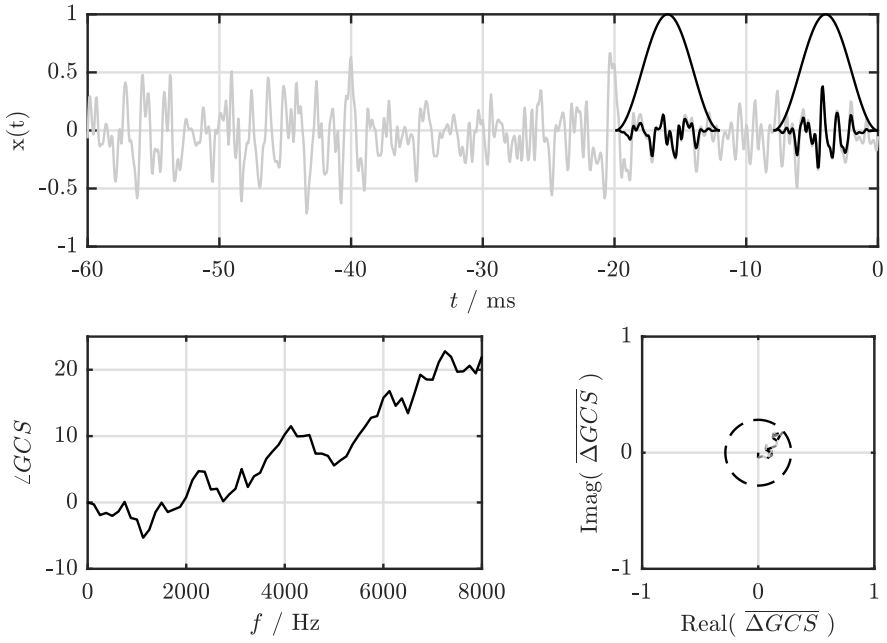
$$\operatorname{sinc}(x) = \begin{cases} \frac{\sin(\pi x)}{\pi x} & \text{for } x \neq 0 \\ 1 & \text{else} \end{cases} \tag{3.41}$$

assumes its maximum value one whereas for arguments $\tau' \neq \tau'_{\mathrm{p}}$, the value is negligibly small.

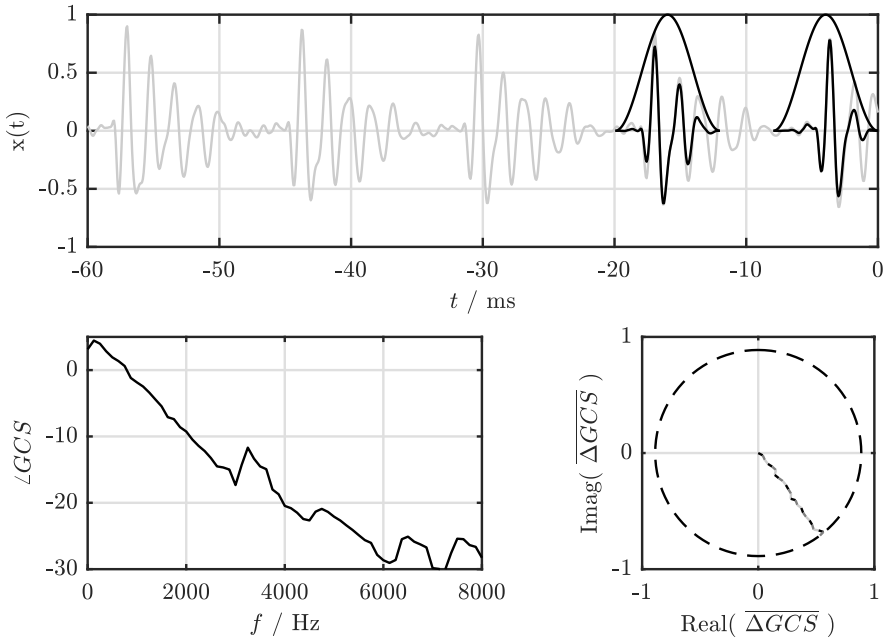
Testing for a linear phase of the cross-spectrum can hence be employed for voiced speech detection. A product of the complex-valued and a complex conjugated normalized cross-spectra at two frequencies

$$\begin{aligned} \Delta GCS(k, \ell, \Delta\ell) &= GCS(k, \ell, \Delta\ell) \cdot GCS^*(k-1, \ell, \Delta\ell) \\ &= e^{j\varphi(k,\ell,\Delta\ell) - j\varphi(k-1,\ell,\Delta\ell)} \\ &= e^{j\Delta\varphi(k,\ell,\Delta\ell)} \end{aligned} \tag{3.42}$$

reflects the phase difference $\Delta\varphi(k, \ell, \Delta\ell)$ between the two bins k and $k-1$. For a linear phase, equal phase differences are observed for all frequencies. Correspondingly, the



(a) Unvoiced



(b) Voiced

Figure 3.18: Illustration for unvoiced and voiced speech: the cross-spectrum's phase $\angle GCS$ for unvoiced speech appears randomly distributed whereas it approaches a linear slope over frequency for voiced speech. Correspondingly, the accumulated complex-valued measure's magnitude $\left|\overline{\Delta GCS}\right|$ is low for unvoiced speech and assumes higher values close to one for voiced speech. The angle $\angle\overline{\Delta GCS}$ further relates to the pitch period.

complex product is constant over frequency. Considering the complex value here is advantageous since ambiguities of the phase differences $\Delta\varphi \hat{=} \Delta\varphi \pm 2\pi$ are avoided.

A weighted sum over frequency

$$\overline{\Delta GCS}(\ell, \Delta\ell) = \frac{\sum_{k=1}^{K-1} w(k, \ell, \Delta\ell) \cdot \Delta GCS(k, \ell, \Delta\ell)}{\sum_{k=1}^{K-1} w(k, \ell, \Delta\ell)} \tag{3.43}$$

reflects the degree of linearity. Its magnitude $\left|\overline{\Delta GCS}(\ell, \Delta\ell)\right|$ is maximized to one for a linear phase cross-spectrum. During absence of harmonic components, the magnitude is lower as the phase differences are distributed rather randomly. In Figure 3.18, the sum is exemplified for a voiced and an unvoiced speech spectrum.

The weighting factors are chosen such that frequency regions that are relevant for voiced speech are emphasized whereas frequencies that are dominated by noise are excluded. A fixed weighting function based on the typical distribution of voiced speech could be employed. Here, it is chosen dynamically based on the spectral magnitude

$$w(k, \ell, \Delta\ell) = \begin{cases} |X'(k, \ell)| & \text{for } 50\,\text{Hz} < k f_\text{s}/N' < 4\,\text{kHz} \\ 0 & \text{else} \end{cases} \tag{3.44}$$

and it is further limited to the relevant frequency region. This weighting function relies on the assumption that the highest spectral bins correspond to speech. Alternatively, one could employ a Wiener filter to put a stronger emphasis on non-stationary signal components.

Using Eq. (3.43), only a pair of frames ℓ and $\ell - \Delta\ell$ is considered for the estimate. Including more frames increases the robustness against noise. Recursive smoothing

$$\overline{\overline{\Delta GCS}}(\ell, \Delta\ell) = \alpha_\text{GCS} \cdot \overline{\overline{\Delta GCS}}(\ell - \Delta\ell, \Delta\ell) + (1 - \alpha_\text{GCS}) \cdot \overline{\Delta GCS}(\ell, \Delta\ell) \tag{3.45}$$

along time can be employed to consider previous results. Since peaks of the harmonic excitation with a distance in the range of $\Delta\ell$ frames are investigated, this distance is also employed for smoothing. Hence signal portions of the frames $\ell, \ell - \Delta\ell, \ell - 2\Delta\ell, \cdots$ are incorporated in the smoothed value.

Magnitude and phase of the smoothed result are related to the degree of linearity and to the temporal shift between the peaks, respectively. The magnitude

$$p_\text{v}(\ell, \Delta\ell) = \left|\overline{\overline{\Delta GCS}}(\ell, \Delta\ell)\right| \tag{3.46}$$

is limited to the interval $0 < p_\text{v}(\ell, \Delta\ell) \leq 1$ where high values close to one indicate a linear phase. High values, hence, can be associated with presence of voiced speech.

Lower values, however, do not necessarily imply absence of voiced speech. Since the frames are very short, they might lie between two impulses of the repetitive excitation without capturing any of them. To overcome this effect and to fuse the results of the different pairs of frames, a post processing is applied.

Holding maxima over $\Delta\ell$ frames

$$\overline{p}_{\mathrm{v}}(\ell, \Delta\ell) = \max_{0 \leq \tilde{\ell} < \Delta\ell} \left(p_{\mathrm{v}}(\ell - \tilde{\ell}, \Delta\ell) \right) \tag{3.47}$$

prevents from gaps that may occur when the current frame is located between two peaks.

The different pairs of frames correspond to different candidates for regions of the pitch period. Finding the most probable pair

$$\widehat{\Delta\ell}(\ell) = \operatorname*{argmax}_{\Delta\ell} \left(\overline{p}_{\mathrm{v}}(\ell, \Delta\ell) \right) \tag{3.48}$$

is the last step for voiced speech detection. Based on this pair, the feature

$$f_{\mathrm{LPCS}}(\ell) = \overline{p}_{\mathrm{v}}(\ell, \widehat{\Delta\ell}(\ell)) \tag{3.49}$$

representing a linear phase cross-spectrum can be derived.

In case that voiced speech is detected, the pitch period can be derived based on the slope of the linear phase. Replacing the magnitude in Eq. (3.46) by an angle operator

$$\widehat{\Delta\varphi}(\ell) = \angle\overline{\overline{\Delta GCS}}(\ell, \widehat{\Delta\ell}(\ell)) \tag{3.50}$$

allows for estimating the slope that corresponds to a temporal shift τ'. This value, however, is gained by correlating the content of two frames corresponding to samples of the signal at two different time intervals. The time shift $\Delta\ell \cdot R$ between both frames is therefore not reflected and must be considered separately

$$\hat{\tau}_{\mathrm{pitch}}(\ell) = \frac{\widehat{\Delta\varphi}(\ell)}{2\pi} N' + \widehat{\Delta\ell}(\ell) \cdot R \tag{3.51}$$

for determining the final estimate of the pitch period.

For the speech example, the novel approach detects voiced parts with a similar accuracy as a baseline ACF with a long frame length as depicted in Figure 3.19. Also the pitch frequency can be estimated, however, the variance is higher compared to the baseline.

3.2.5 Formant structure

Excitation-related aspects of human speech production have been discussed in the previous sections. The corresponding features reflecting energy and voicing in an audio wave are very important for the detection of speech: in many situations both the power and a harmonic structure set apart very well from background noise. In contrast, for speech recognition, the spectral envelope constitutes the main cue that distinguishes one phone from another [Arora and Reetz, 2017]. The different phones are characterized by certain frequency regions that are emphasized by varying settings of cavities in the human vocal tract. This formant structure can also be employed for speech detection [Yoo et al., 2015].

In this section, different traditional representations of the spectral envelope are summarized. All features discussed here target on modeling this envelope by means of multiple

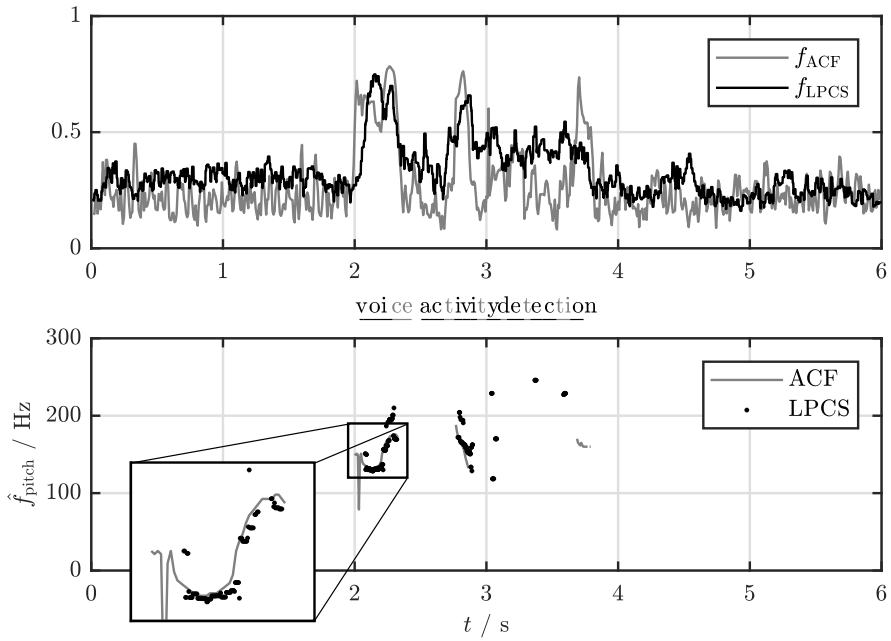


Figure 3.19: VAD feature based on a linear phase and corresponding pitch estimate: the novel feature detects voiced speech similarly as the ACF-based approach. Also the pitch frequency may be estimated, however, the variance of the estimate is higher compared to the ACF.

parameters. For the detection of speech, these feature vectors can be combined to a scalar decision variable, e.g., using a neural network as it will be discussed later in Section 3.4.

According to the source-filter model, the vocal tract filter spectrally shapes the voiced or unvoiced excitation signal and applies the spectral envelope. Spectral envelope estimation based on linear predictive coding (LPC) relies on the opposite direction: an FIR-type filter $H_{\mathrm{LPC}}(\ell)$ is determined that levels out the spectral envelope of the signal. When this filter is applied to the speech signal the resulting output is whitened and correlations are reduced. Conversely, the signal's spectral envelope can be reproduced by filtering a white noise signal with the corresponding IIR filter

$$H_{\mathrm{LPC}}^{-1}(z,\ell) = \frac{1}{H_{\mathrm{LPC}}(z,\ell)} = \frac{1}{1 - \sum_{\tilde{n}=1}^{N_{\mathrm{LPC}}} a_{\mathrm{LPC},\tilde{n}}(\ell) \cdot z^{-\tilde{n}}} \tag{3.52}$$

expressed here by the z-transformation.

The feature vector of filter coefficients

$$\boldsymbol{f}_{\mathrm{LPC}}(\ell) = \underset{\boldsymbol{a}_{\mathrm{LPC}}(\ell)}{\operatorname{argmin}} \; E\left\{ \left(x(\ell R) - \sum_{\tilde{n}=1}^{N_{\mathrm{LPC}}} a_{\mathrm{LPC},\tilde{n}}(\ell) \cdot x(\ell R - \tilde{n}) \right)^2 \right\} \tag{3.53}$$

can be found using linear prediction: the coefficients $\boldsymbol{a}_{\mathrm{LPC}}(\ell) = [a_{\mathrm{LPC},1}(\ell), \cdots, a_{\mathrm{LPC},N_{\mathrm{LPC}}}(\ell)]^T$ are determined such that the mean squared error between the current sample value $x(\ell R)$ and a value predicted based on previous samples $x(\ell R - 1), \cdots, x(\ell R - N_{\mathrm{LPC}})$ is minimized. Implementations typically estimate the signal's ACF $E\{x(\ell R) \cdot x(\ell R - \tilde{n})\}$ and express the prediction by means of Yule-Walker equations. This equation system can be solved efficiently using the Levinson-Durbin algorithm as described, e.g., in [Hänsler and Schmidt, 2004]. As only few filter coefficients, e.g., $N_{\mathrm{LPC}} = 16$ for the experiments in this thesis, are sufficient to describe the speech signal's envelope, this approach is commonly used for speech encoding.

In [Rabiner and Sambur, 1977], the distance between a mean vector and the observed coefficients vector was employed for voiced/unvoiced/silence classification. In this work, the feature vector is processed with a neural network for speech detection.

The relation between plain LPC coefficients and the spectral envelope is quite complex. Given the LPC coefficients, the transfer function Eq. (3.52) has to be evaluated for multiple $z = e^{j\Omega}$ on the unit circle to extract the spectral shape which is a drawback of this approach. Furthermore, even small deviations of the filter coefficients may severely change the corresponding spectrum which makes this representation vulnerable against quantization errors.

To overcome these drawbacks, line spectral frequencies (LSF) can be extracted based on the LPC coefficients [Kabal and Ramachandran, 1986]. For this, the LPC filter polynomial is expressed as a sum of two polynomials

$$H_{\mathrm{LPC}}(z,\ell) = \frac{1}{2}\Bigg(\overbrace{(1 - a_{N_{\mathrm{LPC}}}) - (a_1 + a_{N_{\mathrm{LPC}}-1})z^{-1} \cdots - (a_{N_{\mathrm{LPC}}} - 1)z^{-N_{\mathrm{LPC}}}}^{H_{\mathrm{LPC,symm}}(z,\ell)} + \underbrace{(1 + a_{N_{\mathrm{LPC}}}) - (a_1 - a_{N_{\mathrm{LPC}}-1})z^{-1} \cdots - (a_{N_{\mathrm{LPC}}} + 1)z^{-N_{\mathrm{LPC}}}}_{H_{\mathrm{LPC,anti}}(z,\ell)} \Bigg) \tag{3.54}$$

with symmetric and antisymmetric coefficients respectively. As a beneficial property, all roots $H_{\mathrm{symm/anti}}(z_r) = 0$ of both polynomials are located on the unit circle $z_r = e^{j\Omega_r}$. The filter can hence be described equivalently by the LSFs

$$\boldsymbol{f}_{\mathrm{LSF}}(\ell) = [\Omega_0(\ell), \Omega_1(\ell), \cdots, \Omega_{N_{\mathrm{LPC}}-1}(\ell)]^T \tag{3.55}$$

given by the angles Ω_r of the complex-valued roots. This representation is much closer related to the spectral shape: as illustrated in Figure 3.20, the LSFs cluster around the formant frequencies that are emphasized in the vocal tract. As LSFs are further robust against small deviations, many speech coding systems rely on this representation. For example, the standardized VAD approach according to ITU-T Recommendation G.729 Annex B [ITU, 1996] considers the LSFs to capture the formant structure of speech.

Another feature representing the spectral shape again relies on the cepstrum Eq. (3.22) that was already considered for the detection of a voiced excitation structure. As mentioned before, the cepstrum converts the convolutive mixture in the time domain between

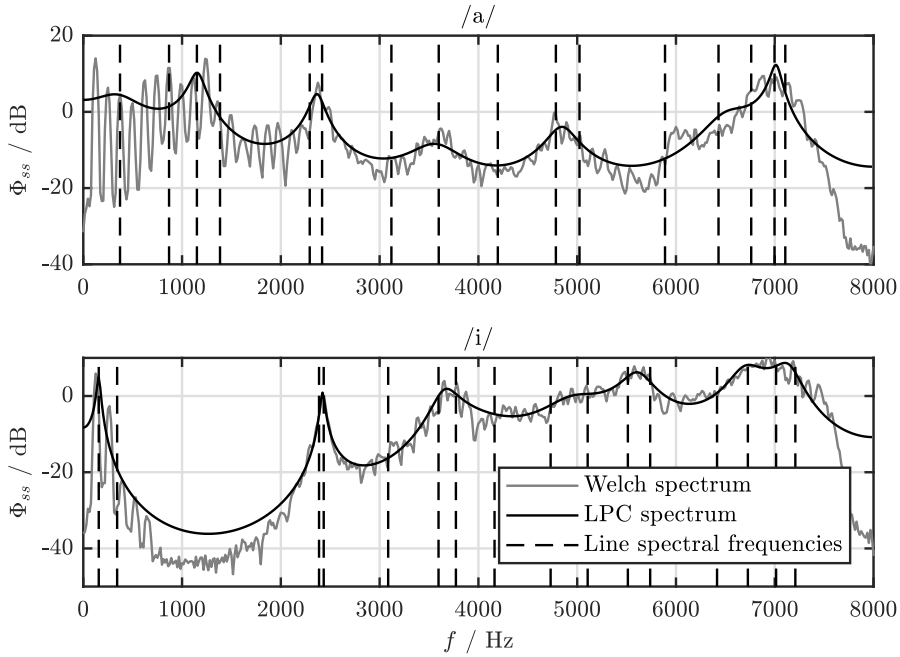


Figure 3.20: Linear predictive coding and line spectral frequencies: the IIR filter configured by LPC coefficients models the spectral envelope of speech. It is depicted here for two different phones /a/ and /i/. The corresponding LSFs cluster around the spectral peaks: prominent peaks correspond to denser LSFs. A Welch spectrum is plotted for reference that does not distinguish between excitation and envelope.

an excitation signal and the vocal tract filter into an additive mixture of two cepstral components. The lower order cepstral coefficients

$$\boldsymbol{f}_{\mathrm{CEP2}}(\ell) = [CEP(0,\ell), CEP(1,\ell), \cdots, CEP(N_{\mathrm{CEP2,low}}-1,\ell)]^T \tag{3.56}$$

corresponding to slowly fluctuating components represent the spectral envelope as illustrated in Figure 3.21 whereas the higher order coefficients carry information on the harmonic excitation. The feature [Haigh and Mason, 1993] is closely related to the spectral envelope which is beneficial for the detection of speech. Calculation of the feature is straightforward particularly when the power spectrum is already available. However, the logarithm for each frequency bin and the DCT are computationally expensive.

In the human auditory system, the resolution of distinguishable frequencies decreases with increasing frequency. Signal processing algorithms can adopt this non-linear perception of frequency using a mel scale instead of a linear frequency axis [Wang and Brown, 2006]. Spectral bins equidistant on the linear frequency axis can be accumulated to mel bands that get broader with increasing frequency as illustrated in Figure 3.22. This operation compresses the spectrum while preserving the relevant envelope information. Based

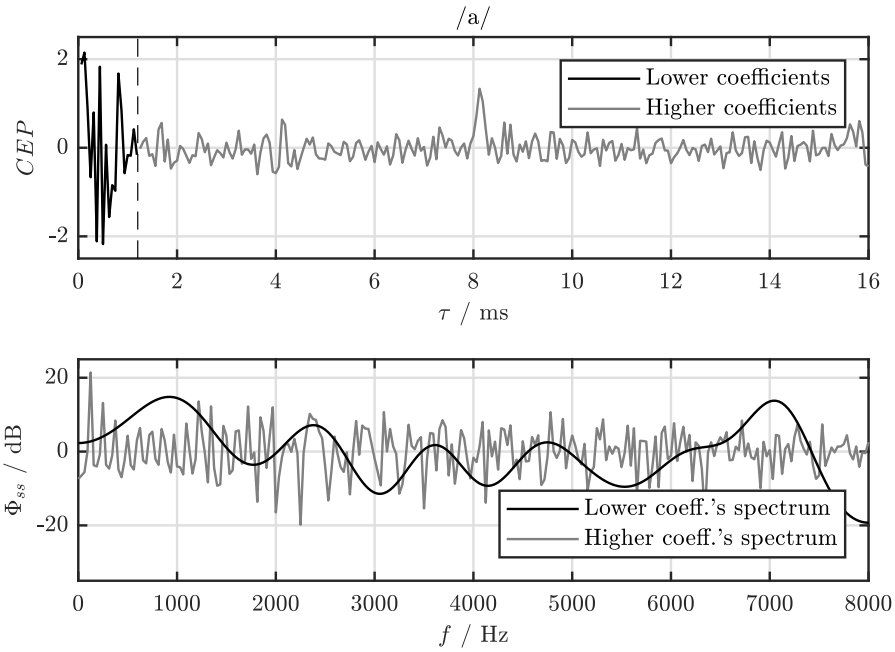


Figure 3.21: Cepstrum and power spectrum: similar to an ACF, a peak in the higher cepstral coefficients indicates the pitch period (8.1 ms in this example) of the voiced excitation. The lower coefficients reflect the spectral envelope which becomes evident when calculating the corresponding spectra for both parts separately.

on this compressed spectrum, mel-frequency cepstral coefficients (MFCCs)

$$\boldsymbol{f}_{\mathrm{MFCC}}(\ell) = [MFCC(0,\ell), MFCC(1,\ell) \cdots, MFCC(N_{\mathrm{MFCC}}-1,\ell)]^T \tag{3.57}$$

can be calculated for VAD [Kinnunen et al., 2007] analog to Eq. (3.22) by applying a logarithm to the mel frequency spectrum followed by a DCT. Some publications, e.g., [Hoyt and Wechsler, 1994] alternatively propose applying the logarithm before the aggregation to mel bands.

3.3 Suprasegmental properties: Sequences of phones

In the previous section, only short intervals were investigated for the detection of speech. The instantaneous characteristics of different phones were discussed without considering their temporal context.

Human speech, however, is based on the concatenation of different phones in order to utter the message. Suprasegmental properties of speech [Hirst, 2006] widen the perspective from isolated phones to sequences of concatenated phones. Since a longer temporal context is considered, these properties typically have a lower overlap with the properties of noise. A higher robustness against noise can therefore be expected from features that exploit

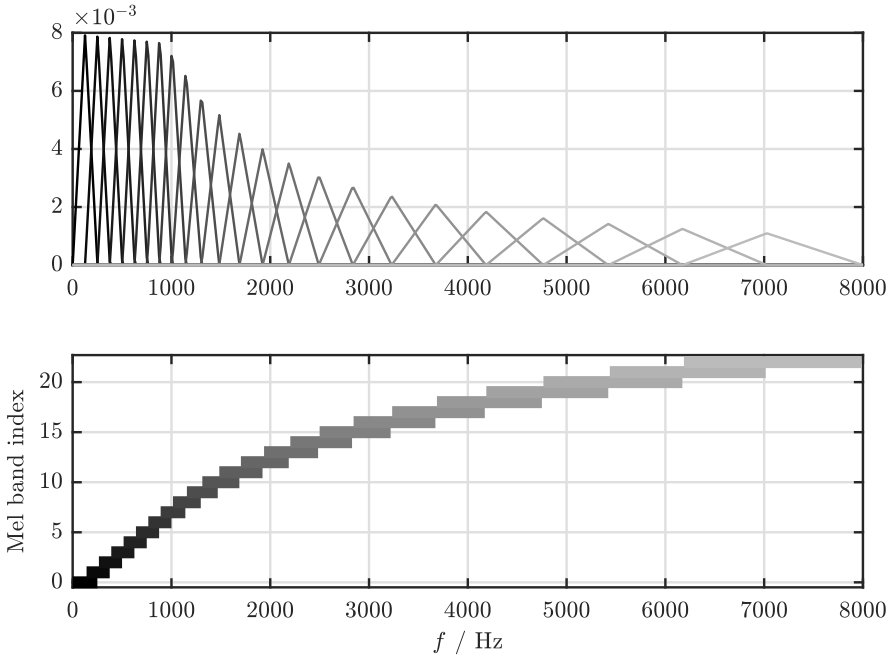


Figure 3.22: Mapping of linear frequency axis to mel bands: groups of frequencies are aggregated using triangular windows that get broader with increasing frequency. This frequency-dependent resolution mimics the human perception of frequency.

suprasegmental properties. However, the increased temporal context goes along with an increased latency of the features: compared to the instantaneous properties of separated phones, more time elapses until the temporal evolution is observable.

3.3.1 Stationarity

While a speech signal can be assumed to be short-time stationary for the duration of a single phone, the signal appears non-stationary when a longer temporal context is considered that spans over a sequence of multiple phones. As different phones are concatenated, the signal's statistics vary over time. This property of speech sets apart well from stationary background noise.

Different features were introduced in literature that reflect the temporal variability of speech signals. Ghosh et al. [2011] employed the temporal entropy

$$H(k,\ell) = -\sum_{\tilde{\ell}=0}^{L_{\mathrm{LTSV}}-1} \frac{\hat{\Phi}_{xx}(k,\ell-\tilde{\ell})}{\overline{\Phi}_{xx}(k,\ell)} \cdot \log\left(\frac{\hat{\Phi}_{xx}(k,\ell-\tilde{\ell})}{\overline{\Phi}_{xx}(k,\ell)}\right) \tag{3.58}$$

to measure the long-term signal variability (LTSV). The power spectrum $\hat{\Phi}_{xx}(k,\ell)$ is normalized with $\overline{\Phi}_{xx}(k,\ell) = \sum_{\tilde{\ell}=0}^{L_{\mathrm{LTSV}}-1} \hat{\Phi}_{xx}(k,\ell-\tilde{\ell})$ such that each frequency bin can be inter-

preted as a probability distribution over L_{LTSV} frames. Rearranging this equation to

$$H(k,\ell) = -\frac{\sum_{\tilde{\ell}=0}^{L_{\mathrm{LTSV}}-1} \hat{\Phi}_{xx}(k,\ell-\tilde{\ell}) \cdot \log\left(\hat{\Phi}_{xx}(k,\ell-\tilde{\ell})\right)}{\overline{\Phi}_{xx}(k,\ell)} + \log\left(\overline{\Phi}_{xx}(k,\ell)\right) \tag{3.59}$$

allows for a more efficient calculation with less divisions per frame.

