IMPERIAL COLLEGE LONDON

FINAL YEAR PROJECT

Robust Speech Detection in High Levels of Background Noise

Author:
Marcin Baginski

Supervisor:
Mike Brookes

This report is submitted in fulfilment of the requirements for the degree of *MEng Information Systems Engineering* in the Department of Electrical and Electronic Engineering Imperial College London

February 2014

Declaration of Authorship

I, Marcin Baginski, declare that this thesis titled, 'Robust Speech Detection in High Levels of Background Noise' and the work presented in it are my own. I confirm that:

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated
- Where I have consulted the published work of others, this is always clearly attributed
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Signed: ______________________________

Date: ______________________________

IMPERIAL COLLEGE LONDON

Abstract

Department of Electrical and Electronic Engineering

MEng Information Systems Engineering

Robust Speech Detection in High Levels of Background Noise

by Marcin Baginski

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas pretium sem nec nisi facilisis, vel consectetur libero rutrum. Curabitur rhoncus commodo leo, nec lobortis ante venenatis vehicula. Duis vel posuere risus. Nulla blandit risus elit, quis eleifend leo lacinia ut. Mauris rutrum vitae orci eu commodo. Suspendisse egestas, ipsum quis interdum rutrum, tortor lectus facilisis lorem, eget mollis ligula arcu at mi. Quisque accumsan orci magna, sit amet interdum lacus commodo non. Mauris elit magna, venenatis in auctor sit amet, laoreet in tortor. Suspendisse et dolor mattis, tempus libero sit amet, tempus lectus. Nunc in dolor et lorem dignissim elementum. Curabitur suscipit lectus lorem. Pellentesque ultrices venenatis neque, vitae consectetur arcu mattis sed. Quisque porta nisl elementum lacus mollis commodo. Sed vestibulum dolor sed lectus interdum eleifend. Quisque in libero ut augue blandit malesuada.

Acknowledgements

I would like to thank . . .

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Voice Activity Detection

Voice Activity Detection (VAD) is a process of identifying parts of an audio recording which contain the presence of human voice as opposed to those which are only comprised of silence or the background noise. VAD is a relatively simple task in recordings which have high signal-to-noise ratios (SNR), in which voice can be distinguished from noise simply by computing the short-time energy of all frames and setting an appropriate threshold for their classification. However, in most modern applications, the signal is almost always corrupted to some extent by a background noise which makes the VAD performance to deteriorate. While some types of noise can be relatively easily dealt with, i.e. those with spectral characteristics different from speech, in the presence of other, it might be very difficult to identify speech segments. One such noise type might be the *babble noise* which consists of speech which is not of particular interest. Additionally, VAD decision is especially difficult for the unvoiced phonemes whose spectrum contains no periodicity and is often similar to the one of white noise [1, 2].

There has been an active research in the VAD area from as early as 1975, when Rabiner and Sambur [3] proposed a VAD algorithm (then referred to as *algorithm for determining the endpoints of isolated utterances*) based on the aforementioned short-time energy and the zero-crossing rate. Such approach works reasonably well for signals with the SNR on the order of 30 dB and is suitable for a variety of applications which are not subject to a constant, high level of background noise, such as telecommunications, when a person speaks to a closely positioned microphone in a relatively calm environment. However since then there has been a need for much better performance, including algorithms whose robustness has to be achieved even at negative SNRs. A person driving a car, trying to communicate with their smartphone through its built-in speech recognition system (e.g. Apple Siri) might be one example of such application. Figure 1.1 shows a comparison of the amplitude of a clean utterance with the same utterance corrupted by a -5 dB car noise. As it can be seen, some parts of the recording are completely submerged in the noise (especially during the second second of the recording) and their detection poses a considerable challenge for any VAD system. In robotics there is often a desire to communicate with the robot by speaking from a distance which naturally decreases power of the received signal. Taking the detrimental effect of the background noise into consideration, the simple algorithms are likely to either fail completely or their performance might significantly drop.