The temporal entropy is maximized for stationary signals where the power spectrum does not change over time. On the other hand, non-stationary signals, such as speech, typically show lower values of the entropy.

The final broadband LTSV feature is defined as the variance over frequency

$$f_{\mathrm{LTSV}}(\ell) = \frac{1}{K}\sum_{k=0}^{K-1}\left(H(k,\ell) - \overline{H}(\ell)\right)^2 \quad \text{with} \quad \overline{H}(\ell) = \frac{1}{K}\sum_{k=0}^{K-1} H(k,\ell). \tag{3.60}$$

The variance drops when similar entropy values are observed in all frequency bins. Consequently, the feature value is low for stationary signals and increases when non-stationary parts lower the entropy for some frequencies. Tsiartas et al. [2013] introduced a multi-band LTSV. This extended approach relies on variances for multiple frequency bands instead of a single variance over the full frequency spectrum.

Instead of the entropy's variance, Ma and Nishihara [2013] proposed calculating an averaged long-term (spectral) flatness measure

$$f_{\mathrm{LSFM}}(\ell) = \frac{1}{K}\sum_{k=0}^{K-1} \log\left(\frac{\left(\prod_{\tilde{l}=0}^{L_{\mathrm{LSFM}}-1} \hat{\Phi}_{xx}(k,\ell-\tilde{\ell})\right)^{1/L_{\mathrm{LSFM}}}}{\left(\sum_{\tilde{l}=0}^{L_{\mathrm{LSFM}}-1} \hat{\Phi}_{xx}(k,\ell-\tilde{\ell})\right)/L_{\mathrm{LSFM}}}\right) \tag{3.61}$$

that captures the temporal variability of the signal by the ratio of geometric and arithmetic mean over L_{LSFM} frames. This ratio always assumes values less than or equal to one, hence the final measure is negative or zero for perfect stationarity. For speech and other non-stationary signal components, the value decreases towards lower negative numbers.

Features that rely on the long-term non-stationarity of speech work well for stationary background noises. However, like energy-based features, it has to be expected that these features are also quite sensitive to non-stationary interferences. In the following sections, features will be discussed that overcome this problem by taking into account more characteristic properties of speech.

3.3.2 Modulation

Human listeners are capable of perceiving speech even in severe noise conditions. Investigating the human speech perception and finding features that partially mimic the underlying mechanisms hence appears promising.

One important factor in human speech perception is the characteristic modulation of speech: the human auditory system reacts particularly sensitively to modulation frequencies in the range of 4 Hz [Edwards and Chang, 2013]. This frequency corresponds to the typical syllable rate of speech [Loizou, 2013].

Interferences usually exhibit modulation structures that differ from the modulation of speech even in cases where other properties are similar. Modulation-based features for VAD can hence be expected to distinguish speech even from highly challenging non-stationary interferences.

Compared to the features discussed before, a longer temporal context in the range of a second is needed for the extraction of modulation. Different approaches are known from literature to assess the spectrum's temporal evolution by means of a spectrogram: buffering multiple frames of the plain spectrum $\hat{\Phi}_{xx}(k,\ell)$ was employed in [Hsu et al., 2013] to capture this evolution. Perceptually motivated representations of the spectrogram were proposed by many authors to compress the amount of data: mel-frequencies [Scheirer and Slaney, 1997] or similar transformations [Tchorz and Kollmeier, 1999] can be applied to accumulate multiple spectral bins for W wider frequency bands. The resulting spectrogram $\tilde{\Phi}_{xx}(w,\ell)$ focuses on the most distinctive frequency regions which lowers the computational complexity due to the reduced number of bands $W < K$. Mesgarani et al. [2006] even introduced a model of the early-stage auditory system for an auditory spectrogram.

Even though speech and music are both non-stationary signals, their rhythm typically differs [Ding et al., 2017]. Therefore, Scheirer and Slaney [1997] considered also modulations in the range of 4 Hz for a feature

$$f_{\mathrm{M4Hz}}(\ell) = \frac{1}{W}\sum_{w=0}^{W-1}\frac{\Psi_{xx,4\,\mathrm{Hz}}(w,\ell)}{\tilde{\Phi}_{xx}^{2}(w,\ell)} \tag{3.62}$$

targeting on discriminating speech from noise. This feature relies on a filtered spectrogram $\Psi_{xx,4\,\mathrm{Hz}}(w,\ell)$ where temporal fluctuations of the spectrum with frequencies around 4 Hz are emphasized. The corresponding filter can easily be implemented, e.g., using an IIR filter. However, as only one filter dedicated to 4 Hz is applied to each band, the approach is limited to a specific syllable rate.

The superior robustness in challenging situations gave rise to a growing interest in modulation features for speech detection during the last decade. Using powerful machine learning techniques, such as neural networks, classifying multi-dimensional modulation patterns is nowadays possible [Hsu et al., 2013]. A wider range of modulation frequencies may be taken into account by applying multiple filters in parallel instead of just a single filter at 4 Hz.

For amplitude modulation spectrograms (AMSs) [Tchorz and Kollmeier, 1999], L_{AMS} spectra are buffered and processed with a DFT over time

$$\Psi_{xx,\mathrm{AMS}}(w,\ell,\kappa) = \left|\sum_{\tilde{\ell}=0}^{L_{\mathrm{AMS}}-1}\tilde{\Phi}(w,\ell-\tilde{\ell})\cdot e^{-j2\pi\kappa\tilde{\ell}/L_{\mathrm{AMS}}}\right|^{2} \tag{3.63}$$

to capture multiple modulation frequencies κ at once. Results for $L_{\mathrm{AMS}}/2$ modulation frequencies per band normalized w.r.t. the temporal average can be collected in vectors

$$\boldsymbol{\Psi}_{xx,\mathrm{AMS}}(w,\ell) = \frac{\left[\Psi_{xx,\mathrm{AMS}}(w,\ell,1),\cdots,\Psi_{xx,\mathrm{AMS}}(w,\ell,L_{\mathrm{AMS}}/2)\right]^{T}}{\Psi_{xx,\mathrm{AMS}}(w,\ell,0)} \tag{3.64}$$

that are stacked for the final feature

$$\boldsymbol{f}_{\mathrm{AMS}}(\ell) = \begin{bmatrix} \boldsymbol{\Psi}_{xx,\mathrm{AMS}}(0,\ell) \\ \vdots \\ \boldsymbol{\Psi}_{xx,\mathrm{AMS}}(W-1,\ell) \end{bmatrix} \tag{3.65}$$

comprising $W \cdot L_{\mathrm{AMS}}/2$ values in total. Bach et al. [2010] chose a long window length of 1 s to capture the relevant modulation frequencies. They considered 29 modulation frequencies and 17 spectral bands resulting in a 493-dimensional feature vector. In Figure 3.23, AMSs are exemplified for a speech and a non-speech signal portion.

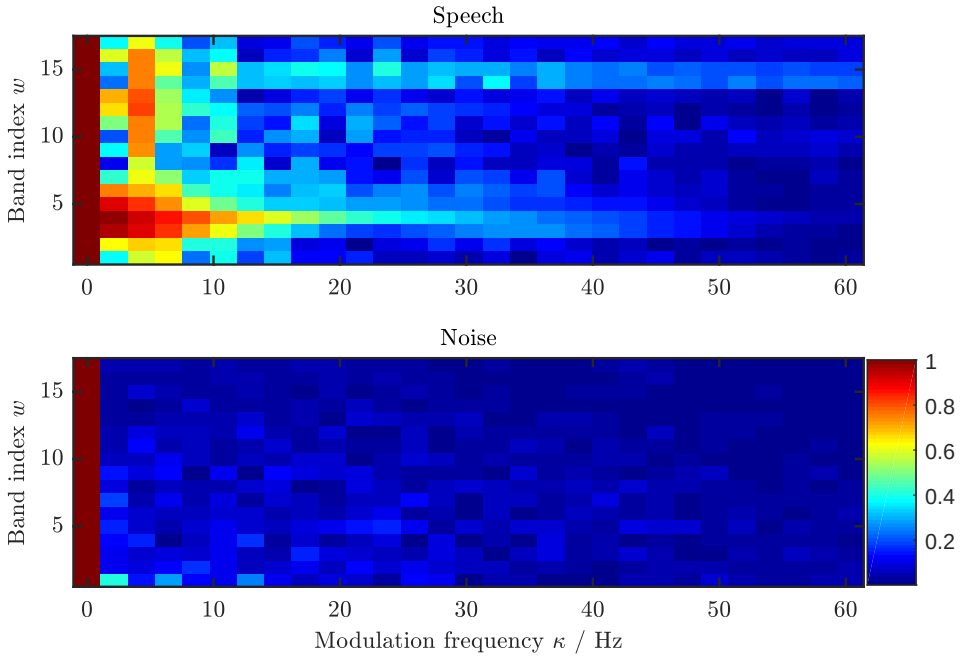


Figure 3.23: Illustration of amplitude modulation spectrogram: the AMS reflects modulations of speech that typically lie in the range of 4 Hz. For the noise example, no such modulations are visible.

The AMS represents temporal modulations more generally but still does not consider dependencies between multiple spectral bands. The temporal evolution of power is captured, however, the feature disregards the harmonic and formant structures of speech. To assess also these characteristic properties of speech, modulation patterns along time and frequency can jointly be taken into account by means of spectro-temporal modulation (STM) [Ezzat et al., 2007].

Certain modulation structures are emphasized in a convolved spectrogram

$$\tilde{\Phi}_{xx,\mathrm{STM}}(k,\ell,\omega,\Omega) = \sum_{\tilde{k}} \sum_{\tilde{\ell}} \hat{\Phi}_{xx}(k-\tilde{k},\ell-\tilde{\ell}) \cdot STMF(\tilde{k},\tilde{\ell},\omega,\Omega) \tag{3.66}$$

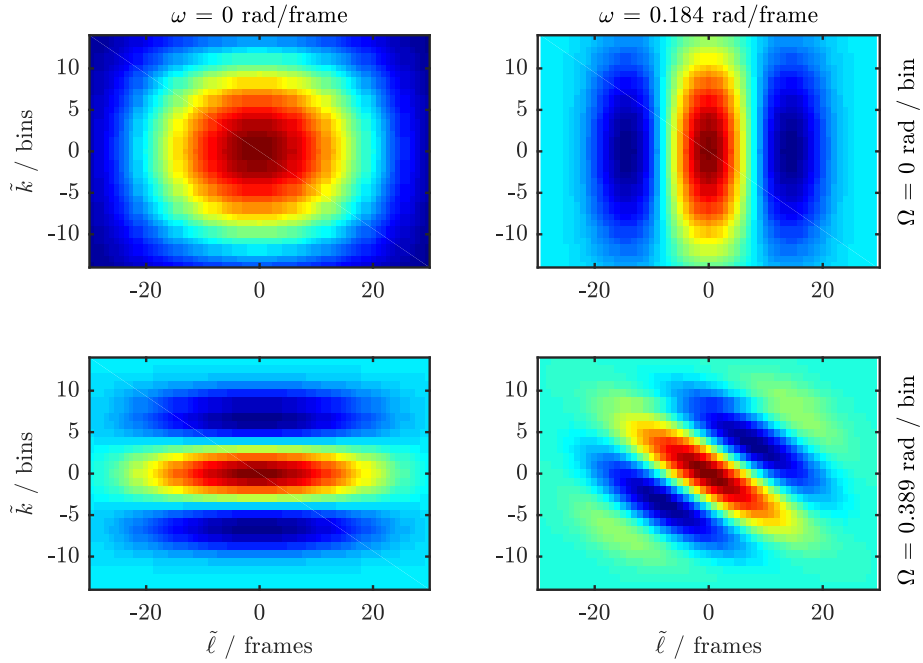


Figure 3.24: Spectro-temporal modulation filters: the two-dimensional STMF patterns are convolved with the spectrogram to find characteristic modulations along time and frequency.

based on two-dimensional spectro-temporal modulation filters (STMFs). These patterns can be configured using two parameters: a rate parameter ω specifies the modulation frequency along time that is emphasized by the pattern whereas modulations along frequency are addressed by a scale parameter Ω as illustrated in Figure 3.24. The resulting two-dimensional FIR filter is softly truncated using an envelope function, e.g., a Hann-window [Schädler et al., 2012].

Usually, multiple rates and scales are evaluated per frequency region. So the total number of dimensions for the feature vector

$$\begin{aligned}\boldsymbol{f}_{\mathrm{STM}}(\ell) = \Big[& \boldsymbol{\Psi}_{xx,\mathrm{STM}}(\ell, \omega_0, \Omega_0), && \cdots, && \boldsymbol{\Psi}_{xx,\mathrm{STM}}(\ell, \omega_0, \Omega_{N_{\mathrm{scale}}-1}), \\ & && \cdots, && \\ & \boldsymbol{\Psi}_{xx,\mathrm{STM}}(\ell, \omega_{N_{\mathrm{rate}}-1}, \Omega_0), && \cdots, && \boldsymbol{\Psi}_{xx,\mathrm{STM}}(\ell, \omega_{N_{\mathrm{rate}}-1}, \Omega_{N_{\mathrm{scale}}-1}) \Big]^T\end{aligned} \tag{3.67}$$

with $\boldsymbol{\Psi}_{xx,\mathrm{STM}}$ subsuming the stacked frequency regions, by far exceeds the size of all other features discussed in this thesis: Mesgarani et al. [2006], for example, evaluated $N_{\mathrm{scale}} = 5$ scales and $N_{\mathrm{rate}} = 12$ rates for 128 bands resulting in a 7680 element feature vector.

Different strategies were proposed to compress the feature vectors while keeping the relevant information. Mesgarani et al. [2006] applied a multidimensional principal component analysis (PCA) to decompose the original feature vector into uncorrelated and hence more

informative components. Subsequently, the decision was taken based on boundaries gained by a support vector machine (SVM). A rather basic but efficient solution was considered in [Hsu et al., 2013]: instead of calculating all the different scales and rates at runtime, they focused on the most prominent pair as discovered in preliminary analyses.

For analyses in this thesis, again a neural network is trained to fuse the feature vector to a scalar decision variable.

3.3.3 Alternating structure of voiced and unvoiced phones

The long-term properties considered so far in this thesis became more and more specific for human speech. While the first features based on non-stationarity treat every signal component as speech that sticks out of the stationary background, the modulation-based features expect a particular temporal structure. By making the features more specific, the robustness against stationary and non-stationary noise has been improved.

When having a look at the spectrogram of speech signals, another even more specific characteristic appears promising for the detection of speech: typically, voiced and unvoiced speech portions do not occur at the same time. Instead, an alternating pattern of excitations of high and low frequencies is observable as illustrated in Figure 3.25. In this section, a new feature introduced in [Graf et al., 2016b] is summarized that explicitly exploits this characteristic.

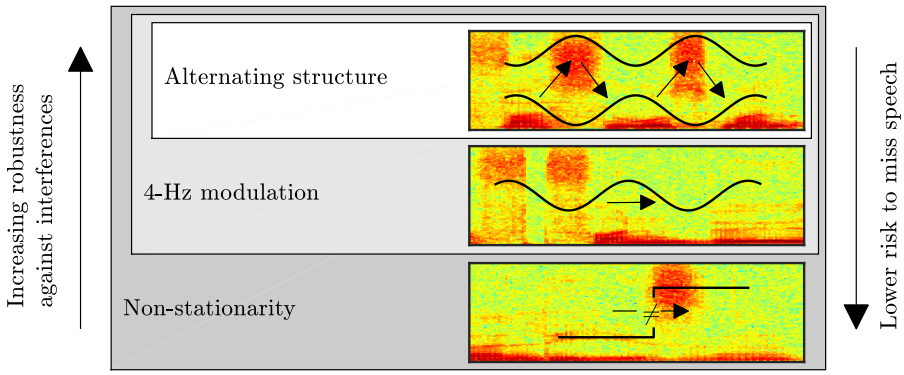


Figure 3.25: Different levels of specialization: non-stationarity indicates speech for *any variation* of the power spectrum over time, modulation further expects a *certain frequency* of fluctuations, whereas the new MPD feature is even more focused on detecting an *interleaved excitation* of low and high frequencies [Graf et al., 2016b].

This feature targets on an interleaved excitation of low and high frequencies corresponding to alternations between voiced and unvoiced speech portions. Interferences typically do not exhibit this structure, hence, the feature is expected to be particularly robust against false alarms.

To capture the presence of voiced and unvoiced phones, spectral magnitudes are averaged over frequency

$$B(w,\ell) = \frac{\sum_{k=k_{\min}(w)}^{k_{\max}(w)} |X(k,\ell)|}{k_{\max}(w) - k_{\min}(w) + 1} \tag{3.68}$$

for frequencies that are dominated by the respective phones. In this thesis, two very broad frequency bands $w \in \{1,2\}$ for voiced and unvoiced speech are chosen as $[200\,\text{Hz}, 2\,\text{kHz}]$ and $[4.5\,\text{kHz}, 8\,\text{kHz}]$ respectively. This aggregation of frequencies makes the feature suitable also for applications that have to cope with a low spectral resolution such as ICC systems. As an alternative to the purely power-based averaging of spectral magnitudes, other features could be taken as basis that explicitly represent the voicing properties of speech: voiced and unvoiced speech portions can be detected by features such as the ones summarized in Section 3.2.3, or the low-complex feature for pitch detection in Section 3.2.4. Temporal alternations of the results can subsequently be assessed by a feature analogous to the approach discussed here.

Only the non-stationary components should be considered by the feature. Removal of the stationary components can be realized by high-pass filters

$$B_{\text{hp}}(w,\ell) = (1+\beta_1) \cdot (B(w,\ell) - B(w,\ell-1))\,/2 + \beta_1 \cdot B_{\text{hp}}(w,\ell-1) \tag{3.69}$$

with a smoothing constant $\beta_1 \hat{=} -48\,\text{dB/s}$ that preserve the modulations in both bands.

As discussed in the previous section, speech typically fluctuates with a modulation frequency of about 4 Hz. This temporal characteristic is adopted here by employing IIR filters

$$B_{\text{mod}}(w,\ell) = (1-\beta_2) \cdot B_{\text{hp}}(w,\ell) + \beta_2 \cdot B_{\text{mod}}(w,\ell-1) \cdot e^{2\pi j \Omega_{\text{mod}}} \tag{3.70}$$

that emphasize frequencies close to $\Omega_{\text{mod}} \hat{=} 4\,\text{Hz}$. An exponentially decaying window is realized using a smoothing constant $\beta_2 \hat{=} -24\,\text{dB/s}$. The output signals of the filters are complex-valued and can be divided into magnitude and phase: the magnitude captures information on the degree of modulation in a similar manner as the modulation features discussed so far. Using the phase, concentration of power can now further be temporally localized and compared between both bands for a detection of interleaved structures.

The magnitude of $B_{\text{mod}}(w,\ell)$ still depends on the scaling of the input signal. By relating the modulated components to the variance

$$B^2_{\text{norm}}(w,\ell) = (1-\beta_2) \cdot B^2_{\text{hp}}(w,\ell) + \beta_2 \cdot B^2_{\text{norm}}(w,\ell-1) \tag{3.71}$$

of all components, smoothed with the same constant β_2 as before, a normalized version

$$\tilde{B}(w,\ell) = \frac{B_{\text{mod}}(w,\ell)}{\sqrt{B^2_{\text{norm}}(w,\ell)}} \tag{3.72}$$

can be derived that is independent from the original scaling.

Averaging the magnitudes of both bands yields a modulation feature

$$MOD(\ell) = \frac{|\tilde{B}(1,\ell)| + |\tilde{B}(2,\ell)|}{2} \tag{3.73}$$

similar to f_{M4Hz} in Eq. (3.62).

Combining the magnitudes of both bands as well as the phase difference between the bands results in the new feature

$$MPD(\ell) = -|\tilde{B}(1,\ell)| \cdot |\tilde{B}(2,\ell)| \cdot \cos\left(\angle \tilde{B}(1,\ell)\tilde{B}^*(2,\ell)\right) \tag{3.74}$$

that indicates speech when the modulation is high in both bands and, additionally, occurrences of high values are temporally separated between both bands. In this case, $\cos\left(\angle \tilde{B}(1,\ell)\tilde{B}^*(2,\ell)\right)$ assumes values close to -1 so that the feature is maximized. An equivalent expression of the feature is given by

$$MPD(\ell) = -\left(\text{Re}\{\tilde{B}(1,\ell)\} \cdot \text{Re}\{\tilde{B}(2,\ell)\} + \text{Im}\{\tilde{B}(1,\ell)\} \cdot \text{Im}\{\tilde{B}(2,\ell)\}\right) \tag{3.75}$$

that can be evaluated very efficiently.

For noise, no such pattern is expected: stationary noise does not exhibit a modulation at 4 Hz and hence generally doesn't trigger modulation-based features. But even non-stationary noise components can be rejected by the new feature due to the very specific criterion. This high robustness against different types of noise is the main benefit of the feature.

Unfortunately, the quite specific pattern is also not permanently observable during presence of speech. On the one hand, there may be sequences with only voiced speech, and on the other hand, short speech pauses may interrupt the pattern. Temporally extending the decision, e.g., by introducing a hangover mechanism, is therefore advisable. A hangover can be realized by holding preliminary detection results or equivalently by taking the maximum value

$$f_{\text{MPD}}(\ell) = \max_{0 \le \tilde{\ell} < L_{\text{MPD}}} MPD\left(\ell - \tilde{\ell}\right) \tag{3.76}$$

of the feature over some previous frames. This mechanism can prevent detection dropouts for capturing longer speech intervals as it will be discussed again in Section 5.2 for the purpose of speech detection in an ICC system.

3.4 Speech detectors

Many publications distinguish between feature extraction and classification or detection [Ramírez et al., 2007]. However, the separation between both categories is not clearly specified and to some extent appears to be fuzzy. In a frequently cited approach by Sohn et al. [1999], for instance, the DFT bins are modeled by Gaussian-distributed random variables for speech and noise. Using a decision-directed method [Ephraim and Malah,

1984], the model parameters, i.e., variances for speech and noise, are estimated. The final decision relies on a likelihood ratio between the models for speech and noise that further incorporates a hidden Markov model (HMM)-based smoothing. For this example, the likelihood ratio could be interpreted as a sophisticatedly normalized feature with a basic threshold-based decision. Just as well, the complete approach could be attributed to the detector with the DFT-bins as a basic feature.

The features as discussed in this work target on characteristic properties of speech. The raw input signal is processed in such a way that a certain property of speech gets emphasized. Conclusions on presence or absence of speech in the noisy audio data can be drawn based on the resulting scalar (f) or vectorial ($\boldsymbol{f}$) value. Taking a binary decision based on one or multiple features is the detector's task. The detector is hence an integral part of any VAD that significantly contributes to the system's performance.

There are numerous ways for designing and optimizing a detector: some detectors are designed heuristically, e.g., based on a meaningful combination of different decisions as in [Marzinzik and Kollmeier, 2002]. Other approaches rely on machine learning techniques that optimize the detector with the help of labeled training data. Multiple weak decisions can be combined to a more robust detector using AdaBoost optimization [Kwon and Lee, 2003, Usukura and Mitsuhashi, 2008]. Some detectors even consider the temporal evolution of feature values by incorporating additional context information, e.g., by means of HMMs [Veisi and Sameti, 2012] or partially observable Markov decision processes [Park et al., 2009].

Artificial neural networks turned into the most favored machine learning approach during the last decade [Krohn et al., 2019]. Even though the basic principles were already investigated in the 1940s [Bishop, 2007], large-scale practical applications became possible in the first place thanks to the tremendous advances in computer power. Nowadays, training deep networks comprising several layers and even recurrent neurons is feasible [Goodfellow et al., 2016].

In this work, an artificial neural network as shown in Figure 3.26 is considered for speech detection. The input feature vector $\boldsymbol{f} = [f_0, f_1, \cdots, f_{I-1}]^T$ is processed with a feed-forward structure. In a single hidden layer comprising $H = 20$ neurons

$$h_h = \varphi_{\text{act,hidden}}(z_h) \text{ with } z_h = b_h + \sum_{i=0}^{I-1} w_{h,i} \cdot f_i, \tag{3.77}$$

the vector elements f_i are weighted with $w_{h,i}$ and accumulated including an additional term b_h that implements a bias. For the subsequent activation function $\varphi_{\text{act,hidden}}$, a hyperbolic tangent function tanh is chosen. These intermediate results are finally merged by neurons in the output layer

$$o_o = \varphi_{\text{act,out}}(z_o) \text{ with } z_o = b_o + \sum_{h=0}^{H-1} w_{o,h} \cdot h_h \tag{3.78}$$

that calculate a weighted sum z_o of the hidden layer's results. Using a soft-max activation

function [Bishop, 2007]

$$\varphi_{\text{act,out}}(z_o) = \frac{\exp z_o}{\sum_{\tilde{o}=0}^{O-1} \exp z_{\tilde{o}}}, \tag{3.79}$$

two probability-like output values between zero and one are calculated that indicate presence o_1 or absence $o_0 = 1 - o_1$ of speech.

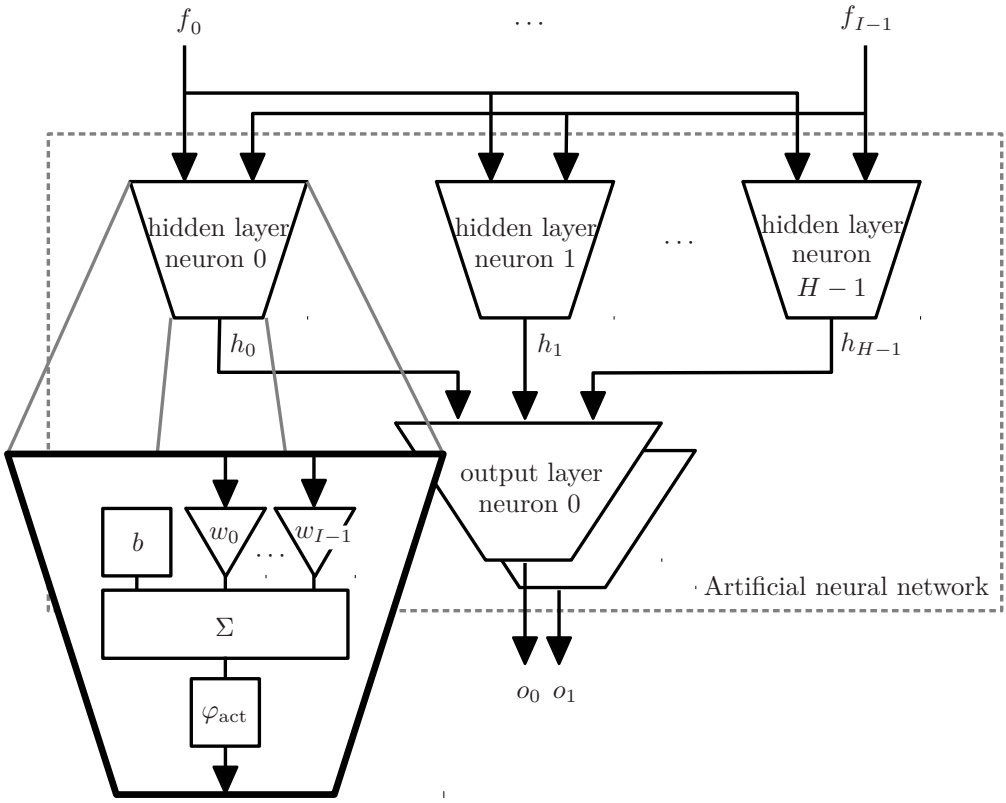


Figure 3.26: Artificial neural network used for feature fusing: I-dimensional input feature vectors are merged by a feed-forward structure with a single hidden layer comprising H neurons. The network's output indicates presence $o_1 = 1 - o_0$ or absence o_0 of speech. This network is applied to any multi-dimensional feature in this thesis for deriving a scalar decision variable.

This basic network structure was chosen in order to ensure a fair comparison of the features without overemphasizing the network's influence: a more elaborated architecture certainly could improve the overall detection results, however, it would shift the attention away from the features towards the detector. Later in the analyses, the common network structure will be applied to any multi-dimensional feature for fusing the feature vectors to scalar decision variables. The input data is based on normalized feature vectors: mean and variance of each element are forced to zero and one respectively. Both statistics can be calculated in advance based on the complete database.

The network parameters – weights w and biases b – were trained using a standard error backpropagation algorithm [Bishop, 2007] based on a mini-batch gradient decent with a fixed step size 10^{-4}: the training data was subdivided into mini batches each covering 900 frames of the feature data along with a corresponding reference for VAD. The frames were randomly selected under the constraint that each mini batch contains an equal number of speech and non-speech frames. This way, the network was forced to focus on the contrast between speech and noise frames irrespective of the a priori probability of presence of speech in the original database. Otherwise, with only a little percentage of frames capturing speech or noise, the network's output tended to be biased towards this fixed a priori probability instead of predicting frame-based results corresponding to the current input data [Lawrence et al., 1998].

The neutral network's output can be interpreted as a scalar representation derived from the original multi-dimensional features. For the analyses in this thesis, a threshold η according to Eq. (2.17) is finally applied either directly to the scalar features or to the scalar desicion variable provided by the network.

Standard approaches

VAD has been an intensively investigated field of research over several decades. During this time, many approaches have been presented for various applications. Apart from academic research, standards were defined for commercial applications, e.g., by the International Telecommunication Union (ITU) and the European Telecommunications Standards Institute (ETSI). Their speech coding standards include precisely specified algorithms for VAD that are accompanied by corresponding reference implementations.

In the following, three standardized techniques will be discussed. By means of these approaches, the typical structure of VAD approaches, feature extraction, detection, and post-processing are exemplified.

An early VAD standard was defined in ITU-T G.729 Annex B [Benyassine et al., 1997, ITU, 1996] for the purpose of speech transmission. During absence of speech, the speech codec reduces the transmission rate dynamically. Comfort noise is inserted that can be modeled using less parameters compared to speech. The underlying VAD approach is a popular reference that has been taken as a baseline in numerous research papers.

The algorithm illustrated in Figure 3.27 generates a VAD result every 10 ms relying on the contributions of four different features: the full-band energy, a low-band energy (addressing frequencies in the range 0-1 kHz [Ramírez et al., 2004a]), the ZCR, as well as a measure for spectral distortions based on 10 LSFs. For each feature, the difference between the instantaneous value and a running average is calculated. In order to capture the background noise characteristics, the average values are updated in case that the VAD finally decides for speech pauses. In the detector, a preliminary decision is made based on 14 boundary conditions that are applied to the four-dimensional feature space. A hangover mechanism is employed for the final VAD result. This post-processing is realized by means of four stages of decision smoothing.

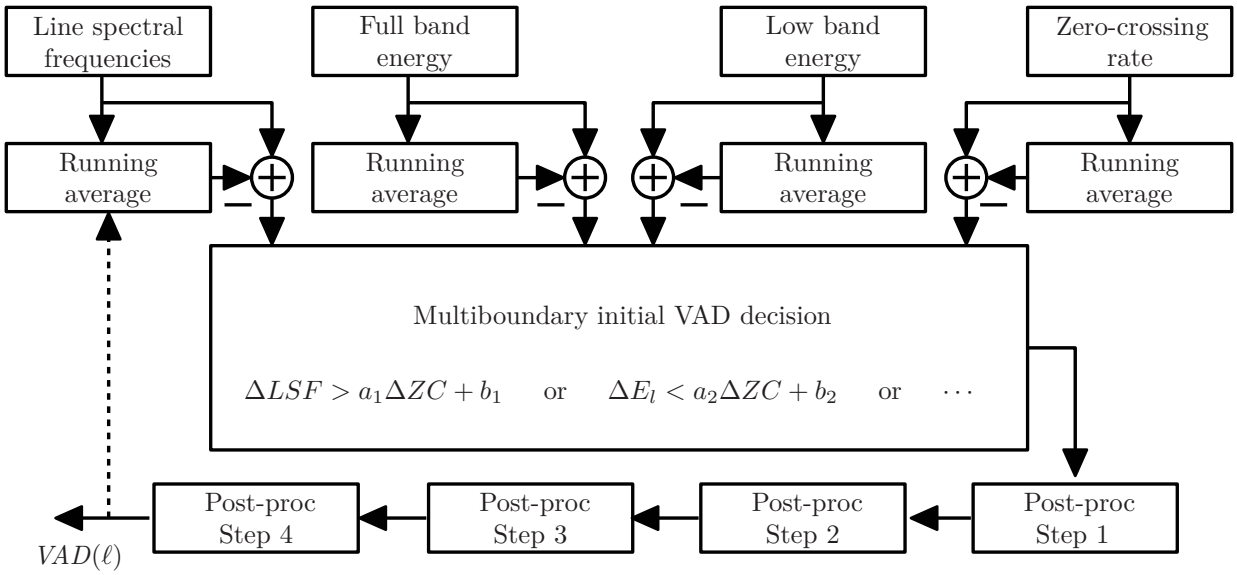


Figure 3.27: Processing steps for VAD as defined in ITU-T G.729 Annex B: four different features are extracted along with their running averages. Based on these features, an initial decision is taken that gets smoothed to the final VAD result in four post-processing steps realizing a hangover.

The original approach was designed for narrowband telephony applications with a sample rate of 8 kHz. An extension was introduced with ITU-T G.729.1 Annex F [ITU, 2012] that supports wideband signals with a sample rate of 16 kHz. To cope with different preferences, this approach provides three modes: a bandwidth saving operating point, a quality-preferred operating point, as well as a balanced operating point.

Two alternatives for the VAD to be used with adaptive multi-rate (AMR) speech traffic channels were defined in [ETSI, 1998]. Both generate a detection result for each 20 ms frame.

Option 1 as illustrated in Figure 3.28 puts a strong emphasis on harmonic components. Voiced speech and other periodic signals are assessed by means of a pitch feature. The feature reuses information calculated during a pitch analysis in the speech encoder. In addition to speech, information tones and other strongly periodic signal components should be detected. Replacing these components in the encoded signal with comfort noise may sound annoying and should hence be avoided. Dedicated features for tone detection and for the detection of correlated components are taken into account. Signal levels for nine different frequency bands are finally calculated. Based on these levels, an SNR and the level of background noise are estimated. A preliminary decision is made by comparing the SNR with a threshold that is chosen dynamically depending on the background noise level. For the final VAD result, a hangover is added that temporally smooths the decision based on statistics over the different features.

Option 2 primarily relies on energy-based features as illustrated in Figure 3.29. Using a DFT, the signal is converted into the frequency domain. Multiple frequencies are aggre-

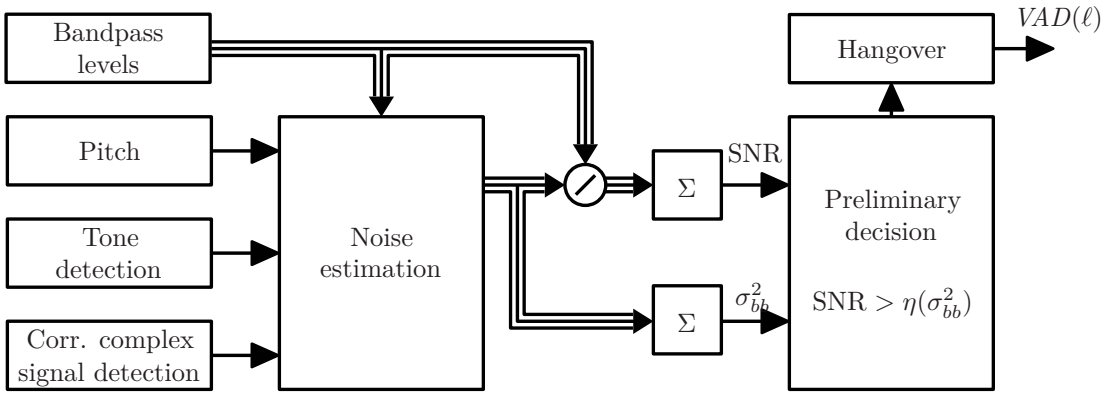


Figure 3.28: The VAD option 1 according to ETSI AMR standard utilizes four features for noise and SNR estimation. The SNR is compared with a noise-dependent threshold for a preliminary decision that successively gets smoothed by a hangover.

gated to wider channels for which the energy as well as the SNR are calculated. Based on the channel SNR, a voice metric is determined that serves as the VAD's primary feature. The estimation of background noise relies on additional features: a pitch feature calculated from the speech encoder's long-term prediction gain and a sinewave flag. The detection threshold applied to the voice metric is chosen dynamically depending on a long-term peak SNR. Similar thresholds are applied to a burst count and to a hangover count that both smooth the detection results in order to capture also the end of speech utterances.

The original AMR standard again was defined for narrowband applications. A wideband extension for 16 kHz was introduced with [3GPP, 2000] that makes use of similar features compared to AMR Option 2.

Another standardized approach was introduced in [ETSI, 2007] for the purpose of distributed speech recognition. The advanced front-end (AFE) feature extraction algorithm incorporates a VAD based on three features illustrated in Figure 3.30: the full-band energy is taken into account primarily for capturing unvoiced speech portions. The lower part of the spectrum containing voiced speech is addressed by the low-band energy as well as the spectral variance within the lower half of the spectrum. For each feature, the second derivative – also called acceleration – is estimated based on the ratio of the instantaneous and a slowly smoothed value. When any of the three acceleration features exceeds the corresponding threshold, a preliminary flag indicating presence of speech is buffered by the detector. Statistics of consecutive speech flags finally control a dynamic hangover mechanism for the resulting VAD decision.

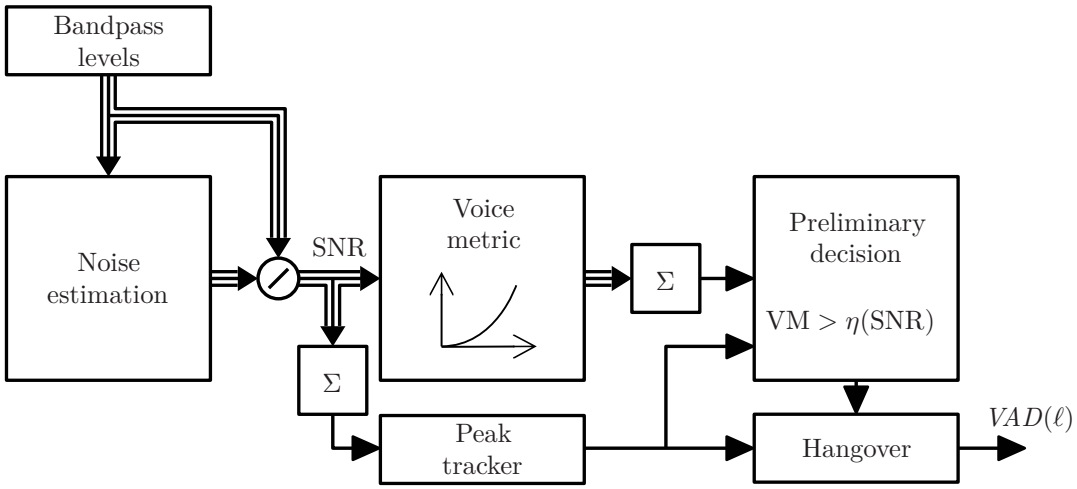


Figure 3.29: Option 2 in ETSI AMR primarily relies on the power extracted for different frequency regions. The SNR is calculated and mapped to a voice metric. This metric is taken as decision variable for a preliminary detection. Again, a hangover stage is applied for the final result. Both the preliminary decision as well as the hangover are parametrized dynamically depending on the SNR's peak.

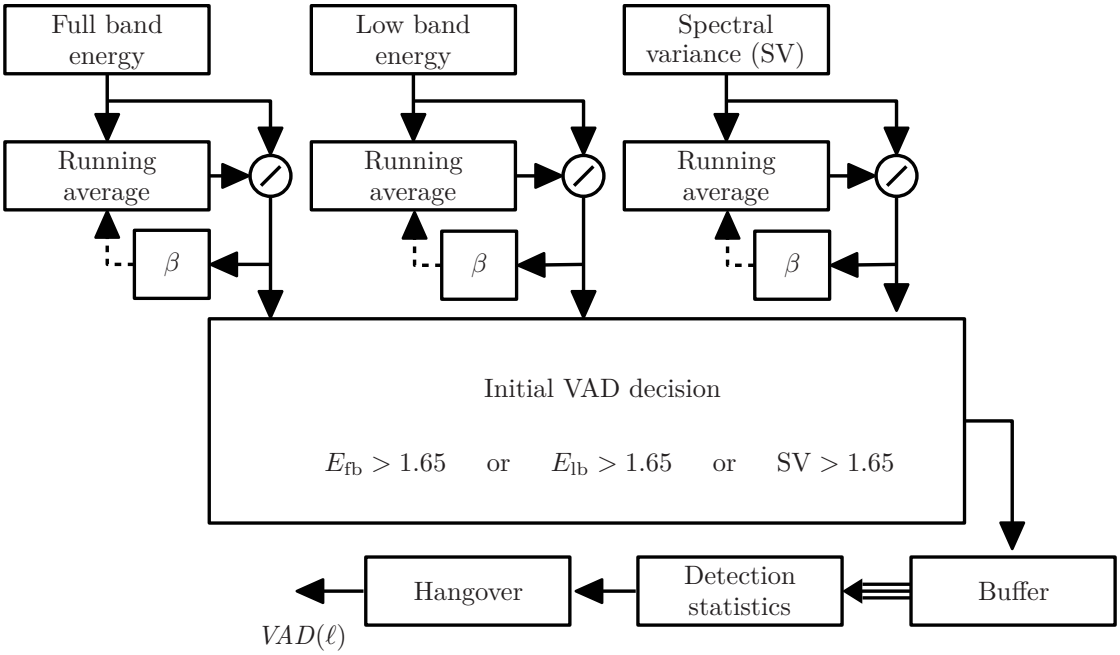


Figure 3.30: Standardized VAD according to ETSI AFE: three features are extracted along with their running averages. The ratios of instantaneous values and their respective averages are taken as decision variables that control the smoothing parameters β as well as the initial VAD decision. In a post-processing, involving statistics over consecutive frames as well as a hangover, the final result is smoothed.

Chapter 4

Evaluation of Speech Detection in Noisy Environments

Many different applications may benefit from accurate VAD results. The VAD is integrated in these applications in conjunction with several other components and controls, e.g., the noise power estimation. The application's performance therefore depends not only on the VAD, but on the interaction of multiple components. In the end, the overall application performance should be improved.

Evaluations of the complete application may be quite complex and time consuming especially when subjective tests with human subjects have to be conducted. It is therefore desirable to initially investigate the behavior of a VAD in an isolated fashion independent from the influence of other components. Even though improvement of the complete application is targeted, evaluations of the pure VAD can help to get an impression of the performance. These evaluations should take into account the specific constraints imposed by the application.

As illustrated in Figure 4.1, VAD can be evaluated by applying the speech detector to known test signals. These test signals must include both speech and noise and should cover representative scenarios the particular application is typically faced with. The signal can be recorded in a noisy environment or can be artificially mixed based on databases containing clean speech and pure noise. Irrespective of the signal generation, it is required for the evaluation that the audio signal is accompanied by a temporally aligned reference that indicates the presence or absence of speech. This reference is taken as ground truth for the evaluation. The detector's performance can be analyzed by comparing the detection results $VAD(\ell)$ with the reference $VAD_{\text{ref}}(\ell)$. Deviations between the observed detections and the expected reference are treated as errors that can be assessed by means of different measures.

Depending on the application, different aspects of the VAD results are decisive. Some applications require a high robustness against noise whereas for other applications a fast reaction to speech onsets is more important. To find an appropriate solution for a certain application, multiple measures that reflect the different errors have to be considered. Measuring these errors helps rating the usability of different VAD approaches for an application

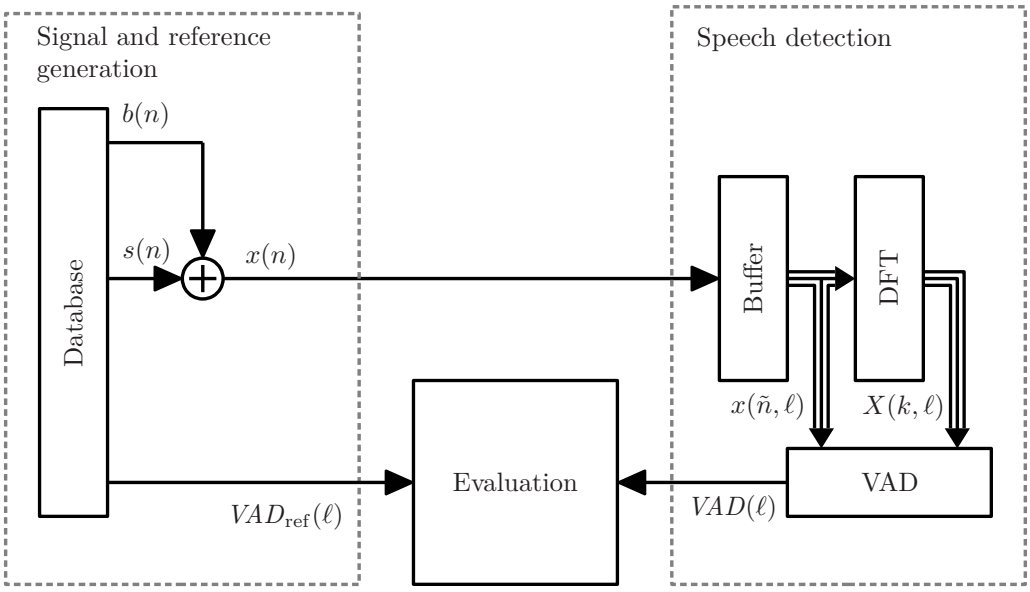


Figure 4.1: Evaluation of VAD results: audio signals as well as a reference for presence of speech are generated based on a database. In the evaluation, the detection results $VAD(\ell)$ are compared to the reference $VAD_{\text{ref}}(\ell)$. Deviations can be rated using different measures.

and to find a tradeoff that matches the application's requirements.

In the following, the generation of test signals and the reference is summarized as well as measures to quantify different types of errors. A comprehensive evaluation of many features is presented at the end of this chapter. This initial analysis covers a wide range of diverse noise conditions that likely exceeds the scope of a single application. Application-specific evaluations are presented in the last chapter of this thesis.

4.1 Signal and reference generation

The choice of the test data depends on many criteria: most important, it should reflect realistic use cases for the final application. The generation of a reference for VAD should further be simple and reproducible. Publicly available data is preferred since it allows for more comprehensible analyses.

The use cases in this thesis focus on applications in mobile and automotive contexts. While in a car the background noise typically is stationary, highly fluctuating babble noise can occur for mobile devices in crowded environments. A comprehensive database is needed to capture the wide range of noise conditions in both situations. Furthermore, the detection of desired speech should not focus on a certain speaker but should perform well for different speakers. To test for a proper generalization, a variety of speakers has to be considered in the analyses.

The subsequent analyses partly rely on recordings that were conducted with a mobile device to capture speech utterances in realistic environments. These recordings perfectly reproduce some real conditions, however, they are limited to only few situations and speakers. Furthermore, the reference for presence of speech is not automatically included and cannot be derived easily. Since only the noisy data is accessible, the reference has to be created manually by hand-labeling the data. These recordings were therefore only used in this work for the final evaluation of noise suppression in a mobile application.

For all other evaluations, artificially mixed signals were employed that are close to reality as subjectively verified by the author. Artificially mixed signals can cover a much wider range of situations and speakers. Since the VAD reference can be generated automatically, these signals are particularly suitable for evaluations. Speech and noise components for the artificial mix can be collected from publicly available databases that will be discussed in the following.

4.1.1 Speech and noise databases

A variety of audio databases is publicly available for signal enhancement research and for algorithm improvement. Some of them are particularly suitable for VAD analyses as they either provide a temporally aligned reference for speech or consider noise conditions of interest. In the following, some databases are briefly summarized that either contain clean speech, pure noise, or mixtures of both.

Speech databases may differ in their extent, the number of languages and speakers, as well as the recording conditions and the quality of the audio signals. To mention just a few important examples:

- TIMIT [Garofolo et al., 1993] is a well-established database that was adopted by a majority of researchers. The clean speech recordings were conducted with 630 native US English speakers (438 male and 192 female) uttering 10 sentences each. The audio data with 16 kHz sample rate is complemented by a time-aligned phonetic transcription that was automatically generated and reviewed by experts in phonetics [Seneff and Zue, 1988]. This transcription can easily be converted into a reference for VAD. Most analyses in this work rely on speech data extracted from the TIMIT corpus thanks to the high audio quality and the precise transcription.

- Keele [Plante et al., 1995] is a speech database dedicated to pitch-related analyses. Audio signals of clean speech were recorded for 10 speakers. Synchronously, vibrations of their vocal folds were recorded using a laryngograph. Based on these signals, the voicing properties of speech can be analyzed in depth.

- SPEECHDAT-CAR [Moreno et al., 2000] and SPEECON [Iskra et al., 2002] both are large-scale databases of noisy speech. Industrial consortia conducted systematic recordings for multiple regions and languages following common schemes. The databases cover several conditions in automotive environments and for consumer devices, respectively. Even though these environments are also investigated in this

work, the databases appeared to be less suitable for the analyses of this thesis: as only the noisy speech is available, generating a reference for VAD is difficult.

Also several publicly available databases covering noise data exist that may be considered for analyses:

- NOISEX-92 [Varga and Steeneken, 1993] is a well-known database even though it covers only a rather small set of noise scenarios. Assorted by the NATO Research Study Group, the adverse noise conditions primarily address military scenarios. Overall, eight recordings capture highly non-stationary machine-gun noise, Lynx-helicopter cockpit noise, F16 fast-jet cockpit noise, car noise, factory noise, office noise, babble noise, as well as an artificially generated colored noise. Most analyses in this work make use of NOISEX-92 data in addition to other databases that are dedicated to other domains such automotive environments.

- QUT-NOISE [Dean et al., 2010] is a corpus of noise data that was explicitly designed for the purpose of VAD analyses. For ten different locations, recordings are available that each cover at least 0.5 h of pure noise. The recording environments are grouped into five categories: cafe, home, street, car, and reverberant. Together with the noise data, scripts are provided for mixing the noise with speech data taken from the TIMIT database and generating a reference for VAD based on a word-level transcription. In this work, a comprehensive analysis of features relies on the noise recordings in QUT-NOISE. In contrast, the scripts were not used since a more fine-grained reference based on the phonetic transcription was preferred.

- UTD-CAR-NOISE [Krishnamurthy and Hansen, 2013] provides an extensive set of noise recordings in automotive environments. Typical driving conditions were captured on a common route for 20 cars, 5 trucks as well as 5 SUVs. In addition to the driving noise, other sounds related to the car were collected such as the indicator's clicking, horn, and noise caused by opening and closing the doors. Here, the noise data is employed for evaluations of the new features that are dedicated to ICC applications.

4.1.2 Reference generation

The analyses in this work rely on a comparison between the detection results $VAD(\ell)$ and a time-aligned reference $VAD_{\mathrm{ref}}(\ell)$. Depending on the speech database, different methods for generating the reference are applicable. Sometimes, the database already provides a transcription that can be adopted for VAD analyses. Otherwise, the reference has to be generated manually or using an automatic labeling approach. In any case, a small uncertainty regarding the exact position of speech boundaries has to be expected: depending on the proficiency of human annotators or the settings of an algorithm for reference, deviations in the range of up to 0.2 s were reported in [Kraljevski et al., 2015].

Table 4.1: TIMIT symbols mapped onto broader speech classes

TIMIT symbol	Speech class
h#, pau	speech pause
epi, bcl, dcl, gcl, pcl, tck, kcl, tcl	word pause
hh,s,sh,f,th,ch,p,t,k,q	unvoiced } speech
all other	voiced } speech

For TIMIT, an accurate reference can easily be generated based on the transcription provided with the database. Each audio file is accompanied by a text file containing phonetic symbols that are temporally aligned with the signal [Seneff and Zue, 1988]. The very granular phonetic symbols can be summarized to broader classes as depicted in Table 4.1 [Graf et al., 2014]. These speech classes are relevant for VAD analyses: during speech pauses, no speech should be detected $VAD_{\text{ref}}(\ell) = 0$. Word pauses inside a continuous utterance have a meaningful role in spoken language and could hence be associated with speech. From a signal processing perspective, they could just as well be treated as pauses due to the missing excitation during these sections. As neither the one nor the other interpretation is perfectly true, here these intervals are simply excluded from the evaluations: $VAD_{\text{ref}}(\ell) = 0.5$. Distinguishing voiced and unvoiced phones is beneficial for some analyses. In this case, only one of both classes is taken into account whereas the other one is excluded from evaluations. This way, detection rates can be measured for voiced and unvoiced phones separately. For most analyses, both classes are jointly considered as speech: $VAD_{\text{ref}}(\ell) = 1$.

For databases containing clean speech such as Keele, a reference can automatically be generated. Applying a simple, e.g., energy-based VAD to the clean signal is sufficient for finding speech intervals. A manual review of the generated reference helps for excluding outliers, e.g., caused by interferences in the signal. After mixing the clean speech and noise, the VAD can be compared to the results based on the clean data. Algorithms are supposed to reproduce the clean VAD results as accurately as possible based on the noisy data to be regarded as robust.

When only noisy speech samples are available, automatically generating a reference appears to be difficult. Automatic labeling approaches would be confronted with exactly the same data used during the analyses without being at an advantage over the algorithms that are to be evaluated finally. Manual labeling is a feasible but tedious option for generating a reference in this case. For this thesis, speech intervals of some selected recordings were marked using a tool shown in Figure 4.2. The visualized time-domain signal and a spectrogram as well as playback of the audio signal helped for finding the relevant sections. Due to the high complexity of this approach, only some real recordings taken with a mobile device in a noisy environment were hand-labeled.

Most analyses in contrast rely on artificially mixed signals based on clean speech and

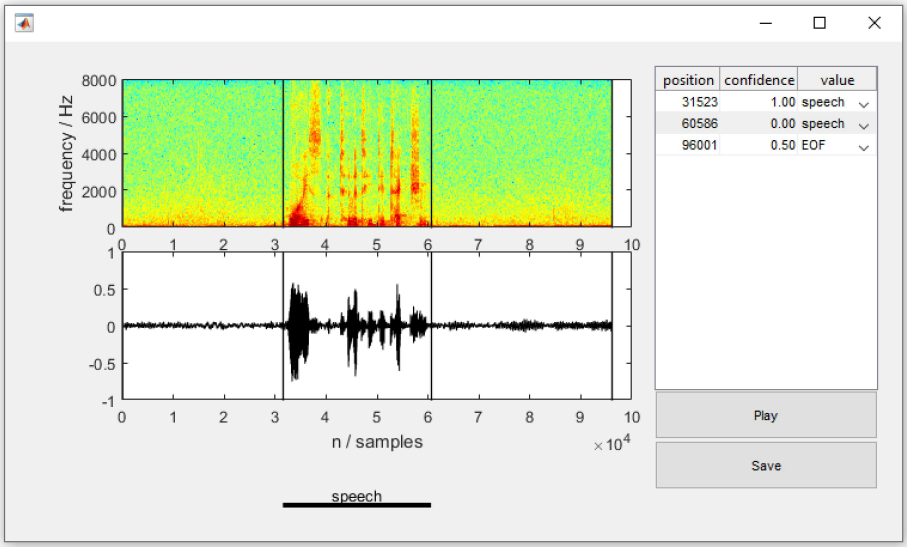


Figure 4.2: Tool used for manual labeling of audio data: spectrogram and time-domain signal as well as playback of the audio assisted during the labeling process. As this approach is yet time-consuming, only some selected signals were labeled this way.

noise. The noisy speech data

$$x(n) = s(n) + b(n) \tag{4.1}$$

is composed using the noise recording $b(n)$ and the speech signal $s(n)$. For realistic data that reflect also the room acoustics, a reverberant speech signal

$$s(n) = \sigma_{\mathrm{s}} \cdot \sum_{\tilde{n}=0}^{N_{\mathrm{RIR}}-1} s_{\mathrm{CT}}(n-\tilde{n}) \cdot h_{\mathrm{RIR}}(\tilde{n}) \tag{4.2}$$

can be generated by filtering close-talk speech recordings $s_{\mathrm{CT}}(n)$ with measured room impulse responses $h_{\mathrm{RIR}}(\tilde{n})$ of length N_{RIR}. This simulation of specific room characteristics will later be applied for some reverberant scenarios where impulse responses are available that match the noise recordings.