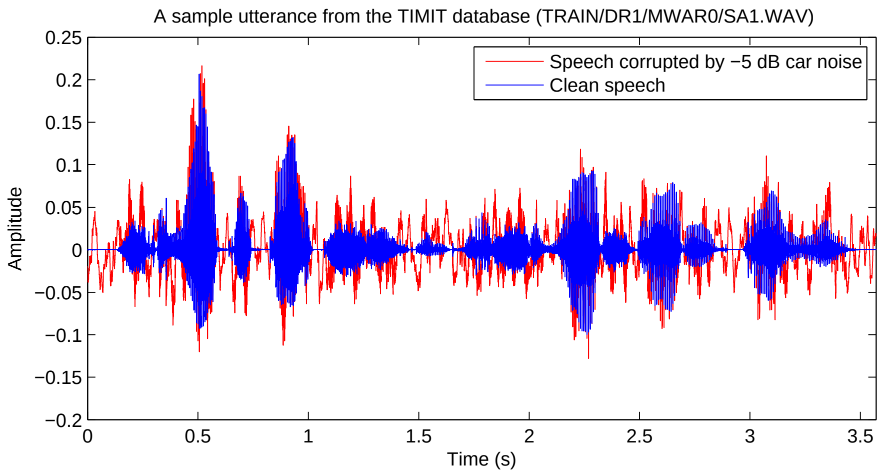

Figure 1.1: A sample utterance from the TIMIT [4] speech corpus corrupted by -5 dB car noise from the NOISEX-92 [5] database

Over the years, numerous VAD algorithms have been proposed, with features based on the energy [6, 7], pitch detection [8], long-term speech information [9], zero-crossing rate [10], higher order statistics [11], periodicity measures [12] and other.

1.2 Applications of VAD

VAD is often the first step in many signal processing applications including speech recognition [9, 13–17], speech coding and transmission [6, 13, 18–20], speech enhancement [13, 21, 22], noise estimation [13] or speaker recognition [23]. In most applications the noise-robust VAD decisions reduce the computational load required by the system and improve its accuracy. The reduced computational load is achieved since the voice-inactive frames are often either transmitted at a much lower bit-rate or not processed at all. At the same time, the clear boundaries of an utterance help to improve the accuracy of some systems (e.g. speech recognition).

1.2.1 Automatic Speech Recognition

In Automatic Speech Recognition (ASR), it is of importance to first extract the voice-active parts of a signal which can then be passed to the actual recognition module. This procedure increases both the accuracy of the ASR system as well as its speed, since the recognition task is not performed on the parts of the signal which do not contain speech. A sample block diagram of an ASR system which uses a VAD module is presented in Figure 1.2 [13]. For ASR, and also most other applications, it is crucial for the VAD module to be able to identify all speech segments in order not to degrade the accuracy of the entire system. Therefore, VAD systems often implement a fail-safe approach which means that if there is an uncertainty in classification of a frame, it is safer to label is as speech than otherwise. Typically, there is a trade-off in VAD performance which can be characterised as maximising the precision while keeping the recall at a steady, high rate.

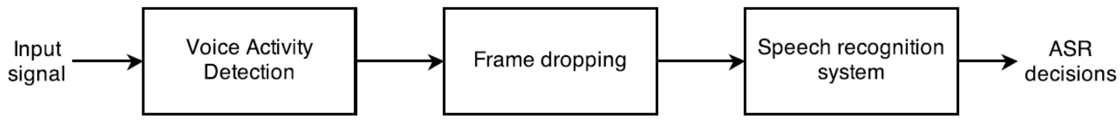

Figure 1.2: Block diagram of an Automatic Speech Recognition system with Voice Activity Detection module [13]

1.2.2 Speech Coding and Transmission

A typical phone conversation involves each person speaking on average no more than 50% of the time [20]. Therefore, signal transmission would be greatly optimised if each transmitter was switched-off half of the time. Such approach could cause the overall system capacity to double. The technique of interrupted transmission during periods of silence is known as discontinuous transmission (DTX). In order to work properly, it requires a precise Voice Activity Detector to direct the operation of a transmitter between being switched on or off. As an alternate method to stopping the transmission, a dual-mode encoding technique could be employed, which uses a higher bit-rate for coding the voice-active frames and lower for silence/noise. The latter is precisely what the popular ITU-T G.729 Annex B [6] standard does, transmitting the voice-active parts at a fixed bit rate of 8 kb/s while the noisy ones at only 15 b/frame.

Figure 1.3 shows a high-level structure of a dual-mode coding and transmission system, in which the VAD module is used to direct the incoming signal into either the active or inactive speech encoder. The noise can be either transmitted at a much lower bit-rate or the transmission might be switched off completely. In case of a stopped transmission, the receiving end often implements a *comfort noise* [6, 13, 20] generation module, which creates a synthetic signal similar to the background noise at the transmitter so that the listener does not notice the rapid, inconvenient switching during the conversation.