During presence of noise, speakers typically raise their voices according to the Lombard reflex [Junqua, 1996]. The main emphasis is shifted towards higher frequencies that set apart more prominently from the background noise. For the artificially mixed data, this effect is not considered but only the loudness is adjusted. A factor σ_{s} is applied to the speech signal in order to control the speech level in the mixed data. Different values of the SNR can be simulated by choosing the factor accordingly.

4.1.3 Perceived SNR and objective measures

Specifying an SNR, the relative contributions of speech and noise in a noisy signal are quantified. When only the noisy mix is available, estimating the SNR is particularly

difficult since speech and noise components have to be separated first, usually relying on a VAD [Vondrášek and Pollák, 2005]. But even for the case that both components can be assessed separately, several definitions of SNR may be found in literature: different values may result depending on how the spectro-temporal distributions of speech and noise are considered.

In a simple implementation

$$\sigma^2_{\mathrm{SNR}} = \frac{\sum_n s^2(n)}{\sum_n b^2(n)}, \tag{4.3}$$

taken as standard reference in the following, the powers of speech and noise are calculated irrespective of their spectral or temporal distributions. According to this definition, an SNR of 0 dB is observed when the energies of both components are equal. The signal duration cancels down since both numerator and denominator are aggregated over the full signal period. Obviously, this approach strongly depends on the percentage of speech in the signal. Persistent speech activity is rewarded with a high SNR whereas the measure drops with an increasing amount of speech pauses. Restricting the estimation of speech power to time intervals that contain speech activity makes the measure independent from the quantity of speech samples. An active speech level can be measured according to ITU-T Recommendation P.56 [ITU, 1993]. By means of a power-based VAD, frames are selected that are taken into account for the calculation of speech level.

When mixing signals for a specific target SNR value, comparability between different scenarios may be desirable. The human perception can be taken as a reference: different signals mixed to the same target SNR should result in the same hearing impression for human listeners irrespective of the spectral distribution of noise. The simple definition according to Eq. (4.3) does not meet this requirement as discussed in [Graf et al., 2015b]. For noises with different spectral distributions, shifted scales of the SNR are expected as illustrated in Figure 4.3.

In order to verify this assumption, subjective listening tests have been conducted for this thesis. For the experiment [Graf et al., 2015b], artificially mixed signals were presented pairwise to the participants. Each pair was based on a common speech component that was mixed with two noises with different spectral distributions. While the first example's SNR was fixed, the test subjects were asked to adjust the noise level of the second example such that the speech set apart from the noise similarly in both examples[1]. Based on the test results, a mismatch between the standard definition of SNR according to Eq. (4.3) and the human perception can be determined.

In total, 20 participants (7 female, 13 male) attended the experiment. They listened to the signals that were played back via two loudspeakers (including a subwoofer dedicated to frequencies below 120 Hz) in a semi-anechoic chamber. All signals were presented with a high sample rate of 44.1 kHz and the loudspeakers' transfer functions were compensated by means of equalizer filters. The speech samples were based on the phonetically balanced

[1]The exact German instruction was: "Stellen Sie [$\cdots$] das Geräusch des rechten Hörbeispiels so ein, dass sich die Sprache ähnlich vom Geräusch abhebt wie im Referenzbeispiel"

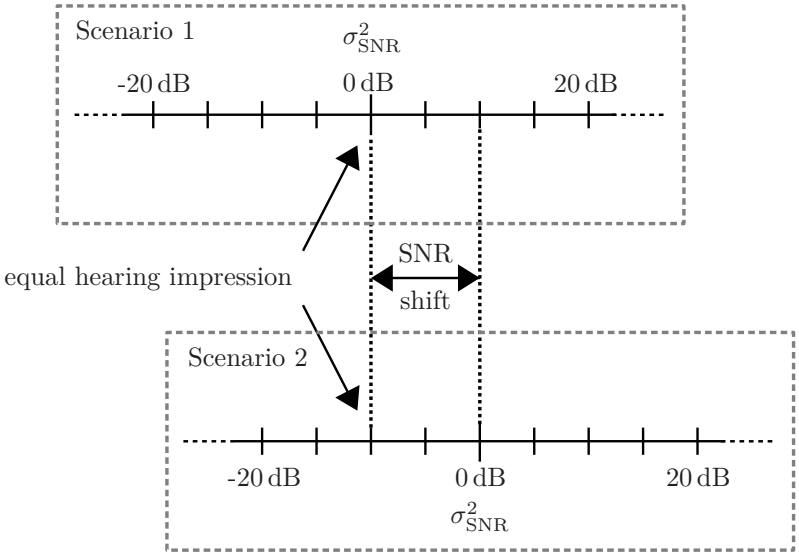


Figure 4.3: Illustration of the expected shift of scales: the measured standard SNR is expected to differ between two noise scenarios that possess different spectral shapes even though human listeners perceive them as being equal.

text "Nordwind und Sonne" [IPA, 1999] read by a female as well as a male speaker. Six noise signals with diverse spectral distributions were considered:

- white noise where all frequencies are equally excited,
- three bandpass filtered noises in the ranges 1-3 kHz, 3-9 kHz, and 9-22 kHz all addressing the upper part of the spectrum which partially overlaps with unvoiced fricatives,
- automotive noise recorded in a car at a speed of 100 kph that primarily affects the low frequencies below 150 Hz,
- and babble noise recorded in a cafe. This signal exhibits a spectral distribution that strongly overlaps with the desired speech's distribution in the frequency range between 200 Hz and 2 kHz. Also some non-stationary components are included whereas all other noise examples considered here are purely stationary.

The results are summarized in Figure 4.4. Obviously, there is a strong mismatch between the simple implementation of SNR and human perception for most scenarios. On average, the SNR was adjusted by -16.4 dB for automotive noise until the ratio was perceived as similar compared to a white noise reference with 0 dB SNR. The same tendency can be observed for the bandpass scenarios that were all adjusted by about -10 dB compared to white noise. For these scenarios, however, a higher inter-quartile range indicates less consistent results. Many participants commented that the high frequency noise could not easily be rated: except for some unvoiced fricatives, there is no spectral overlap between

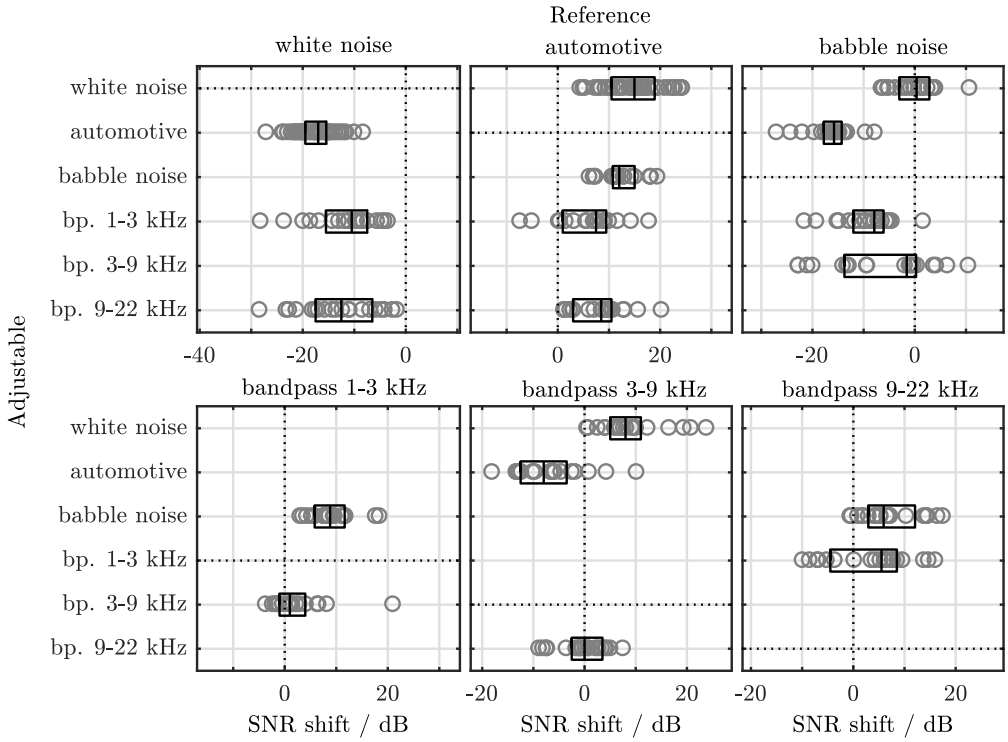


Figure 4.4: Differences of SNR (according to Eq. (4.3)) between reference scenarios (dotted lines) with fixed SNR and scenarios where the SNR was adjustable. The participants adjusted the noise power until the speech similarly set apart from noise in both scenarios. The box plot of median and the quartiles, as well as the individual results of the test subjects indicated by circles are displayed.

speech and noise. Hence the speech component still set apart well even though the noise component already started to get painful. For babble noise, no significant shift is observed compared to white noise. The standard SNR already predicts human perception accurately as the spectral distributions of both noises are similar.

For noises with differing spectral distributions, alternative definitions of the SNR are needed to improve the comparability. These implementations may align the measured values by putting a stronger emphasis on spectral regions that are relevant for speech. Other spectral regions such as the low-frequency components in automotive noise are excluded from the measure.

A well-known psychoacoustical approach for measuring the noise level is to apply an a-weighting function [Zwicker and Fastl, 2013] to the power spectrum before calculating the broadband power. For the SNR, both speech and noise power are determined based on the a-weighted spectra. As depicted in Figure 4.5, the a-weighting considerably improves the comparability between white noise and automotive noise: the mismatch compared to the human perception reduces from 16 dB to only 0.8 dB. For the bandpass-filtered noise examples, however, the mismatch even increases when a-weighting is applied. The static weighting function does not cope well with the high frequency noise.

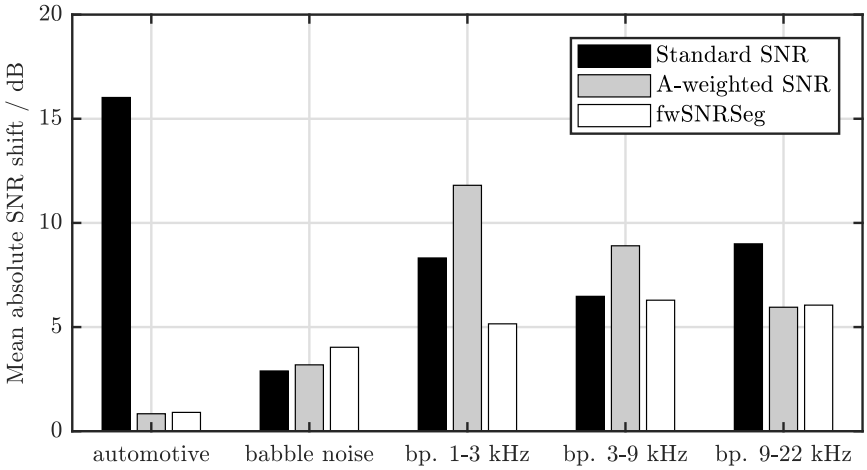


Figure 4.5: Compensation of shifts using different measures: a high mean absolute SNR shift relative to white noise is observable for the standard definition of SNR. By applying A-weighting or calculating a frequency weighted segmental SNR, the shift can be reduced.

Frequency-weighted segmental SNR [Loizou, 2013, Tribolet et al., 1978] hence relies on a dynamic weighting function

$$w_{\text{fwSNRseg}}(k,\ell) = \left(\hat{\Phi}_{ss}(k,\ell)\right)^{0.2} \tag{4.4}$$

that is matched to the current speech spectrum. The measure is based on a local SNR

$$\text{SNR}_{\text{dB}}(k,\ell) = \min\left(\max\left(-10\,\text{dB}, 10\cdot\log_{10}\frac{\hat{\Phi}_{ss}(k,\ell)}{\hat{\Phi}_{bb}(k,\ell)}\right), 35\,\text{dB}\right) \tag{4.5}$$

that is accumulated to a broadband SNR

$$10\cdot\log_{10}\sigma^2_{\text{fwSNRseg}} = \frac{\sum_{k,\ell} w_{\text{fwSNRseg}}(k,\ell)\cdot\text{SNR}_{\text{dB}}(k,\ell)}{\sum_{k,\ell} w_{\text{fwSNRseg}}(k,\ell)} \tag{4.6}$$

using a weighted sum over time and frequency. In contrast to the standard SNR, here the average is calculated over logarithmic values that are limited to the range [-10 dB, 35 dB] in order to reduce the influence of outliers. As shown in Figure 4.5, the measure possesses an improved prediction for automotive noise that is similar to the a-weighting. For the bandpass-filtered noises, the prediction of fwSNRseg outperforms both the standard SNR and the a-weighted SNR.

4.2 Measures for evaluation of VAD results

Noisy speech data along with the corresponding reference for VAD can be used for evaluating VAD approaches: VAD is applied to the audio signal producing a detection result

$VAD(\ell)$ per frame. These results are compared to the reference $VAD_{\text{ref}}(\ell)$ by means of different measures. In the following, traditional as well as a new measure are discussed that address the average detection performance and the transient behavior of VAD.

4.2.1 Receiver operating characteristic

The receiver operating characteristic (ROC) is a well-known measure to investigate the performance of binary detectors. It visualizes the relation between correct detections and false alarms of a detector [Fawcett, 2004]. Using this measure, it is possible to compare the performance of multiple detectors or to find the best parametrization of a tunable algorithm.

In context of VAD, correct detections and false alarms refer to two disjunctive situations where speech is either present or where speech is absent in the signal. During presence of speech, the detection rate (or true positive rate)

$$P_{\text{d}} = \hat{P}(\, VAD(\ell) = 1 |\, VAD_{\text{ref}}(\ell) = 1) \tag{4.7}$$

$$= \frac{\sum_\ell VAD(\ell) \cdot \delta(\, VAD_{\text{ref}}(\ell), 1)}{\sum_\ell \delta(\, VAD_{\text{ref}}(\ell), 1)} \tag{4.8}$$

can be calculated based on the number of frames where the detector indicates speech. High values of P_{d} are desired. Values close to one indicate that all speech was correctly detected whereas lower values correspond to a high amount of missed speech.

Here, the Kronecker-Delta function

$$\delta(u, v) = \begin{cases} 1 & \text{if } u = v \\ 0 & \text{else} \end{cases} \tag{4.9}$$

selects only the frames that contain speech according to the VAD reference. For more detailed analyses, P_{d} can be calculated only on a subset of speech frames, e.g., only for voiced phones. This reveals the detection performance for specific phonetic classes.

Reducing one type of errors typically increases a second type of errors. To find a reasonable tradeoff, multiple types of errors therefore have to be analyzed jointly. Also the measure for correct detections during presence of speech does not fully characterize the detection performance of a VAD. For example, a VAD that always returns "1" irrespective of the actual presence of speech maximizes P_{d}, however, it is obviously not suitable for distinguishing speech from noise. In addition to P_{d}, therefore also its counterpart has to be considered: the false detections during absence of speech. The false-alarm rate (or false positive rate)

$$P_{\text{fa}} = \hat{P}(\, VAD(\ell) = 1 |\, VAD_{\text{ref}}(\ell) = 0) \tag{4.10}$$

$$= \frac{\sum_\ell VAD(\ell) \cdot \delta(\, VAD_{\text{ref}}(\ell), 0)}{\sum_\ell \delta(\, VAD_{\text{ref}}(\ell), 0)} \tag{4.11}$$

is determined based on the number of (false) detections in frames where speech is absent according to the reference. Considering both measures P_d and P_fa jointly, the performances of different VADs can be compared. Optimal detection is achieved when speech is always correctly detected while noise is never accidentally classified as speech ($P_\mathrm{d} = 1$, $P_\mathrm{fa} = 0$).

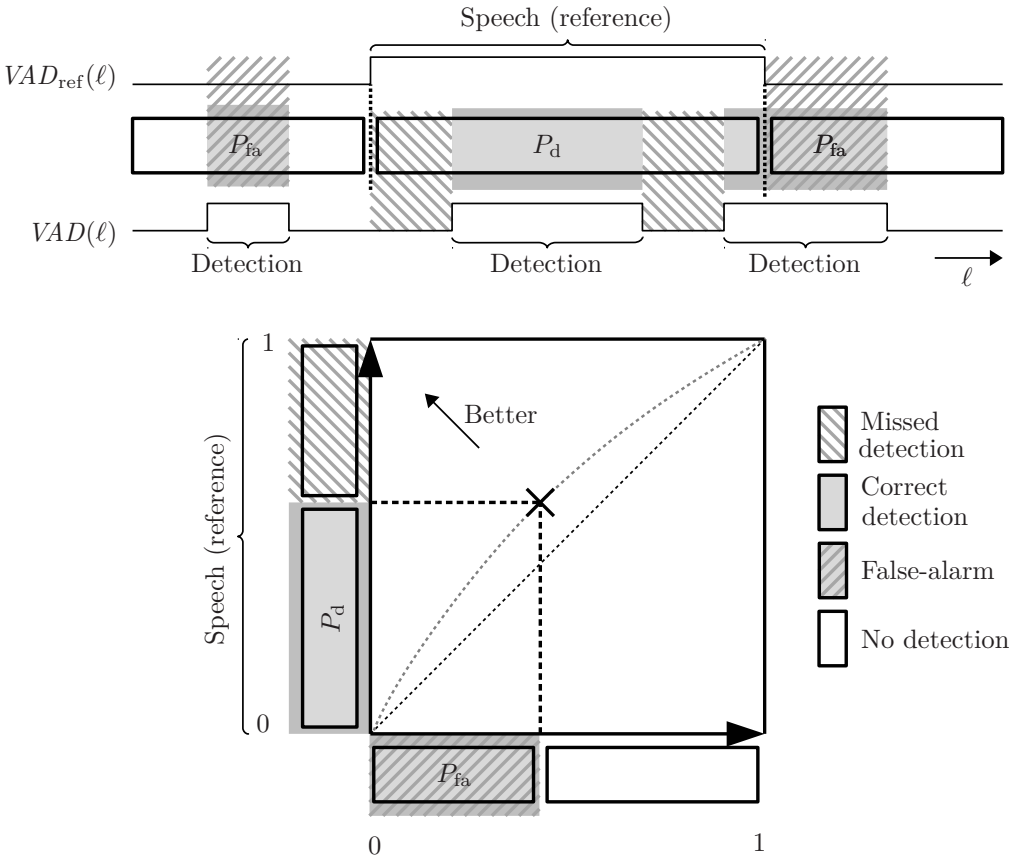


Figure 4.6: Illustration of ROC: detections of $VAD(\ell)$ are compared with a reference $VAD_\mathrm{ref}(\ell)$ (top) to determine errors (striped areas). For ROC (bottom), the rate of correct detections (P_d) is plotted against the rate of false-alarms (P_fa). The operating point (cross) characterizes the performance of a certain configuration of detection. An optimal detector always detects speech correctly ($P_\mathrm{d} = 1$) without any false-alarms ($P_\mathrm{fa} = 0$). The ROC curve (dashed gray line) indicates a set of operating points that can be assumed for varying parametrizations.

Typically, the measures are visualized in a two-dimensional plot as shown in Figure 4.6 with P_fa on the x-axis and P_d on the y-axis. This plot allows for an intuitive comparison of different detectors: operating points close to the upper left corner indicate a good detection performance. For a detector with a fixed parametrization, the performance is indicated by a single operating point. The operating point can be moved by changing the detector's parameters. The basic detector in Eq. (2.17) for example can be parametrized by means of the threshold η. Each threshold value is mapped to a detection rate $P_\mathrm{d}(\eta)$ and false-alarm

rate $P_{\text{fa}}(\eta)$, so that the ROC is a parametric representation depending on the threshold. An operating point further to the left indicates an improved robustness against interferences. However, with increased robustness typically more speech will be missed, so the operating point will also move down.

The envelope curve that encloses all possible operating points for the detector characterizes the detector's performance irrespective of a certain parametrization. The ROC curve connects two extreme cases: the bottom left ($P_{\text{d}} = 0$, $P_{\text{fa}} = 0$) and the top right ($P_{\text{d}} = 1$, $P_{\text{fa}} = 1$) corners corresponding to configurations that never or always detect speech. Both are obviously not suitable for practical applications. A reasonable trade-off has to be found with an operating point that matches the application's requirements. Based on the curve's shape, conclusions on the algorithm performance can be drawn: in general, a curve that approaches the optimum point is desirable. Curves with a steep incline for low false-alarm rates indicate that speech can already be detected to some extent while the noise is still robustly rejected. On the other hand, a little incline can be interpreted as a saturation where the identifiable parts of speech have already been detected and there is no benefit in tolerating additional false-alarms. This is the case for features that are inherently focused on specific parts of speech, e.g., on voiced or unvoiced phones. Even if the detector's sensitivity is increased, it still misses phones that do not belong to the considered group.

ROC curves allow for an intuitive comparison of features, however, with increasing number of features this representation gets cluttered. A measure is desirable that describes the performance of a feature by subsuming the curve into a scalar value. The area that is enclosed by the curve may assume values between zero and one. A high value of the area under curve (AUC) indicates a good performance of the feature. In contrast, the AUC drops to 0.5 when the feature is replaced by a random number. Values lower than 0.5 indicate that the false-alarm rate exceeds the rate of correct detections. In this case, inverting the detection results is advisable as it improves the performance to $\widetilde{AUC} = 1 - AUC > 0.5$.

By means of AUC, the performances of multiple features can easily be compared. The measure reflects the performance by averaging over intervals of speech and speech pauses, respectively. This averaging, however, hides the temporal distribution of errors: errors introduced by a systematically delayed reaction of the detector after speech onsets cannot be distinguished from missed speech during the utterances. To overcome this limitation, additional measures that reveal temporal aspects of the detection performance have to be considered.

4.2.2 Dynamic behavior

Some applications explicitly require a fast reaction of the VAD after speech onsets. For example, for dynamic audio routing in automotive applications, a fast reaction of the VAD is essential to prevent speech onsets from being cut off. In addition to the ROC curve, complementary measures have to be considered to investigate these temporal aspects of the detection and to find features that are suitable for a specific application. Different measures dedicated to the temporal behavior of VAD have been introduced in literature. In Figure 4.7, time intervals are visualized that are relevant for the respective measures.

As discussed, the ROC measure only distinguishes between speech and noise intervals. Detection errors are incorporated in P_d and P_fa irrespective of the temporal position of their occurrence.

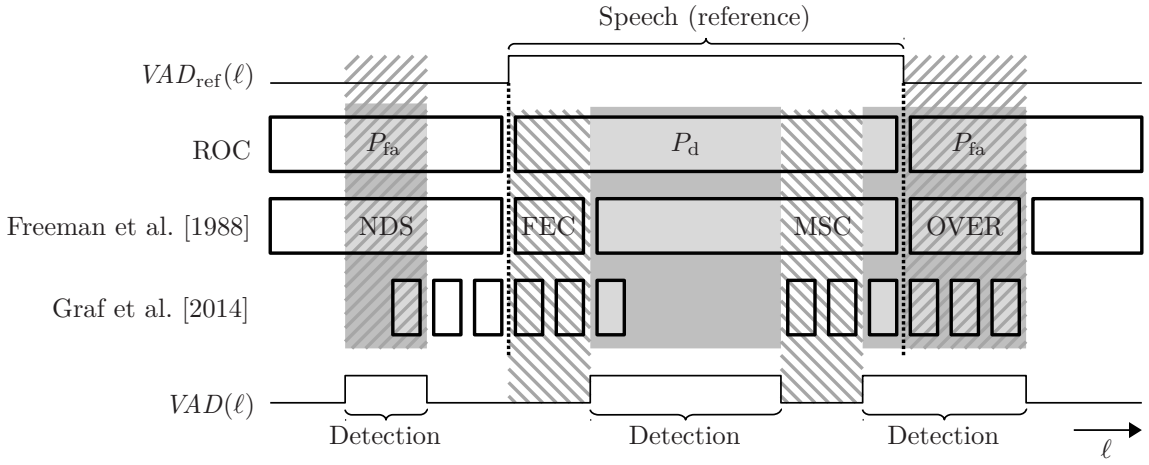


Figure 4.7: Evaluation of the detection (gray areas) of a VAD algorithm by determining errors (striped areas) compared to the reference: for ROC curves, VAD results are averaged over reference speech and noise intervals to determine probability of detection (P_d) and false alarm (P_fa). The measure by Freeman et al. [1988] considers four intervals, bounded by reference and detection end-points, to determine FEC, MSC, OVER and NDS. The fine-grained measure [Graf et al., 2014] increases the temporal selectivity by a frame-wise evaluation around reference speech on- and offsets. The latter captures the dynamic behavior by averaging only over utterances but not over time.

Freeman et al. [1988] therefore introduced four measures that explicitly consider speech onsets and the interval after the end of speech [Beritelli et al., 1998]. Two of these measures are similar to ROC:

- Likewise P_fa, noise detected as speech (NDS) addresses the false alarms during absence of speech.
- Mid speech clipping (MSC) is concerned with speech portions that are missed by the detector, similar to $1 - P_\mathrm{d}$.

Intervals after speech onsets and after the end of speech, however, are excluded from these measures. Instead, these intervals are explicitly considered by two dedicated measures:

- Front end clipping (FEC) addresses the first part of speech utterances that is frequently missed due to a delayed reaction of the detector on speech onsets.
- Hangover after speech (OVER) quantifies false detections after the utterance. This type of error can typically be observed when the detection result is held for a short time to prevent detection dropouts during continuous speech intervals.

These traditional measures may give a first impression of the temporal behavior. The transient behavior, however, cannot be analyzed in detail as the measures still incorporate averaging over time.

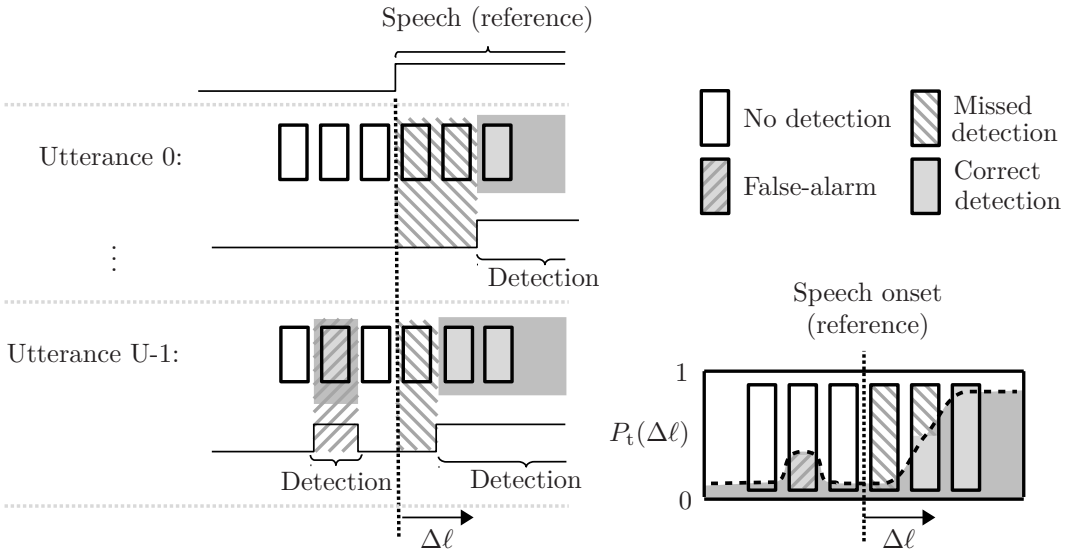


Figure 4.8: Example of transient measure: a time window around speech onsets is analyzed (left) and the results of all utterances are temporally aligned. Statistics over utterances (right) reveal the temporal evolution $P_\mathrm{t}(\Delta\ell)$ of VAD before and after the onset. After the onset, the measure typically stays low for some frames before a rising slope indicates that speech is detected. By means of the measure, the latency introduced by one detector can be evaluated and compared with other algorithms.

In [Graf et al., 2014], a more fine-grained measure was introduced. Instead of averaging over longer time intervals, this evaluation specifically focuses on speech onsets relying on statistics over multiple utterances. As illustrated in Figure 4.8, a time window around the onset is considered for each utterance $u \in \{0, \cdots, U-1\}$. The different windows are then temporally aligned with respect to the onset positions $\ell_\mathrm{on}(u)$. This way, a transient measure

$$P_\mathrm{t}(\Delta\ell) = \hat{P}(\,VAD(\ell)|\ell = \ell_{\mathrm{on},u} + \Delta\ell) \tag{4.12}$$

$$= \frac{1}{U}\sum_{u=0}^{U-1} VAD(\ell_{\mathrm{on},u} + \Delta\ell) \tag{4.13}$$

can be calculated that ultimately refers to a certain instant of time $\Delta\ell$ relative to the onset position. For example, the rate of detection 200 ms after speech onsets can be assessed using this statistics over utterances.

This time-dependent measure preserves information on the temporal evolution of a VAD. The transient behavior can be characterized intuitively by plotting the measure over the time-lag relative to the onset position as exemplified in Figure 4.9.

Before the onset ($\Delta\ell < 0$) speech is absent by definition which corresponds to the situation that is relevant for the false-alarm rate in Eq. (4.10). In this region, the transient measure therefore typically assumes low values close to the false-alarm rate P_{fa}. The measure stays at this low level until speech is detected at or shortly after the onset ($\Delta\ell \geq 0$). A rising slope of the detection rate can be observed that carries information on the transient behavior: the latency introduced by the detector can be determined based on the temporal position and the steepness of the slope. The curve's course during presence of speech finally relates to the detection rate P_{d} in Eq. (4.7). The measure approaches this value whereby in many cases a small overshoot can be observed at the beginning of the utterance that tends to be stressed.

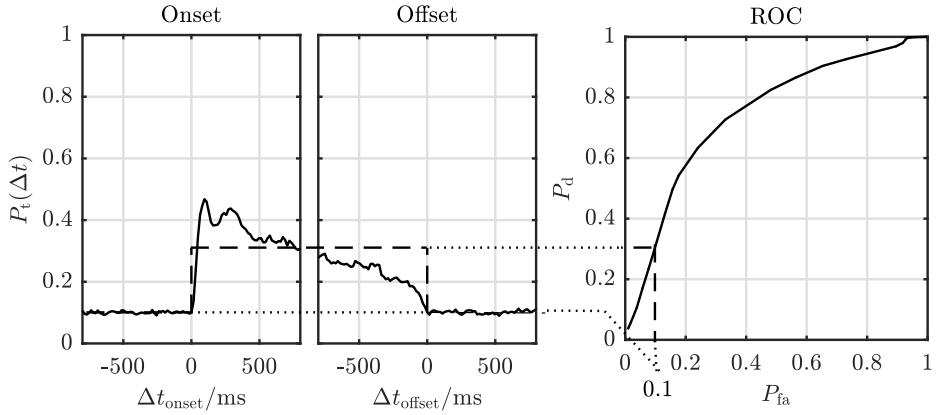


Figure 4.9: Transient measure and ROC curve: the ROC reflects the performance by averaging detections for speech intervals and for intervals where speech is absent. In contrast, the transient measure captures the temporal evolution of detection for one operating point (corresponding to a false-alarm rate $P_{\mathrm{fa}}(\eta_{0.1}) = 0.1$ in this example) by averaging over utterances instead of time intervals. Here, the transient measure around onsets and offsets is depicted whereas in the following only the onsets are considered that are important for a quick reaction of the features.

4.3 Comprehensive evaluation of features for VAD

An overview of features for VAD was presented in Chapter 3. There, the features were categorized according to speech characteristics that are exploited. The behavior during presence of speech was exemplified by means of one utterance. In contrast, the effect of interferences on the feature's performance was so far not considered in this thesis.

In this section, the detection performance for varying conditions is evaluated by means of the discussed measures. Several noise situations are considered for a comprehensive analysis that likely exceeds the scope of a single application. Again, the features are

grouped according to their respective speech characteristics but now a stronger focus is on the influence of noise.

The noise data for this analysis was adopted from two databases: QUT-NOISE [Dean et al., 2010] and NOISEX-92 [Varga and Steeneken, 1993] that were already discussed in Section 4.1.1. Only a subset of NOISEX-92 comprising real noise recordings (i.e., car, factory, operation room, F-16, Lynx, and machine gun) was selected. Due to the rather small extent of NOISEX-92, neural networks for fusing multi-dimensional features to scalar values were trained only on the training dataset of QUT-NOISE. Overall, 16 noise scenarios were considered for the evaluations, whereas the training relied on 10 scenarios.

The TIMIT database [Garofolo et al., 1993] was employed as source for the speech components. Since TIMIT utterances do not contain a significant amount of leading and trailing silence, 1.5 s of speech pause were added before and after each utterance leading to an average duration of 6 s per file. For some reverberant noise conditions, the QUT-NOISE corpus includes corresponding impulse responses. For these examples, the speech was convolved with the impulse responses for a more realistic simulation. In any case, the speech signal was rescaled and mixed to the noise for different SNRs. The Lombard effect [Junqua, 1996] was not considered.

Eight different SNR levels in the range between -5 dB and 15 dB were simulated. For each noise and SNR condition, 40 speech files were randomly selected resulting in 3200 noisy speech utterances for training and 5120 utterances for the evaluation. The final database contained about 14 hours of audio data.

For the analysis, the signals were resampled to a sample rate of $f_\mathrm{s} = 16\,\mathrm{kHz}$. Frames of length $N = 512$ samples were extracted with a shift of $R = 256$ samples between adjacent frames. Features in frequency domain relied on an FFT with a rectangular window that resulted in $K = 257$ complex-valued spectral bins per frame. Power spectra were estimated using Eq. (2.14) with a smoothing constant $\alpha_\mathrm{PSD} \hat{=} 200\,\mathrm{dB/s}$.

All multi-dimensional features in this chapter were fused to scalar decision variables using the neural network approach described in Section 3.4.

4.3.1 Power and SNR

For scenarios with low to moderate noise levels, features based on power and SNR appear to be obvious first candidates for VAD. Interpreting high power signal components as speech is often adequate in this case as discussed in Section 3.2.2. Even a very basic feature using the plain short-term power f_STP in Eq. (2.16) can be employed for speech detection when the SNR is high. The measurement results presented in Table 4.2 show that the performance of VAD drops with decreasing SNR. Applying different kinds of normalization helps improving the detection performance for lower SNRs.

For a normalization according to Eq. (3.2), the power is normalized such that the highest peak in the signal is mapped to a feature value $f_\mathrm{NSTP} = 1$ whereas the lowest power corresponds to a value zero. By doing so, the performance can be slightly improved. Detection results that are comparable to the short-term power feature are achieved for about 2 dB lower SNRs. This procedure, however, is limited in many respects: on the one

Table 4.2: Comparison of power and SNR-based features: the AUC is depicted as a function of SNR with colors on a scale from red (low performance) over yellow (reasonable performance) to green (good performance).

SNR [dB]	-5	0	2	4	6	8	10	15	all
Short-term power Eq. (2.16)	0.58	0.65	0.70	0.74	0.79	0.82	0.86	0.90	0.75
Normalized power (3.2)	0.64	0.71	0.75	0.78	0.82	0.84	0.86	0.88	0.78
SNR1 (3.8)	0.64	0.73	0.76	0.81	0.83	0.87	0.89	0.93	0.81
Power envelope dynamics (3.3)	0.70	0.79	0.83	0.86	0.88	0.90	0.92	0.94	0.85
Long-term spectral div. (3.14)	0.68	0.79	0.83	0.85	0.90	0.92	0.94	0.96	0.86
SNR2 (single frame) (3.12)	0.79	0.86	0.88	0.88	0.87	0.87	0.86	0.84	0.85

hand, the maximum value has to be known in advance which restricts its use to offline applications where a complete file is accessible. On the other hand, varying conditions are not taken into account by the fixed normalization.

More advanced approaches track the noise level and apply a dynamic normalization. For the SNR feature according to Eq. (3.8), the noise estimate is updated in case that the VAD indicates absence of speech. For the feature f_{SNR1}, the short-term power is normalized with respect to the estimated noise level. This recursive procedure results in improvements of detection performance that are similar compared to the first normalization approach. However, the mechanism is more flexible since it adapts to varying noise conditions and it can be applied for online applications without knowledge of the complete signal.

The feature $\boldsymbol{f}_{\mathrm{PED}}$ by Marzinzik and Kollmeier [2002] based on power envelope dynamics tracks both the noise level as well as the speech level for three frequency regions. The 6-dimensional feature vector according to Eq. (3.3) contains the dynamic ranges and the normalized short-term power values. For this analysis, the vector is fused to a scalar value using the neural network described in Section 3.4. An improved detection performance can be observed for this feature. Even for 0 dB SNR the detection performance is acceptable.

Similar improvements can be achieved using the long-term spectral divergence feature f_{LTSD} by Ramírez et al. [2004a] according to Eq. (3.14). This feature explicitly considers a longer time range which was observed to be beneficial for VAD. Many later approaches, e.g., features based on modulation adopt this finding and rely on information that can be derived from the temporal context.

The last feature in this category is based on the power ratio f_{SNR2} between spectral bins containing the highest and the lowest power of the current frame according to Eq. (3.12). These bins are assumed to relate to speech and noise respectively. Normalizing the spectrum by means of the temporal average is essential for the feature. The signal gets whitened and the spectral envelope is flattened. Otherwise the detection might be disturbed by the coloration of noise, e.g., by dominant low-frequency components in automotive noise. A surprisingly good detection performance can be observed in particular for very low levels of SNR. In contrast, the detection performance decreases for high values of the SNR where the contribution of noise vanishes. In this case, even the lowest bins are dominated by speech, however, the algorithm still attributes bins to noise erroneously. This explains

Table 4.3: Comparison of features for detection of voiced and unvoiced speech: the AUC is depicted as a function of voicing with colors on a scale from red (low performance) over yellow (reasonable performance) to green (good performance).

Voicing	voiced	unvoiced	all
Zero-crossing rate Eq. (3.15)	0.55	0.79	0.60
Spectral entropy (3.17)	0.44	0.71	0.49
Max. ACF (3.20)	0.75	0.34	0.66
NN applied to ACF (3.21)	0.84	0.66	0.80
Harmonic product spectrum (3.25)	0.79	0.62	0.76
Cepstral peak (3.23)	0.85	0.80	0.84

the comparatively poor performance for SNRs where other algorithms achieve their best results.

4.3.2 Voicing

The second category of features summarized in Section 3.2.3 addresses the voicing properties of human speech. Evaluation results for these features are shown in Table 4.3 subdivided into voiced and unvoiced speech. In this analysis, the detection rate P_d was evaluated separately for time intervals containing voiced and unvoiced phones. Preliminary analyses revealed that the algorithms may benefit from a noise reduction as a preprocessing. Hence, a Wiener filter with maximum of 6 dB attenuation was applied to the data before calculating the features. The noise spectrum was estimated using a minimum mean squared error (MMSE)-based estimator introduced in [Gerkmann and Hendriks, 2012].

The first two features based on zero-crossing rate f_ZCR and spectral entropy f_SE according to Eq. (3.15) and Eq. (3.17) target on unvoiced speech components. The analysis reveals that both features indeed perform reasonably for unvoiced phones, whereas voiced speech is barely detected. Even though the zero-crossing rate is a low-complex feature in time domain, it outperforms the spectral entropy. However, since only a small fraction of phones is unvoiced, the overall performance on speech signals is rather poor for both features. A VAD should therefore not purely rely on features dedicated to unvoiced speech. Instead, they can be considered jointly, e.g., with features that target on harmonic components of voiced speech.

Most phones including vowels but also many fricatives can be attributed to voiced speech. Detection of these components is therefore a pivotal factor for VAD. Several features were introduced to assess the harmonicity of a signal.

In time domain, the auto-correlation function is a typical approach to deal with harmonic components. Both the degree of voicing as well as the pitch frequency can be determined based on the ACF according to Eq. (3.21). Presence or absence of voiced speech can be detected using the maximum value f_ACF of the normalized ACF as depicted in the table. More information for VAD can be extracted by applying a neural network to the normalized ACF. This way, a significantly higher performance can be achieved.

Table 4.4: Comparison of features based on the formant structure of speech: the AUC is depicted as a function of SNR with colors on a scale from red (low performance) over yellow (reasonable performance) to green (good performance).

SNR [dB]	-5	0	2	4	6	8	10	15	all
LPC coefficients Eq. (3.53)	0.59	0.66	0.67	0.70	0.71	0.71	0.73	0.74	0.69
Line spectral frequencies (3.55)	0.72	0.79	0.82	0.84	0.85	0.87	0.88	0.89	0.83
Cepstrum (3.56)	0.67	0.75	0.80	0.84	0.88	0.90	0.93	0.96	0.85
Mel-filtered spectrum (3.57)	0.71	0.81	0.86	0.90	0.93	0.95	0.97	0.98	0.90

In frequency domain, accurate estimation of pitch frequencies is difficult due to the limited spectral resolution. Detection of voiced speech, however, is possible by means of the maximum value f_{HPS} of the harmonic product spectrum according to Eq. (3.25). An improved performance can be observed compared to the maximum of ACF, however, it is not as good as the neural network approach.

In this analysis, the best detection performance for both voiced and unvoiced speech is achieved using a cepstral-based approach f_{CEP1} according to Eq. (3.23). The cepstrum separates the spectral fine-structure from the spectral envelope. This feature hence reflects not only the harmonic properties but also the formant structure that will be discussed in the following.

4.3.3 Formant structure

The formant structure of speech addresses frequency regions that are emphasized due to resonances in the vocal tract. As discussed in Section 3.2.5, different phones can be distinguished by means of the formant structure. Features based on this property are therefore essential for speech recognition. In this section, the performance of formant-based features for VAD is analyzed. For all features, neural networks were applied to fuse the multi-dimensional features into a scalar value. Evaluation results are depicted in Table 4.4.

Linear predictive coding coefficients $\boldsymbol{f}_{\mathrm{LPC}}$ model the spectral envelope by means of an IIR filter according to Eq. (3.53). These coefficients are commonly used for speech coding as they can be calculated efficiently and represent the spectral envelope in a compact form. For VAD, however, this plain feature appears to be disadvantageous. The neural network does not perform well with the LPCs in this experiment. The relation between filter coefficients and spectral envelope is quite complex and cannot adequately be learned using a single hidden layer with only 20 neurons.

Converting the LPCs into line spectral frequencies $\boldsymbol{f}_{\mathrm{LSF}}$ according to Eq. (3.55) significantly improves the detection performance. Even though both representations are equivalent, the neural network benefits from the different model parameters. The LSFs cluster around the formant frequencies and are hence much closer related to the spectral envelope compared to the original filter coefficients.

As discussed before, the cepstrum covers both the harmonic structure as well as the

formant structure of speech. In this section, the latter is addressed by means of a neural network that is trained to detect speech based on the lower order cepstral coefficients $\boldsymbol{f}_{\mathrm{CEP2}}$ according to Eq. (3.56). On average, a similar performance can be observed compared to LSF. For higher SNRs the cepstrum outperforms the LSF whereas it performs slightly worse for low SNRs.

The feature $\boldsymbol{f}_{\mathrm{MFCC}}$ based on a mel-filtered spectrum shows the best performance in this analysis. This perceptually-motivated representation compresses the spectral envelope by subsuming frequency bins into wider bands according to Eq. (3.57). The resolution varies over frequency: narrow bands are applied for lower frequencies whereas for higher frequencies wider bands are employed. Overall 20 bands were considered for this analysis.

4.3.4 Stationarity and modulation

The features that were analyzed in the previous sections focus on speech characteristics that can instantaneously be detected. It became obvious that different classes of phones require different types of features. So far, the different phones were considered separately. The sequential structure of speech based on concatenated phones was disregarded.

In this section, features are analyzed that employ a longer temporal context for the detection of speech. These features target on non-stationarity or characteristic modulation structures of speech signals. Both properties can be relevant for speech detection as in many cases the background noise is either stationary or shows a modulation structure that differs from the modulation of speech. Different types of background noise were considered for the evaluation results depicted in Table 4.5 to analyze the features' robustness against different types of interferences.

The first two approaches in this section directly address the non-stationarity of speech. The long-term (spectral) flatness measure according to Eq.(3.61) is based on the ratio of geometric and arithmetic mean of spectral bins over time. This ratio is finally averaged over frequency for the feature f_{LSFM}. As shown in the table, the feature performs well for two stationary scenarios (car and pool) whereas the performance drops for non-stationary background noises. A particularly bad performance can be observed for the highly non-stationary noise of a machine-gun.

For the long-term signal variability feature f_{LTSV} according to Eq. (3.60), the stationarity is measured by means of the spectrum's temporal entropy. The variance over frequency is calculated to generate a broadband result. As shown in the table, the feature outperforms f_{LSFM} in most cases. For machine-gun, the best result of all features in this category is achieved. The spectral variance increases the robustness against interferences that affect the full spectrum.

The last three features deal with the modulation structure of speech. All these features are based on a mel-filtered spectrogram with 20 bands. A modulation peak in the range of 4 Hz reflects the typical syllable rate of human speech. This peak is assessed by a basic feature f_{M4Hz} using a bandpass filter according to Eq. (3.62). A similar performance compared to the features dedicated to non-stationarity of speech can be observed by this feature that considers only a single modulation frequency.

Table 4.5: Comparison of features based on non-stationarity and modulation: the AUC is depicted as a function of noise scenario with colors on a scale from red (low performance) over yellow (reasonable performance) to green (good performance).

(a) QUT-NOISE

Noise scenario	cafe	kitchen	car[a]	pool[b]	···	all
Long-term (spectral) flatness Eq. (3.61)	0.79	0.76	0.94	0.94		0.87
Long-term signal variability (3.60)	0.78	0.80	0.96	0.95		0.89
4-Hz modulation (3.62)	0.79	0.78	0.89	0.93		0.86
Amplitude modulation spectrogram (3.65)	0.86	0.84	0.94	0.95		0.90
Spectro-temporal modulation (3.67)	0.95	0.88	0.99	0.99		0.96

[a] closed window
[b] reverberant

(b) NOISEX

Noise scenario	f-16 jet	factory	machine gun	···	all
Long-term (spectral) flatness Eq. (3.61)	0.92	0.86	0.47		0.82
Long-term signal variability (3.60)	0.94	0.85	0.75		0.90
4-Hz modulation (3.62)	0.94	0.86	0.66		0.85
Amplitude modulation spectrogram (3.65)	0.89	0.88	0.63		0.84
Spectro-temporal modulation (3.67)	0.85	0.98	0.64		0.91

Multiple modulation frequencies are jointly considered for an amplitude modulation spectrogram according to Eq. (3.65). In this analysis, again a neural network is applied that fused the feature vector $\boldsymbol{f}_{\text{AMS}}$ to a scalar value. Improvements can be observed for the QUT-NOISE database whereas the highly non-stationary scenarios taken from NOISEX are still challenging.

Spectro-temporal modulation structures are taken into account by the last feature in this analysis. The spectrogram is convolved with 2-dimensional modulation patterns that reflect both the temporal structure as well as the spectral shape of speech. Gabor filters are chosen as suggested in [Schädler et al., 2012]. The feature vector $\boldsymbol{f}_{\text{STM}}$ according to Eq. (3.67) contains 771 elements which by far exceeds the size of all other features discussed here. Correspondingly, the feature outperforms the other approaches in most scenarios.

Comparing the transient behavior for multiple features as depicted in Figure 4.10, another effect of the long temporal context gets visible. As discussed, the modulation-based feature outperforms the other features in terms of detection rate. However, as a drawback the detection is delayed compared to the almost instantaneous reaction of power-based or voicing features.

These advantages and drawbacks of different features may be more or less important depending on the target application. The dissected findings of this chapter hence will subsequently be contextualized in relation to multiple applications with diverse requirements on VAD.

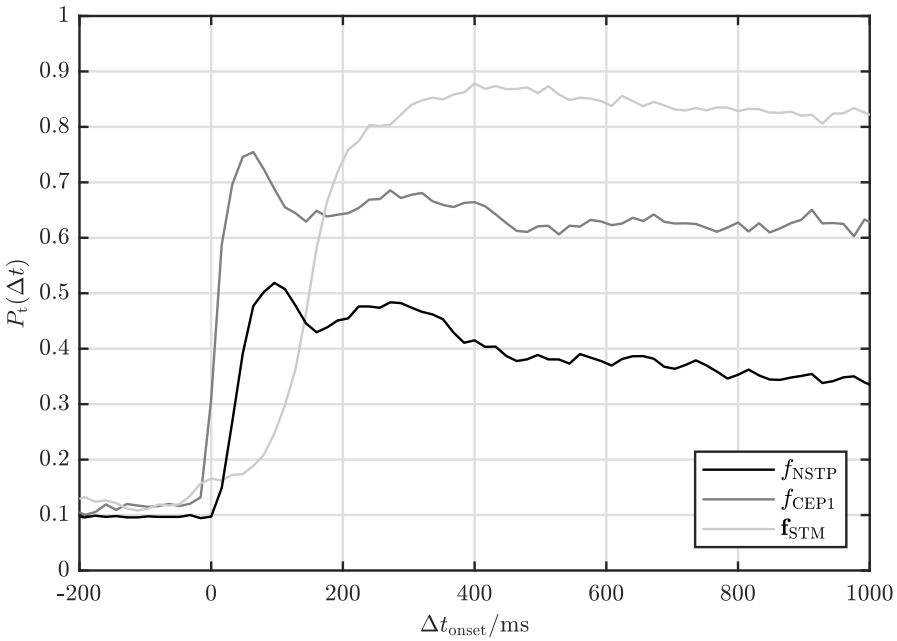


Figure 4.10: Transient measure for three feature classes: the power-based feature f_{NSTP} and the voicing feature f_{CEP1} both indicate the speech onset almost instantaneously. The modulation-based feature $\boldsymbol{f}_{\mathrm{STM}}$ outperforms the others in terms of detection rate, however, this benefit is attended by an increased delay of about 150 ms.

Chapter 5

Application-specific Evaluation

Diverse speech-driven applications make use of VAD as an essential component that controls other algorithms. Two applications with very different requirements on VAD are exemplified in this chapter:

- First, a noise suppression system dedicated to the suppression of babble noise will be discussed. In this system, the VAD has to distinguish the desired speech from babble noise. This detection is particularly challenging since babble noise itself consists of a mixture of undesired interfering speech components. A feature combination will be considered that is robust against this type of interferences. Evaluation results gained by a subjective listening test as well as an objective measure will finally be presented that highlight the benefit of a suitable VAD for the full application.
- The second example concerns with the application of VAD in ICC systems. In automotive context, the typical background noise is rather stationary and hence less challenging compared to the babble noise scenario. The low latency that is needed for the signal processing in ICC systems, however, goes along with very short window lengths and a low spectral resolution. To deal with this particular challenge, dedicated features for VAD are needed as discussed before. Different combinations of these features are introduced and their applicability for algorithms in ICC applications is analyzed.

5.1 Speech detection for babble noise suppression

Noise suppression for mobile devices such as smartphones is difficult due to the huge variety of different noise scenarios. These devices can be employed in almost any environment with the consequence that the noise does not only incorporate stationary but also non-stationary components. In crowded surroundings, the noise might even include interfering speech components which is a challenge particularly for VAD.

Distinguishing the desired foreground speech from interfering speech components in the background necessitates elaborate speech detectors. For finding acoustic cues that can

be employed for the detection, the characteristics of babble noise have to be investigated [Krishnamurthy and Hansen, 2009]. Babble noise consists of a mixture of multiple interfering speech components. Compared to the desired speaker, a larger distance between interfering speakers and the device is expected. Hence the interfering speech components are more reverberant.

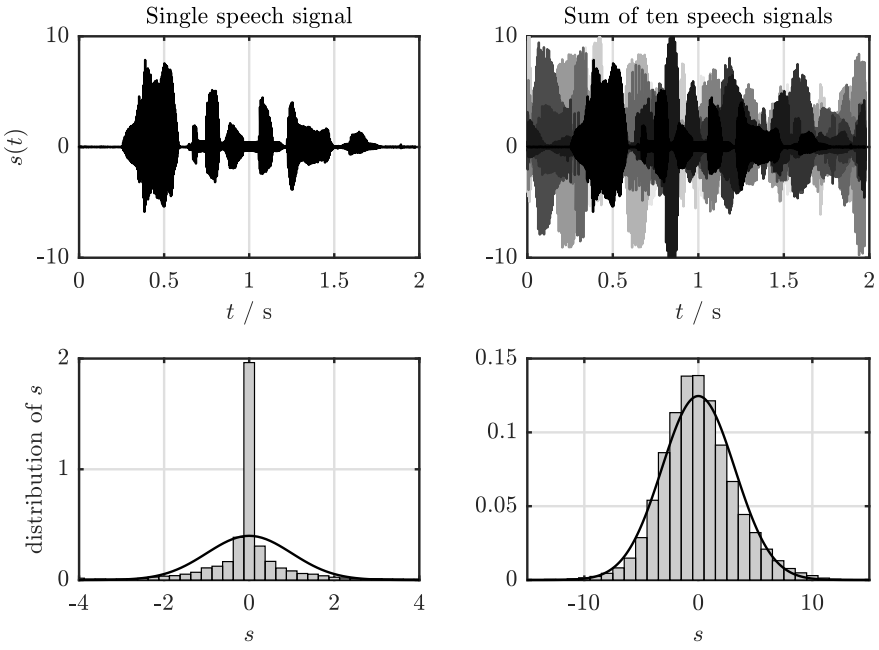


Figure 5.1: Single speech vs. mixture of multiple speech components: a single speech component is sparse with most amplitude values close to zero as indicated by a prominent peak in the histogram. The histogram approaches a Gaussian distribution when mixing multiple speech components.

Reverberation and accumulation of multiple independent components both have an effect on the distribution of the signal's sample values. For a single close-talk signal, most sample values are clustered close to zero with only few outliers. Considering the probability density function (PDF) of such a sparse signal, a peak around zero can be observed. As shown in Figure 5.1, the PDF becomes less sparse and converges to a Gaussian distribution when multiple independent speech components are accumulated. This observation is in accordance with the central limit theorem in probability theory stating that the sum of multiple independent and identically distributed random variables tends to a Gaussian distribution [Bronštejn et al., 2008]. The Gaussianity or non-sparseness of a signal hence is a good indicator for the severity of babble noise [Li and Lutman, 2006].

The Gaussianity of a random variable χ can be quantified by means of the kurtosis

$$\mathrm{KUR}\{\chi\} = \frac{\mathrm{E}\{\chi^4\}}{\left(\mathrm{E}\{\chi^2\}\right)^2} - 3 \tag{5.1}$$

that is based on the ratio between 4th central moment and the squared variance. Here, the random variable is expected to have zero mean $\mathrm{E}\{\chi\} = 0$ which simplifies the expression. For a Gaussian-distributed random variable, the kurtosis is zero. The measure increases for super-Gaussian distributions that have a high peak around zero and long tails.

Several algorithms were introduced in the context of speech processing that make use of estimates of the kurtosis:

- Li and Lutman [2006] investigated the relation between kurtosis and the capability of human speech perception in presence of babble noise. With increasing number of contributors in the babble noise, the kurtosis measure as well as the human ability for perceiving one speaker's contribution decrease. The authors concluded that kurtosis is a good predictor for human speech recognition in babble noise. The kurtosis appears particularly promising for real-time processing since it can easily be estimated based on the time-domain signal.

- Independent component analysis (ICA) [Hyvärinen and Oja, 2000] for blind source separation targets on separating multiple signal components based on recordings of their mixture. Some approaches separate the signals subject to their kurtosis values: dependencies are reduced by finding components with maximum kurtosis.

- Reverberation of the signal also results in a more Gaussian distribution of the sample values and hence a lower kurtosis compared to close-talk speech. This effect was employed in [Hayashida et al., 2014] for close and distant talk discrimination based on the signal's kurtosis. Methods for dereverberation of speech were introduced in [Gillespie et al., 2001] and [Wu and Wang, 2006]. These algorithms process the signal in such a way that the kurtosis is maximized.