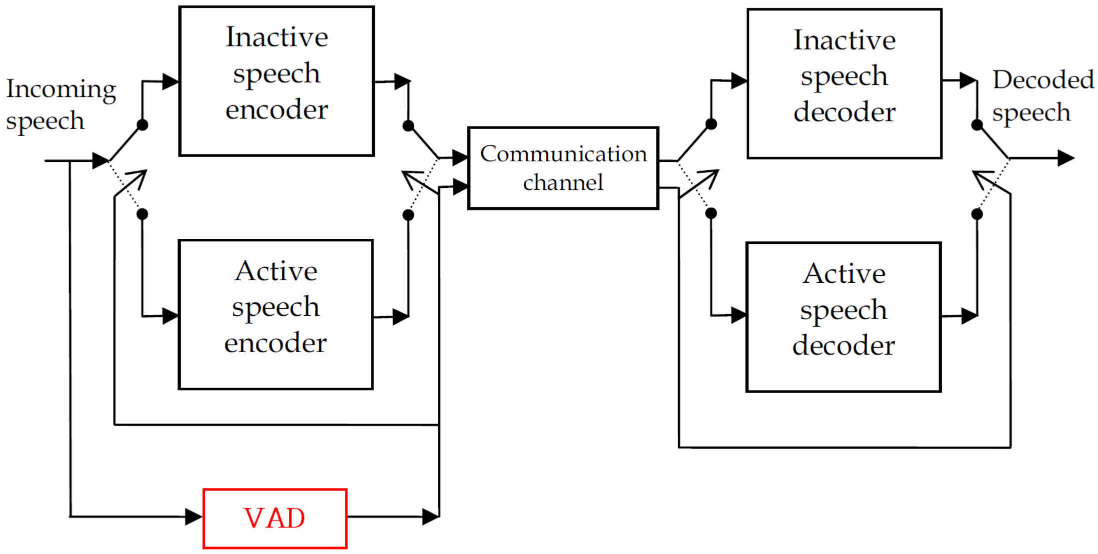

FIGURE 1.3: Block diagram of an dual-mode transmission system with Voice Activity Detection module [6]

1.2.3 Noise Estimation and Speech Enhancement

Speech enhancement aims to improve the intelligibility and quality of speech signals corrupted by additive noise of some kind. Many speech enhancement systems use a technique called *spectral subtraction* [1, 13]. It assumes, that the clean speech can be represented in the frequency-domain in the form:

$$|S(f)| = |Y(f)| - |N(f)| \tag{1.1}$$

where $|Y(f)|$ is the amplitude spectrum of the corrupted speech, $|S(f)|$ of the clean speech and $|N(f)|$ of the noise. In order for this technique to work, one needs to estimate the spectrum of the noise, which in real-world applications where a variety of different, often nonstationary, noise types are encountered, is a difficult problem[1]. A robust Voice Activity Detector can become very useful in this task by identifying the voice-inactive frames of a signal from which the noise statistics could be estimated. A precise VAD can also become useful in applications dealing with slowly varying piecewise stationary noises, where the noise statistics can be adaptively estimated based on the most recent VAD decisions.

1.2.4 Summary

Voice Activity Detection algorithms are utilised in a variety of speech processing tasks. Among others, application of the noise-robust VAD allows to decrease the bandwidth requirements of speech coding and transmission systems, improve accuracy and speed of the speech recognition systems and helps in precise noise statistics estimation and speech enhancement. The bandwidth improvement comes from the lower bit-rate coding of the non-speech frames. In the ASR systems, the voice-inactive frames are often dropped from processing in the core system. In speech enhancement, the noise statistics can often be estimated from the noise-only frames, therefore enabling the use of spectral subtraction and other techniques which require prior knowledge of the noise spectrum.

1.3 Structure of a typical VAD system

Figure 1.4 shows a high-level structure of a Voice Activity Detector, however only the two middle blocks are considered a core of a typical VAD system. In the rest of this section, the operation of each block is going to be described in some detail.

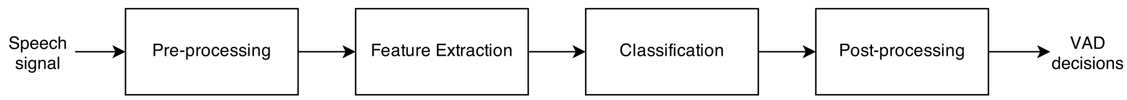

FIGURE 1.4: Block diagram of a typical Voice Activity Detection system

[1] In fact, noise estimation is another, related research field to Voice Activity Detection and some algorithms actually use noise estimation modules in their operation.

1.3.1 Pre-processing

The noisy speech signal is first passed to a pre-processing module which might perform a variety of tasks before the actual voice detection takes place. Most commonly, during pre-processing the input signal is often split into frames which are typically 10-50 ms long, either overlapping or not. Additionally, the module might perform noise estimation and suppression in the signal in order to improve performance of the VAD. Sometimes, average noise statistics are computed during pre-processing for later use.