- Kurtosis was found to be a good predictor of the severity of musical tones that may remain as artifacts after noise suppression, e.g., in [Yu and Fingscheidt, 2012].

- In multiple publications the application of kurtosis for VAD was discussed, e.g., in [Nemer et al., 2001], [Cournapeau et al., 2006], and [Cournapeau and Kawahara, 2007].

In the following, a feature combination for VAD is introduced that is robust against babble noise. This combination is later employed for the suppression of babble noise.

5.1.1 Feature combination

The system for babble noise suppression that is discussed in this section has been introduced first in [Graf et al., 2016a]. The system consists of a speech detector that controls the actual noise suppression. Both detection and suppression are designed for the particular challenges of babble noise scenarios.

In the system, the detection of desired speech components in babble noise relies on a time-dependent estimate

$$KUR(\ell) = \frac{\overline{x^4}(\ell \cdot R)}{\left(\overline{x^2}(\ell \cdot R)\right)^2} - 3 \tag{5.2}$$

of the kurtosis based on the signal in time domain. The 4th moment and the variance are estimated by

$$\overline{x^4}(n) = (1 - \alpha_{\mathrm{KUR}}) \cdot x^4(n) + \alpha_{\mathrm{KUR}} \cdot \overline{x^4}(n-1) \tag{5.3}$$

and

$$\overline{x^2}(n) = (1 - \alpha_{\mathrm{KUR}}) \cdot x^2(n) + \alpha_{\mathrm{KUR}} \cdot \overline{x^2}(n-1) \tag{5.4}$$

respectively. Both estimators are implemented efficiently using recursive smoothing with a smoothing constant $\alpha_{\mathrm{KUR}} \hat{=} -100\,\mathrm{dB/s}$ (≈ 0.9986 for $f_\mathrm{s} = 16\,\mathrm{kHz}$). A zero-mean signal is assumed here. Applying a DC-blocker to the original signal hence might be necessary.

The plain kurtosis measure reacts almost instantaneously to outliers in the signal that are interpreted as speech. Similarly, it drops quickly after speech onsets. To reduce these fluctuations, the feature is smoothed again using a moving average filter

$$f_{\mathrm{KUR}}(\ell) = \frac{1}{2L_{\mathrm{KUR}} + 1} \sum_{\tilde{\ell}=-L_{\mathrm{KUR}}}^{L_{\mathrm{KUR}}} KUR(\ell + \tilde{\ell}) \tag{5.5}$$

that incorporates L_{KUR} frames from the past as well as a look-ahead of the same length. Here, the temporal context was set to $L_{\mathrm{KUR}} = 10\,\mathrm{frames}$ corresponding to 160 ms.

As illustrated in Figure 5.2, the kurtosis captures most of the speech. Presence of desired foreground speech is reflected by high values of the feature. In contrast, the value stays close to zero during presence of babble noise. Vowels such as "/u/" that primarily excite low frequencies, however, are not reliably detected. Considering a second feature dedicated to voiced speech is therefore beneficial.

In Section 4.3.2, the cepstum was identified to provide valuable information on presence of speech. Voiced speech but also unvoiced speech can be detected by means of the cepstral peak. A similar feature is employed here for the detection of desired speech in babble noise.

The cepstrum is calculated according to Eq. (3.22) and smoothed along both frames and cepstral coefficients

$$\overline{CEP}(\tau, \ell) = \frac{1}{(2T_{\mathrm{CEP3}} + 1) \cdot L_{\mathrm{CEP3}}} \sum_{\tilde{\tau}=-T_{\mathrm{CEP3}}}^{T_{\mathrm{CEP3}}} \sum_{\tilde{\ell}=0}^{L_{\mathrm{CEP3}}-1} CEP(\tau + \tilde{\tau}, \ell - \tilde{\ell}) \tag{5.6}$$

with $T_{\mathrm{CEP3}} = 1$ and $L_{\mathrm{CEP3}} = 4$ frames. The maximum value of $\overline{CEP}(\tau, \ell)$ in the range between 60 Hz and 300 Hz is employed and smoothed again for the final cepstral peak feature

$$f_{\mathrm{CEP3}}(\ell) = (1 - \alpha_{\mathrm{CEP3}}) \cdot \left(\max_{\tau} \left(\overline{CEP}(\tau, \ell) \right) - \overline{CEP}_{\mathrm{offset}} \right) + \alpha_{\mathrm{CEP3}} \cdot f_{\mathrm{CEP3}}(\ell - 1) \tag{5.7}$$

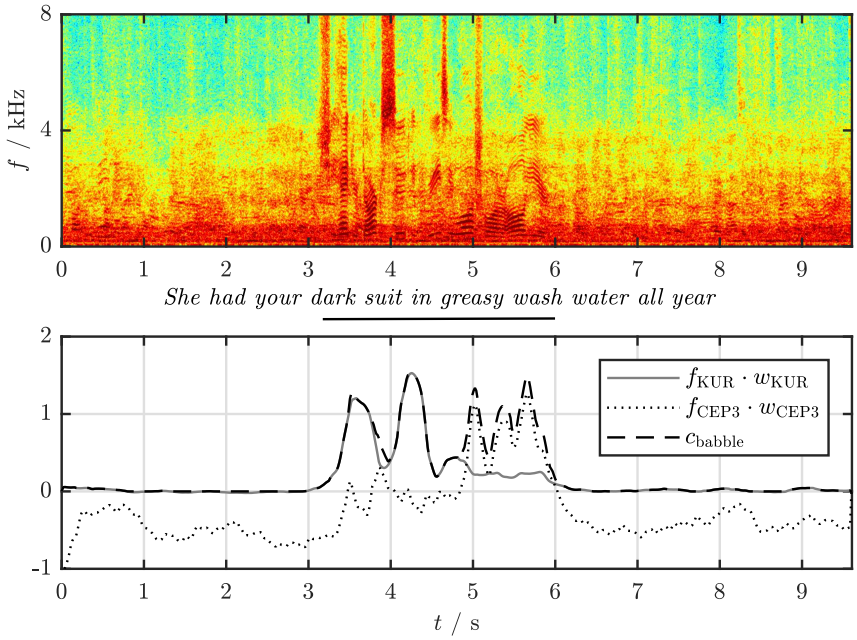


Figure 5.2: Kurtosis and combination with voicing feature: using the feature f_{KUR}, already most of the speech can be detected. For voiced speech in the lower frequencies (as shown in the spectrogram in the upper plot), however, the feature f_{CEP3} is more suitable. A combination c_{babble} of both features is capable of detecting the complete TIMIT speech utterance during presence of babble noise.

using a smoothing constant $\alpha_{\mathrm{CEP3}} = 0.9 \mathrel{\hat{=}} -30\,\mathrm{dB/s}$. As the feature should be below zero during absence of desired speech, an experimentally determined offset $\overline{CEP}_{\mathrm{offset}} = \frac{1}{6}$ is subtracted.

Both original features, kurtosis and cepstral peak, are not normalized. In order to make the values comparable, the features are rescaled with $w_{\mathrm{KUR}} = 1$ and $w_{\mathrm{CEP3}} = \frac{2}{3}$. A combined feature for speech detection in babble noise

$$c_{\mathrm{babble}}(\ell) = w_{\mathrm{KUR}} \cdot \max(0, f_{\mathrm{KUR}}(\ell)) + w_{\mathrm{CEP3}} \cdot \max(0, f_{\mathrm{CEP3}}(\ell)) \tag{5.8}$$

is then calculated based on the weighted sum of both features.

This combined feature is applied in a noise suppression system that is dedicated to the suppression of babble noise. The speech detection is employed to reduce the system's aggressiveness dynamically in order to prevent distortions of the desired speech. The noise suppression system as shown in Figure 5.3 is based on a Wiener filter that was briefly discussed in Section 2.2.1. Here, some modifications of the filter according to Eq. (2.13) are introduced that enable the system to suppress babble noise more aggressively. In Figure 5.4, the effects of the different modifications are exemplified for one spectrogram of speech in babble noise.

Overestimating the noise is a common method to increase the Wiener filter's aggressive-

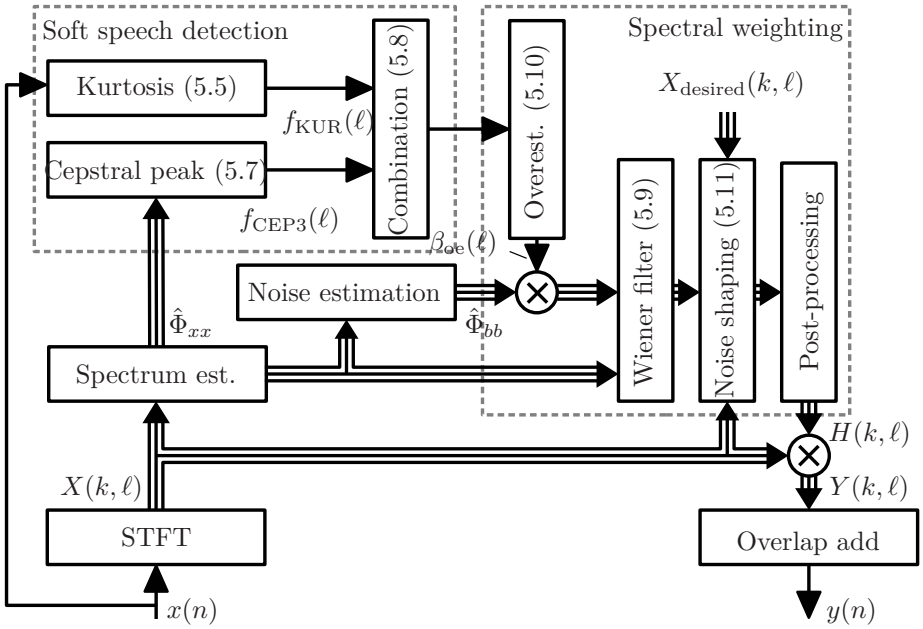


Figure 5.3: Structure of the full babble noise suppression system including speech detection and noise suppression. Some connections are neglected for simplicity.

ness [Hänsler and Schmidt, 2004]. For this, the noise power spectrum $\hat{\Phi}_{bb}(k,\ell)$ estimated according to Eq. (2.15) is scaled in the filter with an overestimation factor $\beta_{\mathrm{overest}}(\ell) \geq 1$ such that the resulting filter weights

$$H_{\mathrm{overest}}(k,\ell) = \max\left(H_{\mathrm{floor}}(k,\ell), 1 - \frac{\beta_{\mathrm{overest}}(\ell)\cdot\hat{\Phi}_{bb}(k,\ell)}{\hat{\Phi}_{xx}(k,\ell)}\right) \tag{5.9}$$

are biased towards lower values. A stronger attenuation of noise is achieved compared to the results with $\beta_{\mathrm{overest}}(\ell) = 1$ that corresponds to the original Wiener filter. The attenuation is limited by a spectral floor $H_{\mathrm{floor}}(k,\ell)$ to prevent unnaturally sounding holes in the output spectrum.

Setting the overestimation factor to a fixed value helps reducing the babble noise. However, this aggressive approach might also cause distortions of the desired speech. Lowering the factor during presence of foreground speech is therefore desirable. The combined feature can be mapped to a dynamic overestimation factor

$$\beta_{\mathrm{overest}}(\ell) = \min\left(\beta_{\max}, 1 + \frac{1}{c_{\mathrm{babble}}(\ell) + \epsilon}\right) \tag{5.10}$$

that approaches one for high values of the combined feature and $\beta_{\max} = 21$ in case that the feature does not indicate presence of desired speech. A small regularization constant ϵ is added to the feature value to prevent from division by zero. As shown in Figure 5.4, a more aggressive attenuation of the babble noise is achieved using dynamic noise overestimation

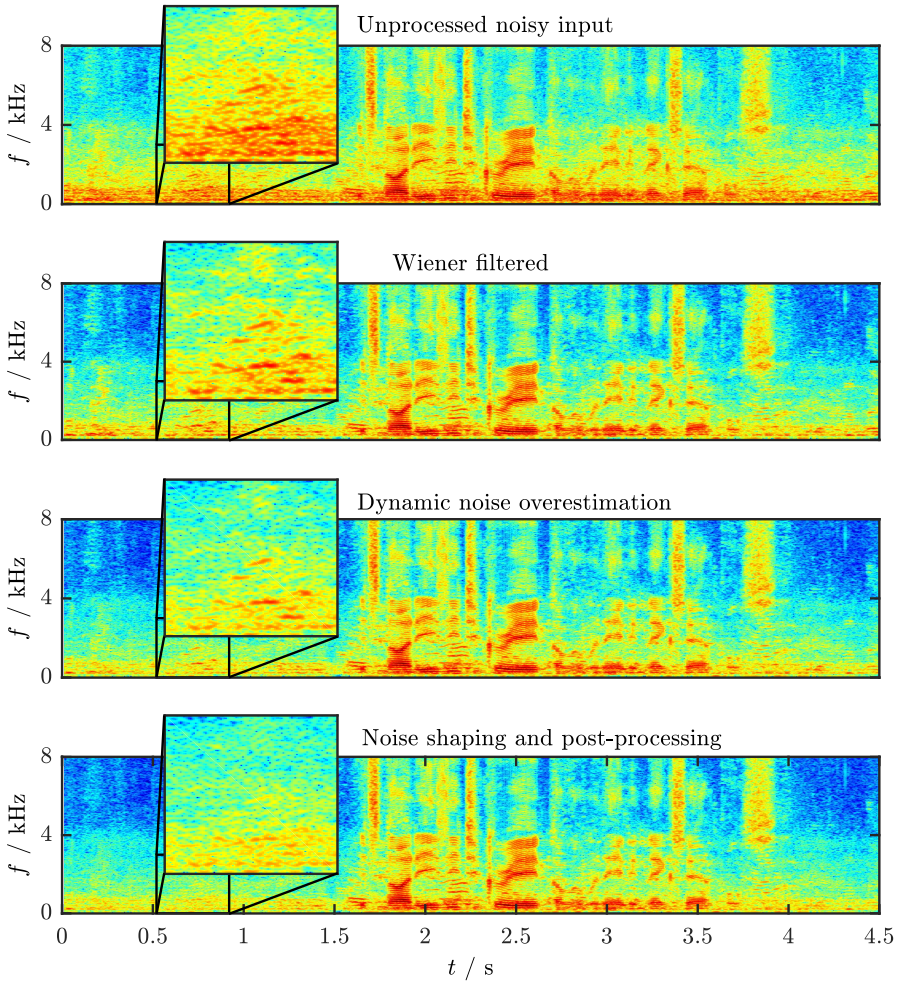


Figure 5.4: Spectra after different processing stages: the unprocessed signal is corrupted with partially non-stationary babble noise. The Wiener filter removes the stationary parts, however, the non-stationary interferences remain. These parts are reduced by additional stages incorporating dynamic noise overestimation, noise shaping, and a post-processing that dynamically remove the noise more aggressively.

compared to the Wiener filter without overestimation. Speech distortions are avoided since the overestimation factor is reduced dynamically during presence of the foreground speech.

The spectral floor controls the maximum attenuation that can be applied to noise. Choosing a fixed value $H_{\text{floor,fixed}}$, noise is lowered by a constant factor. As an effect, babble noise is attenuated, however, the spectro-temporal characteristic remains unchanged: the spectrum is still highly fluctuating with non-stationary components that stick out of the stationary background noise.

Noise shaping [Rajan et al., 2014] can be applied to reduce these fluctuations of the noise. The spectral floor is modified such that an extra attenuation is applied to non-stationary components. The floor is calculated dynamically

$$H_{\text{floor}}(k,\ell) = H_{\text{floor,fixed}} \cdot \min\left(1, \frac{X_{\text{desired}}(k,\ell)}{|X(k,\ell)|}\right)^{c} \tag{5.11}$$

by comparing the currently observed spectral magnitude $|X(k,\ell)|$ with a desired magnitude value $X_{\text{desired}}(k,\ell)$. For an exponent $c = 1$, the approach levels out any fluctuation of noise in the output signal. In contrast, the fixed floor without any noise shaping is applied for $c = 0$. As trade-off, $c = 0.5$ is chosen for the application in this thesis which reduces the non-stationarity but preserves natural fluctuations of the background noise. The desired magnitude is chosen as a slowly temporally smoothed version $X_{\text{desired}}(k,\ell) = \sqrt{\overline{\hat{\Phi}_{bb}}(k,\ell)}$ of the background noise estimate.

In a final post-processing stage, more filter weights may be forced to the dynamic floor. Filter weights are identified

$$I(k,\ell) = \begin{cases} 1 & \text{if } H_{\text{overest}}(k,\ell) > H_{\text{floor,fixed}} \\ 0 & \text{else} \end{cases} \tag{5.12}$$

that were not already set to the floor by Eq. (5.9). The respective bins are set to the dynamic floor in the final filter weights

$$H_{\text{final}}(k,\ell) = \begin{cases} H_{\text{floor}}(k,\ell) & \text{if } I(k,\ell) = 1 \text{ and } \overline{I}(k,\ell) < 0.4 \\ H_{\text{overest}}(k,\ell) & \text{else} \end{cases} \tag{5.13}$$

in case that the majority of neighboring frequency bins

$$\overline{I}(k,\ell) = \sum_{\tilde{k}=-K_{\text{babble}}}^{K_{\text{babble}}} I(k+\tilde{k},\ell) \cdot \frac{K_{\text{babble}} - |\tilde{k}|}{K_{\text{babble}}^2} \tag{5.14}$$

is attenuated to the floor. Here, a triangular window function with $K_{\text{babble}} = 20$ is applied to aggregate the results of the neighboring frequency bins[1].

[1]Bins outside the valid frequency range are taken into account using an expansion

$$I(k,\ell) = \begin{cases} 2I(0,\ell) - I(-k,\ell) & \text{for } k < 0 \\ 2I(N-1,\ell) - I(2N-2-k,\ell) & \text{for } k \geq N \end{cases} \tag{5.15}$$

of the indicator function corresponding to the implementation in MATLAB®'s `filtfilt` function.

Applying noise shaping and the post-processing, a more stationary noise spectrum with less non-stationary artifacts is achieved as shown in Figure 5.4. In the following section, the performance of the speech detector will be investigated. Further, the effect of the different modifications on the remaining noise is investigated by means of a subjective listening test.

5.1.2 Evaluation

The evaluation of the babble noise suppression system is split into two parts: first the speech detector is analyzed by means of ROC curves isolated from other components. Afterwards, a subjective listening test and the results are discussed that target on the complete system's performance.

For evaluation of the speech detector, the audio data was artificially mixed based on two databases. Again, speech recordings were randomly chosen from the TIMIT database [Garofolo et al., 1993] and the babble noise was excerpted from the NOISEX-92 corpus [Varga and Steeneken, 1993]. The SNR was varied between 0 dB and 10 dB to simulate a realistic range of noise conditions. Overall, 3.8 hours audio data were considered for this experiment. ROC curves of different features and combinations are summarized in Figure 5.5.

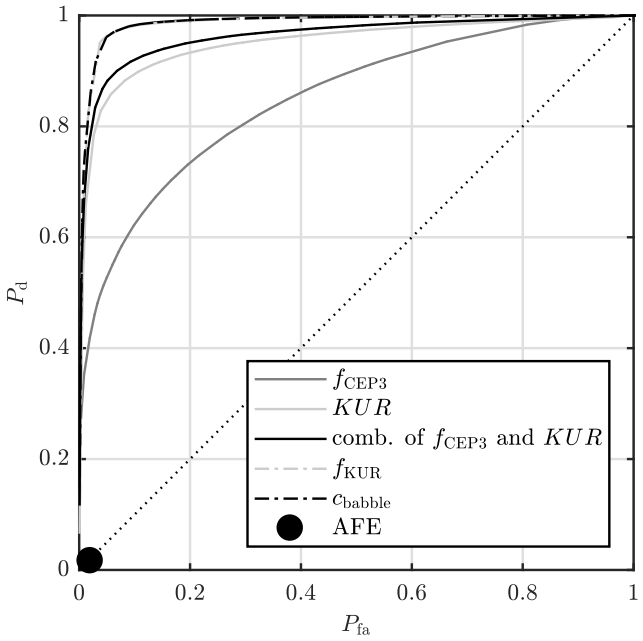


Figure 5.5: ROC curves of features for babble noise suppression: kurtosis alone already shows a reasonable performance that can be further improved by additionally considering a voicing feature f_CEP3 for a combination c_babble. The scenario is particularly challenging such that the standardized detector ETSI-AFE does not detect any speech.

As a baseline, the standardized VAD according to ETSI-AFE [ETSI, 2007] was applied.

This approach missed most of the desired speech in the experiment, however, it was also not erroneously triggered by babble noise. The unfavorable operating point in the lower-left corner of the ROC underlines the challenge of speech detection during presence of babble noise.

The plain kurtosis according to Eq. (5.2) without additional smoothing already shows a reasonable performance that can be further improved by means of different mechanisms. Considering the cepstral maximum according to Eq. (5.7), primarily voiced speech portions are detected. This feature alone hence performs worse than the kurtosis. Combining both features, the performance can be slightly improved.

Smoothing according to Eq. (5.5) and combining kurtosis with a cepstral maximum according to Eq. (5.8) both clearly improve the performance. The ROC of the final combination is close to the upper-left corner indicating a good detection performance. For this reason, the combination is adopted for the complete system that will be evaluated in the following.

A subjective listening test similar to "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)" [ITU, 2015] was conducted to assess the system's performance. This type of test allows for comparing multiple audio examples at once relative to a specified reference. The effect of multiple processing stages as perceived by human listeners can be investigated. For the discussed system, the test targeted on the acceptability of the remaining noise after the suppression of babble noise.

Before starting the actual test, the participants were encouraged to listen to and to familiarize themselves with all the audio examples that occurred later in the test. This way, the variety of processing results was assessable already before the first rating such that there was no need for establishing the scale of rating during the test.

In each part of the test, four audio examples were compared mutually and with a reference example as shown in Figure 5.6. The four randomly ordered audio examples were based on different processing variants: the unprocessed signal, Wiener-filtered signals without overestimation and with dynamic overestimation, as well as an example with noise shaping and post-processing. The unprocessed signal was always presented as reference.

The participants were asked to rate the acceptability of the remaining background noise in the four audio examples. Using sliders, the noise should be rated as more, less, or equally pleasant compared to the noise in the reference. The underlying numeric scale between -10 and 10 allowed for a fine-grained rating considering even subtle differences between the examples.

In the test, 10 different examples were considered with one example occurring twice for checking the consistency of the ratings. These examples included artificially mixed signals as discussed before, as well as real speech recordings taken in a crowded restaurant.

The group of test subjects consisted of 21 persons (male and female) with a majority having experience in audio signal processing. The rating results are summarized in Figure 5.7. The box plots show the median as well as the 25% and 75% quartiles of ratings for the four processing variants.

Almost all participants correctly rated the unprocessed signal as "equal" to the reference. This hidden reference apparently differed from the other variants that incorporated

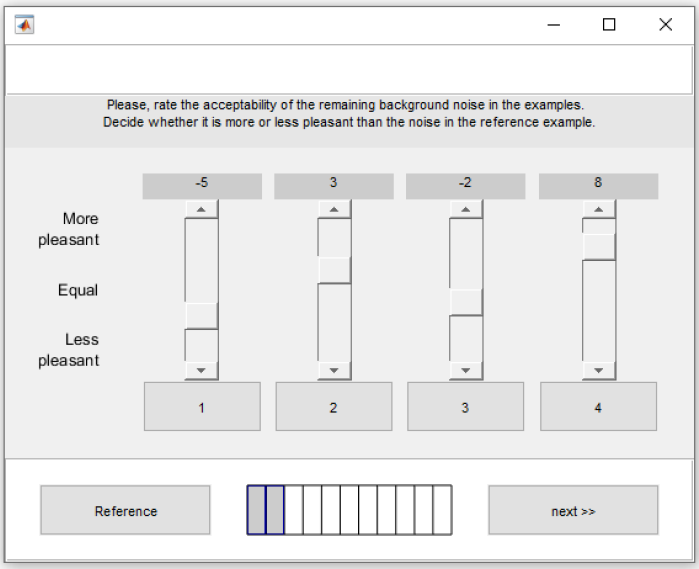


Figure 5.6: Subjective listening test: the participants were asked to rate the acceptability of the background noise for four processing variants mutually and relative to the noisy reference.

noise suppression. In most cases, the Wiener-filtered signal without noise overestimation was preferred to the unprocessed signal. However, some subjects were bothered by the remaining artifacts and rated the processed signal as "less pleasant". Dynamic noise overestimation clearly increased the acceptability of the background noise whereas only small additional improvements were achieved using noise shaping and post-processing.

The subjective listening test intentionally focused on the perceived acceptability of the background noise that can hardly be quantified by means of objective measures. In contrast, the influence of the signal processing on the desired speech signal was not explicitly addressed. These speech distortions can easily be measured using an objective criterion which was hence favored over a second time-consuming subjective test.

For the objective test, the noise suppression algorithm was applied to the mixed speech signal. The filter weights were stored during processing and applied to the clean speech components afterwards. Using this white box model [Steinert et al., 2009], the distorted speech $\tilde{s}(n)$ after noise suppression can be measured and compared with the original clean speech signal $s(n)$. The respective measure is based on a power ratio

$$\sigma^2_{\mathrm{DSR}} = \frac{\sum_n (s(n) - \tilde{s}(n))^2}{\sum_n s^2(n)} \tag{5.16}$$

between speech distortions and the undistorted speech signal. Noise components should not be addressed by this test. Therefore, the speech and noise signals have to be available separately. Since real recordings capture only the noisy speech but not the individual

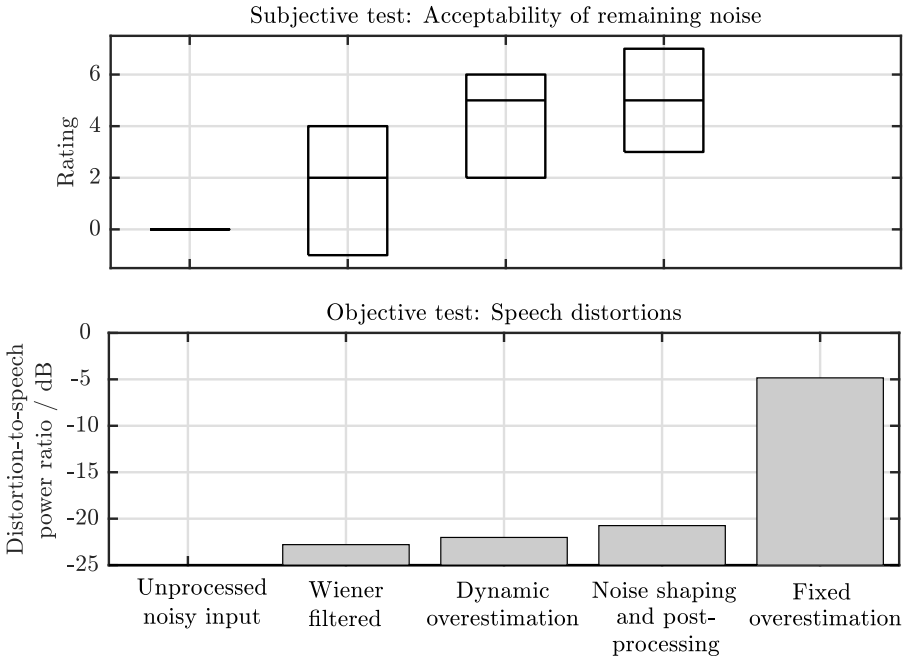


Figure 5.7: Subjective and objective evaluation of babble noise suppression: the remaining noise for different processing variants was rated by human listeners on a scale from less pleasant (-10) to more pleasant (10) compared to the unprocessed noisy input. Complementary, an objective measure was employed to assess the speech distortions introduced by the processing. For comparison, the distortions introduced by a fixed overestimation by far exceed the distortions of the VAD-based system.

components, this analysis again relies on artificially mixed signals based on TIMIT and NOISEX data.

The results are shown in Figure 5.7 together with the results that were obtained by the subjective test. Obviously, the unprocessed signal is completely undistorted, corresponding to a distortion-to-speech ratio of zero ($\hat{=} -\infty\,\mathrm{dB}$). Even though the noise suppression system was designed to act less aggressively during presence of desired speech, speech distortions cannot be perfectly avoided. The basic Wiener filter without overestimation already causes a ratio in the range of -23 dB. The succeeding processing steps, dynamic overestimation and post-processing, each introduce additional distortions in the range of 1 dB.

These speech distortions appear acceptable, given the significant improvements that were observed in the subjective listening test. The dynamically controlled overestimation indeed prevents speech distortions: using a fixed overestimation $\beta_{\mathrm{overest}}(\ell) = \beta_{\max}$, much higher speech distortions in the range of -5 dB are measured.

5.2 Speech detection for in-car-communication systems

Conversations between passengers in a car can be exhausting when their voices are masked by loud background noise. The occupants either have to lean towards each other or shout against the noise which is both inconvenient. In this situation, ICC systems may support the communication as discussed in Section 2.2.2: the passenger's speech is reinforced via loudspeakers close to the listeners.

To ensure a consistently good audio impression over a wide range of noise conditions and for varying speakers, different signal processing techniques are applied. Many of them make use of VAD, however, the exact requirements on the speech detector vary for the different algorithms:

- A high speech detection rate is essential for *noise estimation* as no speech shall be incorporated in the noise estimate. Otherwise, noise suppression might affect desired speech components that were accidentally attributed to noise components. Pausing the estimate's adaptation during presence of speech prevents from these speech distortions. Contrariwise, there are moderate restrictions regarding a low false-alarm rate: the noise estimate gets adapted slowly and it is generally not capable of tracking non-stationary noise components. Pausing the estimation for non-stationary noise components is hence acceptable.

- In contrast, for *AGC* a low false-alarm rate is desirable. The AGC estimates the speech power level and controls the ICC gain such that a desired output level of the speech signal is achieved. Interferences should not accidentally influence the gain which necessitates a robust VAD. On the other hand, the requirements on the detection rate may be relaxed: missing parts of utterances can be tolerated as long as the speech level can still be estimated based on the remaining speech portions that were captured.

- In an ICC system that supports multiple speakers, *dynamic audio routing* is needed to facilitate communication between speakers and listeners. During a conversation, the occupants alternately act as speakers and listeners, so the system has to continuously monitor the situation and enable only the processing unit dedicated to the current speaker. Other units shall be muted to prevent cross-talk between opposite processing units. The dynamic selection of the relevant processing direction requires a spatial detection of the current speaker. A power-based speaker activity detection (SAD) as introduced by Matheja et al. [2013] can be adopted for this spatial detection. Combining SAD and VAD may be beneficial to prevent accidental switching in case of interfering noise. A fast detection of speech onsets is crucial as the mechanism directly affects the signal path. Missed onsets are likely cut off which causes audible distortions. A low delay VAD is therefore needed to prevent from these artifacts.

These different requirements have to be considered when designing a VAD for a specific algorithm. An appropriate feature combination must be chosen that matches the constraints. The performance measures discussed in Section 4.2 help finding such a combination: complementary aspects regarding the detection performance can be quantified and compared to the algorithm's specification.

In the following, different combinations of features are introduced that all take into account the special conditions of ICC systems. The feature combinations are then evaluated subject to the algorithm's requirements. By means of ROC curves as well as the dynamic measures, the feasibility of the combinations for the different algorithms is determined.

5.2.1 Feature combination

The algorithms in context of ICC systems are typically constrained by short frames to ensure low processing delays as discussed in Section 2.2.2. This challenging condition particularly affects signal processing algorithms in the frequency domain. Given short frames, the spectral resolution is low and hence the spectral fine-structure is inaccessible. This has to be taken into account for VAD features.

In this thesis, multiple features suitable for ICC applications have been introduced and discussed. These features do not rely on a high spectral resolution but are capable of dealing with the limited frequency resolution due to short frames. In particular, the following features are promising candidates for feature combinations that will be the subject matter of this section:

- For the long-term signal variability feature f_{LTSV} according to Eq. (3.60), first the spectrum's entropy over time is determined before the variance is applied for fusing the results of different frequency bins. A high spectral resolution is not required during the latter step so the method can directly be applied for ICC applications. The analyses in Section 4.3.4 revealed a good detection performance for many scenarios such that the feature will be applied as baseline in this section.

- Temporal modulation as discussed in Section 3.3.2 does not rely on a high spectral resolution. Features based on this property can hence be applied in ICC systems without restrictions. Modulation focuses on fluctuations of the power spectrum with a particular frequency that is characteristic for speech. The feature is therefore expected to be less sensitive against interferences. Here, an implementation according to Eq. (3.73) is considered that measures the magnitude of modulations around 4 Hz.

- The feature f_{MPD} introduced in Section 3.3.3 also relies on the modulation but it takes into account the phase term in addition to the magnitude. The spectrogram is aggregated to wider bands before phase differences of the modulation between the bands are measured. Alternating patterns of voiced speech in lower frequencies and high frequency fricatives are expected. These patterns appear to be characteristic for human speech whereas they barely occur in noise which makes them robust against

most interferences. Making use of this feature together with others can improve the combination's robustness [Graf et al., 2016b].

- Pitch detection usually requires long frame lengths or a high spectral resolution which conflicts with the basic conditions of ICC systems. In Section 3.2.4, a pitch detection technique was introduced that overcomes this problem by taking multiple low resolution spectra into account. Using the feature f_{LPCS}, it is possible to consider the harmonic structure of voiced speech for VAD in an ICC system. Since large parts of speech are voiced, pitch detection is usually capable of detecting speech quickly while capturing even speech portions that otherwise might be missed.

In the following, different combinations of these features are introduced, each aiming at the particular requirements of an algorithm in ICC context:

- For noise estimation, a high speech detection rate is desirable. Features such as f_{LTSV} that reflect non-stationary components in the signal appear promising in this case. A combination with modulation-based features is conceivable, however, since the robustness against interferences is secondary for this application, already the single feature

$$c_{\mathrm{noiseest}}(\ell) = f_{\mathrm{LTSV}}(\ell) \tag{5.17}$$

  is expected to be sufficient. By choosing a low threshold, a sensitive detector can be designed that captures most of the speech.

- In contrast, for AGC, a higher robustness against non-stationary interferences is requested. The feature f_{MPD} can contribute to the combination's robustness. As it is rarely triggered by interferences, modulation phase difference (MPD) can be applied as a temporal mask that prevents from false-alarms. Calculating the product

$$c_{\mathrm{AGC}}(\ell) = f_{\mathrm{MPD}}(\ell) \cdot MOD(\ell) \tag{5.18}$$

  with 4 Hz modulation focuses the combination to frames that almost certainly contain speech. The long hangover in MPD coarsely selects intervals that certainly contain speech. The modulation is then used to subdivide these long intervals into the actual speech sections. Using this combination, some speech portions may be missed. The detection of speech is hence not optimal, which can be tolerated for the AGC algorithm.

- Finally, algorithms for dynamic audio routing require a VAD that detects speech onsets almost instantaneously. Considering only non-stationary components complies with this requirement. However, it also increases the vulnerability against interferences. For the feature combination discussed here, the harmonic structure of speech is hence considered. The pitch reflects an important property of speech that captures most phones including vowels and voiced fricatives. On the other hand, the feature is

robust against non-stationary interferences that typically do not exhibit a harmonic structure in the frequency range of human speech. Using

$$c_{\text{routing}}(\ell) = \max\left(c_{\text{AGC}}(\ell), f_{\text{LPCS}}(\ell) - \gamma_{\text{LPCS}}\right), \tag{5.19}$$

the advantages of the robust combination discussed for AGC and the pitch information can be combined. The combination indicates speech when at least one of both parts assumes a high value. To align the contributions of both features and to move the combination's operating point closer to the optimum, an offset between the features is compensated. During preliminary analyses, a value $\gamma_{\text{LPCS}} = 0.4$ was found to be appropriate. The resulting combination is expected to detect speech onsets faster but still to be robust against non-stationary interferences.

In the subsequent evaluation, the discussed feature combinations will be analyzed subject to their applicability for the respective algorithms in an ICC system.

5.2.2 Evaluation

Different aspects may be decisive for the applicability of a VAD depending on the algorithm at hand. This includes complementary criteria such as a high speech detection rate, a low false-alarm rate, or a quick detection of speech onsets. By means of the corresponding measures discussed in Section 4.2, the proposed feature combinations will now be analyzed.

First, the performance is assessed in terms of ROC curves before temporal properties of the feature combinations are considered. Taking into account multiple criteria finally helps judging the applicability of detectors for certain speech enhancement algorithms.

Two databases are employed for the following analyses: again, speech data is taken from TIMIT [Garofolo et al., 1993] but now the speech is artificially mixed with noise data based on the UTD-CAR-NOISE corpus [Krishnamurthy and Hansen, 2013]. The latter database contains noise recordings captured in diverse vehicles. Primarily stationary driving noise as well as some non-stationary interferences such as indicator clicking or the horn cover many scenarios that are relevant for an ICC system. For the simulation, an SNR of about -10 dB is chosen. As discussed in Section 4.1.3, human listeners perceive this low value for low-frequency automotive noise as being similar to 6 dB SNR for white noise.

The noise database also includes example speech sequences addressing digits between zero and nine. Based on these digits, the detection performance for short utterances can be analyzed. The lack of temporal context information usually complicates VAD as it increases the risk of missing those speech components. Including short utterances in the analyses hence may provide more detailed insights into the benefits and drawbacks of different approaches.

The basic conditions of a typical ICC system are taken into account for the evaluation. For a sample rate $f_{\text{s}} = 16\,\text{kHz}$, short frames of length $N = 128$ samples with a shift of $R = 32$ samples between consecutive frames are assumed. This results in very short frames of 8 ms and a temporal resolution of the spectrogram of 2 ms similar to [Franzen and Fingscheidt, 2017] that fulfill the low delay requirement. The system is analyzed in

the open loop while disregarding feedback components that might occur in a closed loop system.

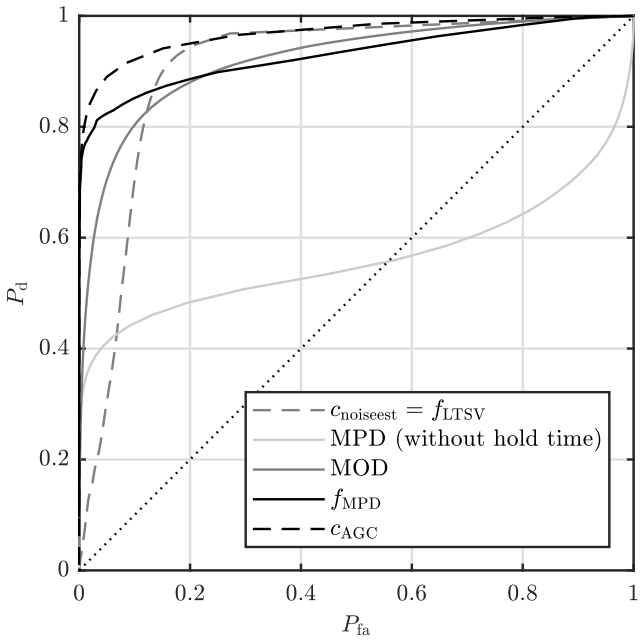


Figure 5.8: ROC curve for the MPD feature f_MPD, f_LTSV serves as a baseline.

In a first analysis, different features and the feature combination c_AGC are compared. The ROC curves in Figure 5.8 give an impression of the performance: the feature f_LTSV reflecting non-stationary components in the signal reaches a high detection rate. However, also a high false-alarm rate has to be expected for this feature. For false-alarm rates lower than 0.2, the detection rate decreases rapidly. So the LTSV feature can be applied for algorithms such as noise estimation that require a high detection rate but that can tolerate some false-alarms.

The curves for features based on MPD show a different trend: even for very low false-alarm rates, good detection rates are achieved but then for higher false-alarms, the detection rate stays below the curve of LTSV. This observation highlights the robustness of these features but it also reveals their drawback: they are rarely triggered by interferences but on the other hand some speech portions will be missed when relying solely on MPD.

In Figure 5.9, this observation is analyzed in more detail. By considering spoken digits, the detection performance for short utterances is assessed. The detection rate of various features disaggregated by different digits is shown for a fixed and very low false-alarm rate $P_\mathrm{fa} = 0.001$. LTSV almost never detects speech when such a low false-alarm rate is desired. The modulation feature takes into account a more specific property of speech and hence reaches higher detection rates that are similar for all digits. Since MPD relies on a very specific pattern of alternating voiced and unvoiced phones, it is more robust

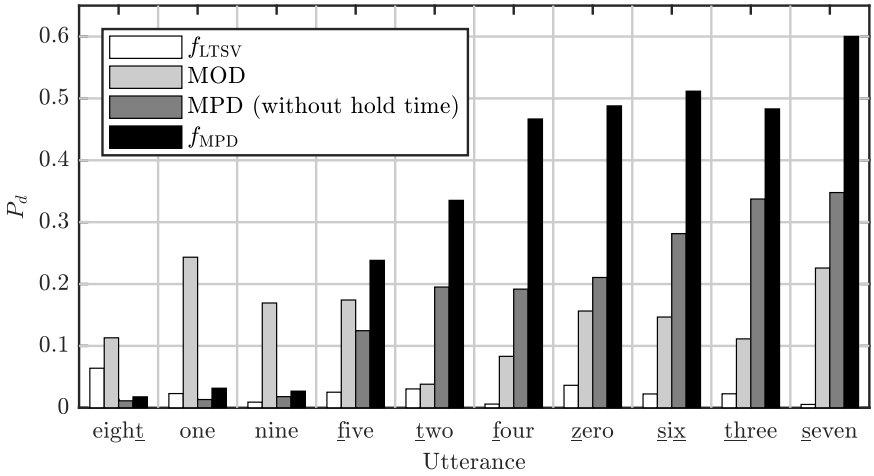


Figure 5.9: Comparison of MPD feature f_{MPD} with stationarity-based f_{LTSV} and modulation-based features for a detector targeting on a very low false-alarm rate $P_{\mathrm{fa}} = 0.001$.

against interferences but it does not perform equally well for all digits. For example the spoken digit "one" does not contain any unvoiced fricative and hence does not trigger the feature. On the other hand, "six" perfectly matches the expected pattern as it starts with a fricative, has a vowel in the middle, and ends again with a fricative. Consequently, MPD is capable of detecting the latter short utterance. MPD may fully leverage its strengths for longer utterances as supposed by AGC where the expected pattern is usually observable. A VAD in this case may benefit from the robustness of MPD.

Using a combination c_{AGC} based on MPD and a modulation feature results in a good compromise as shown again in Figure 5.8: MPD robustly detects portions of speech that are extended to long time intervals by means of a hangover mechanism for f_{MPD} in Eq. (3.76). These long intervals get shrunk back to the actual speech based on the modulation. Depending on the chosen detection threshold, low false-alarm rates or high detection rates comparable to LTSV can be achieved. A detector for AGC can be designed with an operating point close to the optimum position in the upper left corner.

The long-term features MPD and modulation both focus on sequences of phones. A first phone marking the beginning of an utterance is not sufficient to trigger the features as they rely on alternations of multiple phones. Speech onsets are hence likely missed which causes a delayed detection. For dynamically controlling the audio routing, a combination with another feature is desirable to better capture speech onsets. Power-based features could improve the detection delay, however, they tend to be vulnerable against interferences. Pitch-based features are more suitable since the harmonic structure can be detected quickly and segregates speech well from most interferences.

The ROC curves in Figure 5.10 show that the pitch feature is robust against most interferences. Even for a very low false-alarm rate, the detection rate reaches 0.5 but then it increases slowly. The curve lies far below the combination c_{AGC} as only voiced phones

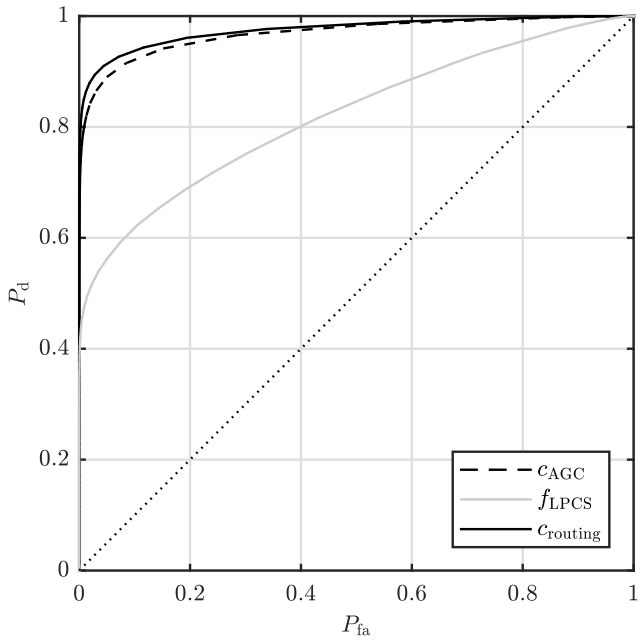


Figure 5.10: Combining c_AGC with a pitch feature f_LPCS slightly improved the result c_routing in terms of ROC curve.

are detected whereas the feature inherently misses unvoiced phones. The ROC curve shows only small improvements when combining both c_AGC and f_LPCS. Most speech components have already been captured by c_AGC hence the additional detections due to pitch are only marginal compared to the total number of detections for long utterances.

Speech onsets, however, are better detected as shown in the temporal analysis in Figure 5.11. The feature combination c_AGC introduces a high detection delay. A detection rate of 0.5 is reached after about 0.3 s but even after this it increases only slowly. For comparison, the pitch feature reaches a detection rate of 0.5 after 0.15 s which is already the average detection rate. In this analysis, the combination's advantage becomes more obvious: c_routing adopts the lower delay from the pitch feature but then it reaches a high average detection rate similar to the previous combination. The combination is hence suitable for algorithms such as dynamic audio routing that require a fast detection but also a high robustness against interferences.

Evaluations as discussed here provide insights into different aspects of the features' performance. Based on complementary measures, configurations can be identified that fulfill particular requirements imposed by algorithms and applications. Depending on the application, some features can be preselected as being generally suitable. For the considered ICC application, this narrowed the options to features that can cope with short frames. Other features that do not comply with the general framework can be discarded immediately. Also the choice of audio data for evaluation depends on the final application: typical

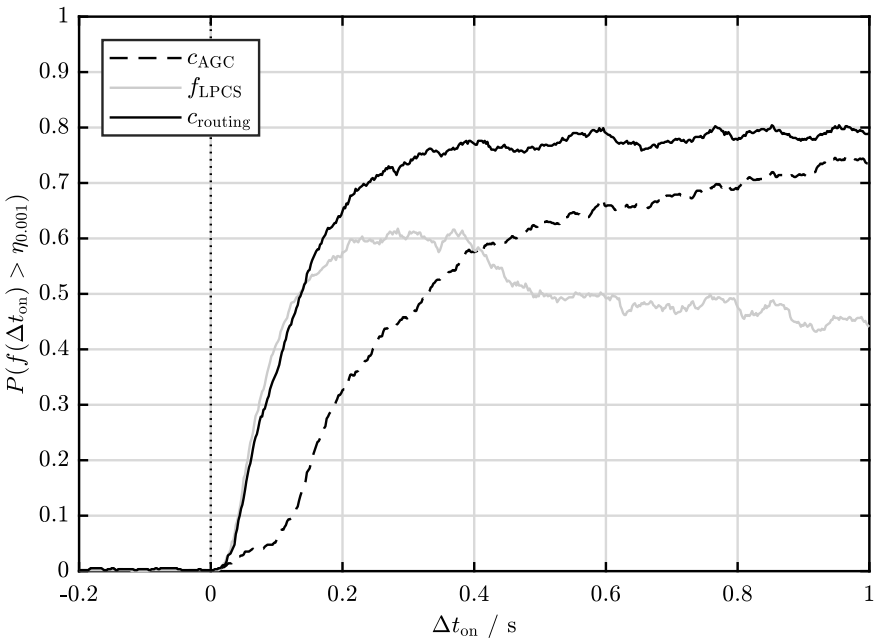


Figure 5.11: The onset plot (for $P_{\mathrm{fa}} = 0.001$) shows a swift reaction of c_{routing} due to the combination of c_{AGC} with f_{LPCS}.

use cases have to be covered such that the analyses reflect the performance realistically. In the evaluation, specific needs of an algorithm regarding the VAD can be taken into account. Some require a low false-alarm rate or a high detection rate. For others, the speed of detection is more important. Using corresponding measures, features or combinations thereof can be chosen that match the algorithm's requirements. Features such as f_{MPD} and f_{LPCS} were found to be capable of contributing to accurate VAD results in an ICC system as exemplified with different heuristic combinations. More advanced combinations, e.g., based on machine learning techniques are conceivable. Also these combinations and even complete speech detectors can be evaluated analogously using the same methods of analysis.

Chapter 6

Conclusion and Outlook

Several decades of research in the field of speech signal processing and particularly of VAD opened up a continuously growing diversity of use cases for speech-driven applications. Users of early speech applications still had to be content with controllable environments sealed off from noise. The first telephony users for example were bound to silent spaces while only science-fiction authors dreamed from voice-controlled computers. In those days, the requirements on signal processing were low thanks to the nearly optimal recording conditions with only little background noise. However, the situation has undergone a fundamental change in the meantime.

6.1 Summary and conclusion

Nowadays, speech applications can be found in almost any situation in daily life: at home, voice-controlled smart speakers provide access to information or allow controlling devices in the home environment. In automotive context, passengers and particularly the driver may utilize speech-driven applications to communicate with conversational partners at the far-end via hands-free telephony or to interact with the vehicle while staying focused on the road. In recent years, ICC systems facilitating communication among passengers in the same car become more and more popular.

Numerous algorithms in all the different speech-driven applications rely on VAD as an important component. Using VAD, speech distortions may be avoided in noise suppression algorithms by dynamically reducing the aggressiveness when desired speech is detected. Contrariwise, noise characteristics can be estimated more easily during absence of speech for time intervals where pure noise is assessable. Also dynamic audio routing for an ICC application may be controlled using VAD to enable only the relevant signal path dedicated to the current speaker.

Though the list of algorithms could go on and on, VAD always deals with the same fundamental question: the decision whether speech is present in an audio signal during a certain time interval or not. Various approaches have been introduced in literature to tackle this problem.

In this thesis, several different VAD approaches known from literature were summarized: they were grouped according to the underlying speech properties and compared in terms of their detection performances. Energy and SNR-based features turned out to be a reasonable basis for VAD in moderate noise conditions. Thanks to their low complexity, extracting these features is usually feasible even when the computational capacities of the target hardware are limited.

The harmonic structure of voiced speech corresponds to very specific repetitive signal portions that may be detected using features in time or in frequency domain. Detecting harmonic signal components has been proven in the experiments of this thesis to increase the robustness against many interferences that do not exhibit a harmonic structure. Standard techniques generally require long frame lengths to observe this signal property. This limitation impedes using well-known approaches for detection of voiced speech in low-latency applications such as ICC systems that operate with shorter frames. A new feature introduced with this work overcomes this restriction by considering multiple short frames jointly. This way, the effective frame length can be extended without introducing additional latency in the processed audio signal.

As a general conclusion of the present study, there is some evidence that extending the temporal context for feature extraction is beneficial with respect to the VAD's robustness. On a longer time scale, concatenation of different phones, vowels and consonants, induces a rhythm that is characteristic for human speech. Modulation-based features exploit this speech property by analyzing the temporal evolution of power. A new feature introduced with this thesis additionally targets on alternations of voiced and unvoiced speech portions. These features performed very well even for non-stationary noise conditions. This robustness, however, is gained at the expense of an increased reaction time of the detector and a higher risk of missing speech.

Even though several approaches showed promising detection results in the experiments of this thesis, no ultimate recommendation for a single approach can be made that equally fits for all target applications: depending on the specific algorithm, the requirements on VAD may differ considerably. Some algorithms need a VAD that is particularly robust against interferences, others depend on not missing speech, and for still others, a VAD is preferable that is triggered quickly after speech onsets. VAD hence has to be designed carefully with special attention to constraints imposed by the target application.

For this purpose, an evaluation measure introduced in this work allows for a granular temporal assessment of VAD approaches that may be utilized in addition to classical ROC curves. Whereas the latter only consider averaged detection results, the new measure provides insights into the transient behavior around speech onsets. Using this measure, a detector's reaction time after speech onsets can be quantified which allows for a better founded comparison of different candidates for VAD. This evaluation framework was finally applied for the selection of features and combinations thereof for two applications: a noise suppression system dedicated to the suppression of babble noise as well as multiple algorithms in the context of an ICC system.

For the detection of desired foreground speech during presence of babble noise, kurtosis turned out to be appropriate. This feature captures large parts of speech, however, low-

frequency vowels are sometimes missed. A harmonicity-based feature extracted from the cepstrum was therefore taken into account additionally. This combination was applied in the final babble noise suppression system. A subjective listening test as well as objective measures confirmed the benefits of a VAD-controlled noise suppression in the babble noise scenario.

Low-latency requirements in an ICC system affect almost any algorithm including the VAD. A low spectral resolution complicates extraction of features that rely on the spectral fine structure. Using the novel features that explicitly take this limitation into account, presence of voiced speech and a distinctive pattern of alternating voiced and unvoiced phones can be detected. Combinations of these and some established features were introduced and evaluated for three different algorithms with very different needs.

Though the features and evaluation criteria introduced with this thesis were motivated and exemplified by means of two quite specific applications, they can be easily adopted also for other use cases. The features are promising whenever a high robustness of VAD is targeted or when the application framework only provides low-resolution spectra. The evaluation measure is further implementation agnostic: since it purely relies on the comparison of detection results and a reference, it can be applied to any VAD even without knowledge of the detector's internals.

6.2 Outlook

Thinking ahead, the number of applications that make use of speech and voice control will even increase. Whereas currently the microphones are usually still placed close to the speakers in a rather shielded environment, e.g., mounted inside a noise-reduced car cabin above the occupants' heads, those conditions may no longer be present for future applications.

Smartphones are often used in an acoustically adverse pose already today: holding the device with an outstretched arm to face the display, e.g., for video telephony, increases the spatial distance between mouth and microphone. This situation is disadvantageous in two respects: on the one hand, the desired speech is not optimally captured while on the other hand, the device is excessively exposed to environmental influences. Similarly, microphones mounted on vehicles without a shielding cabin, such as motorbikes, are heavily impacted by wind and other interferences that mask large parts of the desired speech. As a consequence, an increased robustness of VAD against interferences will become even more crucial in future.

The contributions of this thesis explicitly target on some challenging scenarios. Novel features were introduced that are robust against many interferences but that at the same time can be extracted with relatively little effort. In foreseeable future, promising technical innovations in the field of deep learning will help tackling further challenges. Data-driven approaches that automatically find informative patterns, e.g., based on the plain spectrum or even the raw waveform [Zazo et al., 2016] and traditional features will complement each other. For applications running on platforms with limited processing power or memory,

however, smaller neural networks or other combination mechanisms that rely on sophisticated features will most likely remain the first choice for the time being. The compact structure of hearing aids, for example, sets a natural limit regarding the hardware dimensions including the battery which prohibits costly calculations. Condensing information on presence of speech in an early feature extraction stage further lowers the amount of data that is needed for neural-network training and reduces the training efforts in general.

The evaluation framework introduced with this work can be applied to any VAD irrespective of the underlying technology. Whenever a quick detection of onsets is crucial, e.g, for algorithms that directly affect the signal path, the transient behavior can help judging the detection performance. Including similar measures in the training of neural networks is conceivable to favor and reward a quick reaction of the final detection in future VAD approaches.

Bibliography

3GPP. 3gpp ts 26.194 : AMR wideband speech codec; voice activity detection (VAD). 2000.

Vipul Arora and Henning Reetz. Automatic speech recognition: What phonology can offer. *The Speech Processing Lexicon: Neurocognitive and Behavioural Approaches*, 22: 211, 2017.

Jörg-Hendrik Bach, Birger Kollmeier, and Jörn Anemüller. Modulation-based detection of speech in real background noise: Generalization to novel background classes. In *Proc. of ICASSP*, pages 41–44, Dallas, Texas, USA, 2010.

Adil Benyassine, Eyal Shlomot, Huan-yu Su, Dominique Massaloux, Claude Lamblin, and Jean-Pierre Petit. ITU-T Recommendation G.729 Annex B: A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications. *IEEE Communications Magazine*, pages 64–73, September 1997.

Francesco Beritelli, Salvatore Casale, and Alfredo Cavallaero. A robust voice activity detector for wireless communications using soft computing. *IEEE Journal on Selected Areas in Communications*, 16(9):1818–1829, 1998. ISSN 0733-8716.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007. ISBN 0-387-31073-8.

Michael Brodersen, Achim Volmer, and Gerhard Schmidt. Signal enhancement for communication systems used by fire fighters. *EURASIP Journal on Audio, Speech, and Music Processing*, 2019(1):21, December 2019. ISSN 1687-4722.

Ilja N. Bronštejn, Konstantin A. Semendjajew, Gerhard Musiol, and Heiner Mühlig. *Taschenbuch der Mathematik*. Verlag Harri Deutsch, 2008. ISBN 978-3-8171-2017-8.

Philipp Bulling. *Rückkopplungsunterdrückung für Innenraumkommunikationssysteme*. PhD thesis, Christian-Albrechts Universität Kiel, 2018.

Israel Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5): 466–475, 2003. ISSN 1063-6676.

Israel Cohen and Baruch Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Letters*, 9(1):12–15, 2002. ISSN 1070-9908.

James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

David Cournapeau and Tatsuya Kawahara. Evaluation of real-time voice activity detection based on high order statistics. In *Proc. of Interspeech*, pages 2945–2948, Antwerp, Belgium, 2007.

David Cournapeau, Tatsuya Kawahara, Kenji Mase, and Tomoji Toriyama. Voice activity detector based on enhanced cumulant of LPC residual and on-line EM algorithm. In *Proc. of Interspeech*, Pittsburgh, PA, USA, 2006.

David B. Dean, Sridha Sridharan, Robert J. Vogt, and Michael W. Mason. The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. In *Proc. of Interspeech*, Makuhari, Japan, 2010.

Nai Ding, Aniruddh D. Patel, Lin Chen, Henry Butler, Cheng Luo, and David Poeppel. Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, 81: 181 – 187, 2017. ISSN 0149-7634.

B Elan Dresher. The phoneme. *The Blackwell companion to phonology*, pages 1–26, 2011.

Erik Edwards and Edward F Chang. Syllabic ($\sim$ 2–5 Hz) and fluctuation ($\sim$ 1–10 Hz) ranges in speech and auditory processing. *Hearing research*, 305:113–134, 2013.

Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(6):1109–1121, 1984. ISSN 0096-3518.

Miquel Espi, Shigeki Miyabe, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. Analysis on speech characteristics for robust voice activity detection. In *Spoken Language Technology Workshop (SLT)*, pages 151 –156, Berkeley, California, USA, 2010.

ETSI. Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels. 1998.

ETSI. Advanced front-end feature extraction algorithm. 2007.

Tony Ezzat, Jake Bouvrie, and Tomaso Poggio. Spectro-temporal analysis of speech using 2-D Gabor filters. In *Proc. of Interspeech*, volume 7, pages 506–509, Antwerp, Belgium, 2007.

Gunnar Fant. *Acoustic theory of speech production*. Number 2. Walter de Gruyter, 1970.

Tom Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1), 2004.

Jan Franzen and Tim Fingscheidt. A delay-flexible stereo acoustic echo cancellation for DFT-based in-car communication (ICC) systems. In *Proc. of Interspeech*, pages 181–185, Stockholm, Sweden, 2017.

Jan Franzen, Inka Meyer zum Alten Borgloh, and Tim Fingscheidt. On the benefit of a stereo acoustic echo cancellation in an in-car communication system. In *Proc. of ITG Conference on Speech Communication*, pages 1–5, Oldenburg, Germany, 2018. VDE.

Daniel K. Freeman, C.B. Southcott, and I. Boyd. A voice activity detector for the Pan-European digital cellular mobile telephone service. In *IEE Colloquium on Digitized Speech Communication via Mobile Radio*, pages 6/1–6/5, London, United Kingdom, 1988.

Takashi Fukuda, Osamu Ichikawa, and Masafumi Nishimura. Improved voice activity detection using static harmonic features. In *Proc. of ICASSP*, pages 4482–4485, Dallas, Texas, USA, 2010.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallet, and Nancy L. Dahlgren. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM, 1993.

Timo Gerkmann and Richard C. Hendriks. Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1383–1393, 2012. ISSN 1558-7916.

Houman Ghaemmaghami, Brendan J. Baker, Robert J. Vogt, and Sridha Sridharan. Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. In *Proc. of Interspeech*, Makuhari, Japan, 2010.

Prasanta Kumar Ghosh, Andreas Tsiartas, and Shrikanth Narayanan. Robust voice activity detection using long-term signal variability. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):600–613, 2011.

Bradford W. Gillespie, Henrique S. Malvar, and Dinei AF Florêncio. Speech dereverberation via maximum-kurtosis subband adaptive filtering. In *Proc. of ICASSP*, Salt Lake City, Utah, USA, 2001.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Improved Performance Measures for Voice Activity Detection. In *Proc. of ITG Conference on Speech Communication*, Erlangen, Germany, September 2014.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015(91), November 2015a. ISSN 1687-6180.

Simon Graf, Anne Theiß, Tobias Herbig, and Gerhard Schmidt. Listening Test to Determine the Mismatch Between Signal-to-Noise Ratio and Human Perception. In *Proc. of DAGA*, Nürnberg, Germany, 2015b.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Kurtosis-Controlled Babble Noise Suppression. In *Proc. of ITG Conference on Speech Communication*, Paderborn, Germany, 2016a.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Voice Activity Detection Based on Modulation-Phase Differences. In *Proc. of ITG Conference on Speech Communication*, Paderborn, Germany, 2016b.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Low-Complexity Pitch Estimation Based on Phase Differences Between Low-Resolution Spectra. In *Proc. of Interspeech*, Stockholm, Sweden, August 2017a.

Simon Graf, Nabeel Zaidi, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Detection of Voiced Speech and Pitch Estimation for Applications with Low Spectral Resolution. In *Proc. of DAGA*, Kiel, Germany, 2017b.

J. A. Haigh and John S. Mason. A voice activity detector based on cepstral analysis. In *Proc. of Eurospeech*, Berlin, Germany, 1993.

Eberhard Hänsler. *Statistische Signale: Grundlagen und Anwendungen*. Springer Berlin Heidelberg, 2001. ISBN 9783540416449.

Eberhard Hänsler and Gerhard Schmidt. *Acoustic Echo and Noise Control: A Practical Approach*. John Wiley & Sons, 2004. ISBN 0-471-45346-3.

Kohei Hayashida, Makoto Nakayama, Takanobu Nishiura, Yukihiko Yamashita, T. K. Horiuchi, and Toshihiko Kato. Close/distant talker discrimination based on kurtosis of linear prediction residual signals. In *Proc. of ICASSP*, pages 2327–2331, Florence, Italy, 2014.

Daniel Hirst. Prosodic Aspects of Speech and Language. *Encyclopedia of Language & Linguistics (Second Edition)*, 2006.

John D. Hoyt and Harry Wechsler. Detection of human speech in structured noise. In *Proc. of ICASSP*, pages 237–240, Adelaide, Australia, 1994.

Chung-Chien Hsu, Tse-En Lin, Jian-Hueng Chen, and Tai-Shih Chi. Voice activity detection based on frequency modulation of harmonics. In *Proc. of ICASSP*, pages 6679–6683, Vancouver, Canada, 2013.

Guoning Hu and DeLiang Wang. Segregation of unvoiced speech from nonspeech interference. *The Journal of the Acoustical Society of America*, 124(2):1306–1319, August 2008. ISSN 00014966.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411 – 430, 2000. ISSN 0893-6080.

IPA. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. 1999.

Kentaro Ishizuka and Tomohiro Nakatani. Study of noise robust voice activity detection based on periodic component to aperiodic component ratio. In *Proc. of Statistical and Perceptual Audition (SAPA)*, pages 65–70, Pittsburgh, Pennsylvania, USA, 2006.

Dorota J. Iskra, Beate Grosskopf, Krzysztof Marasek, Henk van den Heuvel, Frank Diehl, and Andreas Kiessling. SPEECON-Speech Databases for Consumer Devices: Database Specification and Validation. In *LREC*, 2002.

ITU. ITU-T Recommendation P.56 (Objective measurement of active speech level), 1993.

ITU. ITU-T Recommendation G.729 Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming To Recommendation V.70. 1996.

ITU. ITU-T Recommendation G.729.1 Annex F: New annex F with voice activity detector using ITU-T G.720.1 annex A. 2012.

ITU. Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems. 2015.

ITU. ITU-T Recommendation P.1150: In-car communication audio specification. 2020.

Jean-Claude Junqua. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, 20(1):13–22, 1996.

Peter Kabal and Ravi P Ramachandran. The computation of line spectral frequencies using Chebyshev polynomials. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6):1419–1426, 1986.

Tomi Kinnunen, Evgenia Chernenko, Marko Tuononen, Pasi Fränti, and Haizhou Li. Voice activity detection using MFCC features and support vector machine. In *Proc. of Int. Conf. on Speech and Computer (SPECOM), Moscow, Russia*, pages 556–561, 2007.

Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4):320–327, 1976.

Ivan Kraljevski, Zheng-Hua Tan, and Maria Paola Bissiri. Comparison of Forced-Alignment Speech Recognition and Humans for Generating Reference VAD. In *Proc. of Interspeech*, pages 2937–2941, Dresden, Germany, September 2015.

Mohamed Krini and Gerhard Schmidt. Spectral refinement and its application to fundamental frequency estimation. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 251–254, 2007.

Mohamed Krini and Gerhard Schmidt. Method for temporal interpolation of short-term spectra and its application to adaptive system identification. In *Proc. of ICASSP*, pages 45–48, Kyoto, Japan, 2012.

Mohamed Krini and Gerhard Schmidt. Refinement and Temporal Interpolation of Short-Term Spectra: Theory and Applications. In Gerhard Schmidt, Huseyin Abut, Kazuya Takeda, and John H.L. Hansen, editors, *Smart Mobile In-Vehicle Systems: Next Generation Advancements*, pages 139–166. Springer New York, New York, NY, 2014. ISBN 978-1-4614-9120-0. DOI: 10.1007/978-1-4614-9120-0_9.

Nitish Krishnamurthy and John H. L. Hansen. Babble Noise: Modeling, Analysis, and Applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7): 1394–1407, September 2009. ISSN 1558-7916.

Nitish Krishnamurthy and John H. L. Hansen. Car noise verification and applications. *International Journal of Speech Technology*, December 2013. ISSN 1381-2416, 1572-8110.

Trausti Kristjansson, Sabine Deligne, and Peder Olsen. Voicing features for robust speech detection. In *Proc. of Interspeech*, pages 369–372, Lisbon, Portugal, 2005.

Jon Krohn, Grant Beyleveld, and Aglaé Bassens. *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. The Addison-Wesley data & analytics series. Addison Wesley, 2019. ISBN 9780135116692.

Frank Kurth and Alessia Cornaggia-Urrigshardt. Detection of Audio Events with Repetitive Structure Using Generalized Autocorrelations. In *Proc. of ITG Conference on Speech Communication*, Erlangen, Germany, September 2014.

Oh-Wook Kwon and Te-Won Lee. Optimizing speech/non-speech classifier design using adaboost. In *Proc. of ICASSP*, pages I–436, Hong Kong, 2003. IEEE.

Steve Lawrence, Ian Burns, Andrew Back, Ah Chung Tsoi, and C Lee Giles. Neural network classification and prior class probabilities. In *Neural networks: tricks of the trade*, pages 299–313. Springer, 1998.

Guoping Li and Mark E. Lutman. Sparseness and speech perception in noise. In *Proc. of Statistical and Perceptual Audition (SAPA)*, Pittsburgh, PA, USA, 2006.

Philipos C. Loizou. *Speech Enhancement: Theory and Practice, Second Edition*. CRC Press, 2 edition, April 2013. ISBN 1-4665-0421-8.

Christian Lüke, Halil Özer, Gerhard Schmidt, Anne Theiß, and Jochen Withopf. Signal processing for in-car communication systems. In *5th Biennial Workshop on DSP for In-Vehicle Systems*, Kiel, Germany, 2011.

Yanna Ma and Akinori Nishihara. Efficient voice activity detection algorithm using long-term spectral flatness measure. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–18, 2013.

Nilesh Madhu. Note on measures for spectral flatness. *Electronics letters*, 45(23):1195–1196, 2009.

Rainer Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504–512, 2001. ISSN 1063-6676.

Mark Marzinzik and Birger Kollmeier. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing*, 10(2):109 –118, February 2002. ISSN 1063-6676.

Timo Matheja, Markus Buck, and Tim Fingscheidt. Speaker activity detection for distributed microphone systems in cars. In *Proceedings of the 6th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems*. Seoul, 2013.

Stefan Meier and Walter Kellermann. Artificial Neural Network-Based Feature Combination for Spatial Voice Activity Detection. In *Proc. of Interspeech*, pages 2987–2991, San Francisco, USA, 2016.

Nima Mesgarani, Malcolm Slaney, and Shihab A. Shamma. Discrimination of speech from nonspeech based on multiscale spectro-temporal Modulations. *IEEE Transactions on Audio, Speech and Language Processing*, 14(3):920–930, May 2006. ISSN 1558-7916.

Parth Mishra, Serkan Tokgöz, and Issa MS Panahi. Efficient modeling of acoustic feedback path in hearing aids by voice activity detector-supervised multiple noise injections. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3549–3552. IEEE, 2018.

Asunción Moreno, Børge Lindberg, Christoph Draxler, Gael Richard, Khalid Choukri, Stephan Euler, and Jeffrey Allen. SPEECH DAT CAR. A Large Speech Database for Automotive Environments. *Proc. of LREC 2000*, 2000.

Douglas J. Nelson and Joseph Pencak. Pitch-based methods for speech detection and automatic frequency recovery. In *SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation*, pages 92–100, San-Diego, California, USA, 1995.

Elias Nemer, Rafik Goubran, and Samy Mahmoud. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*, 9(3):217–231, 2001.

Marco Orlandi, Alfiero Santarelli, and Daniele Falavigna. Maximum likelihood endpoint detection with time-domain features. In *Proc. of Interspeech*, Geneva, Switzerland, 2003.

Douglas O'Shaughnessy. *Speech Communications*. IEEE Press, 2000.

Chiyoun Park, Namhoon Kim, and Jeongmi Cho. Voice activity detection using partially observable Markov decision process. pages 2227–2230, Brighton, United Kingdom, 2009.

Joseph Pencak and Douglas Nelson. The NP Speech Activity Detection Algorithm. In *Proc. of ICASSP*, pages 381–384, Detroit, Michigan, USA, 1995.

Fabrice Plante, Georg F. Meyer, and William A. Ainsworth. A pitch extraction reference database. In *Proc. of Eurospeech*, Madrid, Spain, 1995.

Lawrence R. Rabiner and M. R. Sambur. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, 54(2):297–315, 1975.

Lawrence R. Rabiner and Marvin R. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(4):338–343, 1977.

Vasudev Kandade Rajan, Christin Baasch, Mohamed Krini, and Gerhard Schmidt. Improvement in Listener Comfort Through Noise Shaping Using a Modified Wiener Filter Approach. In *Proc. of ITG Conference on Speech Communication*, Erlangen, Germany, 2014.

Javier Ramírez, José C. Segura, Carmen Benítez, Ángel de la Torre, and Antonio Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42:271–287, 2004a.

Javier Ramírez, José Carlos Segura, Carmen Benítez, A. de La Torre, and A. Rubio. A new voice activity detector using subband order-statistics filters for robust speech recognition. In *Proc. of ICASSP*, page 849, Montreal, Canada, 2004b.

Javier Ramírez, Juan M. Górriz, and José C. Segura. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. In Michael Grimm and Kristian Kroschel, editors, *Robust Speech Recognition and Understanding*, pages 1–22. IntechOpen, 2007. ISBN 978-3-902613-08-0.

Seyed Omid Sadjadi and John H. L. Hansen. Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux. *IEEE Signal Processing Letters*, 20(3):197–200, March 2013. ISSN 1070-9908, 1558-2361.

Marc René Schädler, Bernd T. Meyer, and Birger Kollmeier. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *The Journal of the Acoustical Society of America*, 131:4134, 2012.

Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. of ICASSP*, pages 1331–1334, Munich, Germany, April 1997.

Gerhard Schmidt and Tim Haulick. Signal processing for in-car communication systems. *Signal processing*, 86(6):1307–1326, 2006.

Stephanie Seneff and Victor Zue. Transcription and alignment of the timit database. *TIMIT CD-ROM Documentation*, 1988.

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1 –3, January 1999. ISSN 1070-9908.

Kai Steinert, S Suhadi, and Tim Fingscheidt. A comparison of instrumental measures for wideband speech quality assessment of hands-free systems in echoic condition. In *Proc. of NAG/DAGA*, Rotterdam, Netherlands, 2009.

Falco Strasser and Henning Puder. Adaptive Feedback Cancellation for Realistic Hearing Aid Applications. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2322–2333, December 2015. ISSN 2329-9290, 2329-9304.

Zheng-Hua Tan and Børge Lindberg. High-accuracy, low-complexity voice activity detection based on a posteriori SNR weighted energy. In *Proc. of Interspeech*, pages 2231–2234, Brighton, UK, 2009.

Jürgen Tchorz and Birger Kollmeier. Speech detection and SNR prediction basing on amplitude modulation pattern recognition. In *Proc. of Eurospeech*, Budapest, Hungary, 1999.

José M. Tribolet, Peter Noll, Barbara J. McDermott, and Ronald E. Crochiere. A study of complexity and quality of speech waveform coders. In *Proc. of ICASSP*, pages 586–590, Tulsa, Oklahoma, USA, 1978.

Andreas Tsiartas, Theodora Chaspari, Nassos Katsamanis, Prasanta Ghosh, Ming Li, Maarten Van Segbroeck, Alexandros Potamianos, and Shrikanth S. Narayanan. Multi-band long-term signal variability features for robust voice activity detection. In *Proc. of Interspeech*, pages 718–722, Lyon, France, 2013.

R. Tucker. Voice activity detection using a periodicity measure. *Communications, Speech and Vision, IEE Proceedings I*, 139(4):377–380, 1992. ISSN 0956-3776.

Tohru Usukura and Wataru Mitsuhashi. Voice activity detection using AdaBoost with multi-frame information. In *2nd International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–8. IEEE, 2008.

Stefaan Van Gerven and Fei Xie. A comparative study of speech detection methods. In *Proc. of Eurospeech*, volume 97, Rhodos, Greece, 1997.

Toon Van Waterschoot and Marc Moonen. Fifty years of acoustic feedback control: State of the art and future challenges. *Proceedings of the IEEE*, 99(2):288–327, 2010.

Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993. ISSN 0167-6393.

Peter Vary and Rainer Martin. *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, 2006. ISBN 9780470031759.

Hadi Veisi and Hossein Sameti. Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement. *IET Signal Processing*, 6(1):54–63, 2012. ISSN 1751-9675.

Damjan Vlaj, Marko Kos, and Zdravko Kačič. Quick and efficient definition of hangbefore and hangover criteria for voice activity detection. In *International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–4, Bratislava, Slovakia, 2016.

Martin Vondrášek and Petr Pollák. Methods for Speech SNR Estimation: Evaluation Tool and Analysis of VAD Dependency. *Radioengineering*, 14(1), 2005.

DeLiang Wang and Guy J Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley, September 2006. ISBN 978-0-471-74109-1.

Jochen Withopf, Laila Jassoume, Gerhard Schmidt, and Anne Theiß. A modified overlap-add filter bank with reduced delay. In *Proc. DAGA 2012, Darmstadt, Germany*, 2012.

Mingyang Wu and DeLiang Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (3):774–784, May 2006. ISSN 1558-7916.

In-Chul Yoo, Hyeontaek Lim, and Dongsuk Yook. Formant-Based Robust Voice Activity Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23 (12):2238–2245, December 2015. ISSN 2329-9290, 2329-9304.

Huajun Yu and Tim Fingscheidt. Black box measurement of musical tones produced by noise reduction systems. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4573–4576. IEEE, 2012.

Ruben Zazo, Tara N. Sainath, Gabor Simko, and Carolina Parada. Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection. In *Proc. of Interspeech*, pages 3668–3672, San Francisco, USA, September 2016.

Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and Models*. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2013. ISBN 9783662095621.

Own Publications

Markus Buck, Tobias Herbig, Simon Graf, and Christophe Ris. Methods and apparatus for speech segmentation using multiple metadata. U.S. Patent Application US 000010229686 B2, International Patent Application WO 2016028254 A1, 2016.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Improved Performance Measures for Voice Activity Detection. In *Proc. of ITG Conference on Speech Communication*, Erlangen, Germany, September 2014.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015(91), November 2015a. ISSN 1687-6180.

Simon Graf, Anne Theiß, Tobias Herbig, and Gerhard Schmidt. Listening Test to Determine the Mismatch Between Signal-to-Noise Ratio and Human Perception. In *Proc. of DAGA*, Nürnberg, Germany, 2015b.

Simon Graf, Tobias Herbig, and Markus Buck. Babble noise suppression. U.S. Patent Application US 000010783899 B2, European Patent Application EP 000003411876 B1, International Patent Application WO 002017136018 A9, 2016a.

Simon Graf, Tobias Herbig, and Markus Buck. Voice activity detection feature based on modulation-phase differences. U.S. Patent Application US 020190139567 A1, International Patent Application WO 002017196422 A1, 2016b.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Kurtosis-Controlled Babble Noise Suppression. In *Proc. of ITG Conference on Speech Communication*, Paderborn, Germany, 2016c.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Voice Activity Detection Based on Modulation-Phase Differences. In *Proc. of ITG Conference on Speech Communication*, Paderborn, Germany, 2016d.

Simon Graf, Markus Buck, and Tobias Herbig. System and Method for Speech Detection Adaptation. International Patent Application WO 2017119901 A1, 2017a.

Simon Graf, Tobias Herbig, and Markus Buck. Low complexity detection of voiced speech and pitch estimation. U.S. Patent Application US 000011176957 B2, International Patent Application WO 002019035835 A1, 2017b.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Low-Complexity Pitch Estimation Based on Phase Differences Between Low-Resolution Spectra. In *Proc. of Interspeech*, Stockholm, Sweden, August 2017c.

Simon Graf, Nabeel Zaidi, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Detection of Voiced Speech and Pitch Estimation for Applications with Low Spectral Resolution. In *Proc. of DAGA*, Kiel, Germany, 2017d.

Acronyms

ACF Auto-correlation function

AFE Advanced front-end (feature extraction)

AGC Automatic gain control

AMR Adaptive multi-rate codec

AMS Amplitude modulation spectrogram

ASR Automatic speech recognition

AUC Area under (receiver operating characteristic) curve

CCF Cross-correlation function

DCT Discrete cosine transform

DFT Discrete Fourier transform

EACF Extended auto-correlation function

ETSI European Telecommunications Standards Institute

FEC Front end clipping

FFT Fast Fourier transform

FIR Finite impulse response

GCC Generalized cross-correlation function

HMM Hidden Markov model

HPS Harmonic product spectrum

IBM Ideal binary mask

ICA Independent component analysis

ICC In-car-communication

IIR Infinite impulse response

IMCRA Improved minima controlled recursive averaging

ITU International Telecommunication Union

LPC Linear predictive coding

LSF Line spectral frequencies

LTSD Long-term spectral divergence

LTSV Long-term signal variability

MCRA Minima controlled recursive averaging

MFCC Mel-frequency cepstral coefficient

MMSE Minimum mean squared error

MPD Modulation phase difference

MSC Mid speech clipping

MSE Mean squared error

MUSHRA MUlti Stimulus test with Hidden Reference and Anchor

NDS Noise detected as speech

OVER Hangover after speech

PCA Principal component analysis

PDF Probability density function

PSD Power spectral density

ROC Receiver operating characteristic

SAD Speaker activity detection

SNR Signal-to-noise ratio

STFT Short-time Fourier transform

STM Spectro-temporal modulation

STMF Spectro-temporal modulation filter

SVM Support vector machine

VAD Voice activity detection

ZCR Zero-crossing rate

Notation and Important Symbols

Notation

$\boldsymbol{x}$ Bold letters denote vectors and matrices, Page 26
$\square^*$ Conjugate of a complex-valued variable, Page 10
$\widehat{\square}$ The variable represents an estimated value, Page 14
$\mathrm{E}\{\square\}$ Mean value of a random variable, Page 12
$\square'$ The variable refers to a shorter frame, Page 36
$\square^T$ Transposed vector or matrix, Page 26

Indices and sampling

f_s Sample rate, Page 8
K Number of frequency bins, Page 10
k Frequency bin index, Page 10
ℓ Frame index, Page 10
N Frame length, Page 10
n Sample index, Page 8
$\tilde{n}$ Relative sample index, Page 10
R Frameshift, Page 10
τ Correlation lag, Page 32

Signals and spectra

$b(n)$ Background noise signal, Page 8
$\hat{\Phi}_{bb}(k,\ell)$ Estimated power spectral density of the background noise, Eq. (2.15), Page 14
$\hat{\Phi}_{xx}(k,\ell)$ Estimated power spectral density of the noisy signal, Eq. (2.14), Page 14
$s(n)$ Speech signal, Page 8
$x(n)$ Noisy audio signal, Eq. (2.1), Page 8

Detection and evaluation

η Detection threshold, Page 19
f Scalar feature, Page 18
$\boldsymbol{f}$ Feature vector, Page 18
P_d (Correct) detection rate, Eq. (4.7), Page 77
P_fa False-alarm rate, Eq. (4.10), Page 77
$P_\mathrm{t}(\Delta\ell)$ Transient measure, Eq. (4.12), Page 81
$VAD(\ell)$ Speech detection result, Page 10
$VAD_\mathrm{ref}(\ell)$ Reference for speech detection evaluation, Page 67

VAD Features

f_{ACF} Auto-correlation maximum feature, Eq. (3.20), Page 32, Page 85

$\boldsymbol{f}_{\mathrm{AMS}}$ Amplitude modulation spectrum feature, Eq. (3.65), Page 56, Page 88

f_{CEP1} Cepstral peak feature, Eq. (3.23), Page 33, Page 86

$\boldsymbol{f}_{\mathrm{CEP2}}$ Cepstral coefficients (lower order) feature, Eq. (3.56), Page 51, Page 87

f_{CEP3} Cepstral peak feature for babble noise suppression, Eq. (5.7), Page 93

f_{HPS} Harmonic product spectrum feature, Eq. (3.25), Page 34, Page 86

f_{KUR} Kurtosis feature for babble noise suppression, Eq. (5.5), Page 93

$\boldsymbol{f}_{\mathrm{LPC}}$ Linear predictive coding coefficients feature, Eq. (3.53), Page 49, Page 86

f_{LPCS} Pitch detection feature based on a linear phase of a cross-spectrum, Eq. (3.49), Page 48, Page 104

$\boldsymbol{f}_{\mathrm{LSF}}$ Line spectral frequencies feature, Eq. (3.55), Page 50, Page 86

f_{LSFM} Long-term spectral flatness measure feature, Eq. (3.61), Page 54, Page 87

f_{LTSD} Long-term spectral divergence feature, Eq. (3.14), Page 29, Page 84

f_{LTSV} Long-term signal variability feature, Eq. (3.60), Page 54, Page 87, Page 103

f_{M4Hz} 4 Hz modulation feature, Eq. (3.62), Page 55, Page 87

$\boldsymbol{f}_{\mathrm{MFCC}}$ Mel-filtered cepstral coefficients feature, Eq. (3.57), Page 52, Page 87

f_{MPD} Modulation-phase difference feature, Eq. (3.76), Page 60, Page 103

f_{NSTP} Normalized short-term power feature, Eq. (3.2), Page 26, Page 83

$\boldsymbol{f}_{\mathrm{PED}}$ Power envelope dynamics feature, Eq. (3.3), Page 26, Page 84

f_{SE} Spectral entropy feature, Eq. (3.17), Pages 30, 31, Page 85

f_{SNR1} Signal-to-noise power feature, Eq. (3.8), Page 27, Page 84

f_{SNR2} Signal-to-noise power feature, Eq. (3.12), Page 29, Page 84

$\boldsymbol{f}_{\mathrm{STM}}$ Spectro-temporal modulation feature, Eq. (3.67), Page 57, Page 88

f_{STP} Short-term power feature, Eq. (2.16), Page 18, Page 83

f_{ZCR} Zero-crossing rate feature, Eq. (3.15), Page 30, Page 85