1.3.2 Feature Extraction

The purpose of the feature extraction module is to compute the speech features for each frame which are suitable for the speech/non-speech classification. These can be time domain [24, 25], frequency domain [8, 9, 18, 26–30], cepstral-domain [31] features or other, depending on the specific VAD algorithm. Ideally, in order to achieve the overall VAD system robustness, the selected signal characteristics should not be easily corruptible by the background noise. Additionally, the algorithms for feature extraction should be of low computational complexity for their potential usefulness in real-time applications. The output of this module is therefore a vector $\mathbf{x}$ of features for each of the frames computed in the pre-processing stage.

1.3.3 Classification

The classification module assigns a binary class (speech/non-speech) to each frame based on the feature vector $\mathbf{x}$ received from the previous processing stage. Classification might be based on a number of decision rules, ranging from simple thresholding [6] to more advanced methods such as statistical likelihood ratio tests [17, 18, 29] or machine learning [32, 33]. Obviously, the classifier performance degrades with the increasing power of the background noise, therefore there is a need also at this stage for a robust decision making rule. Some researchers [24] considered a combination of multiple decision rules in making the final classification.

1.3.4 Post-processing

The last module, post-processing, often tries to *smooth* the VAD decisions (i.e. perform hanging-over) in order to reduce the number of false positives and false negatives. Smoothing is important in order to precisely detect the beginnings and endings of speech bursts, which often have much lower energy than the rest of the signal. Additionally, if among 50 consecutive frames, each of 20 ms duration, only one is classified as speech, the post-processing module might change the decision for this particular frame, since it is highly unlikely for speech to be active during such short-time window. The hang-over schemes proposed in the literature are often based on simple heuristics [6] or other techniques such as Hidden Markov models [18].

The hang-over scheme is especially important for the VAD algorithms which are based on the harmonicity of the voiced speech. Since the unvoiced parts do not have a fundamental frequency, the pitch-based algorithms by definition cannot classify such parts of the signal as speech. However, since the unvoiced phonemes are almost always surrounded by the voiced ones, an efficient hang-over scheme can help in reducing the false-negative rate (i.e. classification of speech segments as noise).

1.4 Report organisation

The rest of this document is organised as follows:

- Chapter 2 contains a literature survey of both the standardised as well as recently proposed VAD algorithms from various sources such as conference proceedings or scientific journals
- Chapter 3 contains the project objectives and planning

Chapter 2

Literature Survey of VAD algorithms

2.1 Standard VAD algorithms

Being an important tool in many speech processing applications, a number of VAD algorithms have been subject to standardisation by various organisations such as the International Telecommunication Union (ITU-T), European Telecommunications Standards Institute (ETSI), Telecommunications Industry Association (TIA) or Electronic Industries Alliance (EIA). Most standardised algorithms use the energy of the input signal as a Voice Activity Detection feature. It is important to note that the standardised VAD approaches have been developed for use in the telecommunications industry, with particular emphasis on the application for discontinuous transmission (DTX), which may make them less appropriate for other speech processing tasks such as speech recognition. Nevertheless, these algorithms often serve as a benchmark for the newly developed VAD features, whose performance is often compared to the standard ones.

In the rest of this section, three standard VAD algorithms are going to be described:

- ITU-T G.729 Annex B [6] which is an extension to the G.729 speech coder with an aim to achieve an improved bit rate during the noise-only periods
- ETSI AMR1 and AMR2 [7] for application to the Global System for Mobile Communications (GSM)

- TIA/EIA IS-733 [34] for application to the Wideband Spread Spectrum Communication Systems

2.1.1 ITU-T G.729 Annex B

The well-known ITU-T G.729 Annex B VAD has been developed as an extension to the G.729 speech coding algorithm [10] transmitting each frame at a fixed bit rate of 8 kb/s. Application of the Voice Activity Detector allows to identify the noise-only frames in a continuous stream of data and adopt a compressed transmission at only 15 b/frame which contains information about the background noise for reproduction by the Comfort Noise Generator (CNG) at the receiving end. This approach for speech/noise coding allows to reduce the average bit-rate of the entire coder from 8 kb/s to only 4 kb/s while keeping the transmission quality unchanged.

The block diagram of the VAD algorithm is presented in Figure 2.1. It starts with computation of four main *instantaneous parameters* for the current frame which describe the energy and spectral content of the signal:

- Set of Line Spectral Frequencies (LSF)
- Full-band energy (E_f)
- Low-band (0 to 1 kHz) energy (E_l)
- Zero-crossing rate (ZCR)

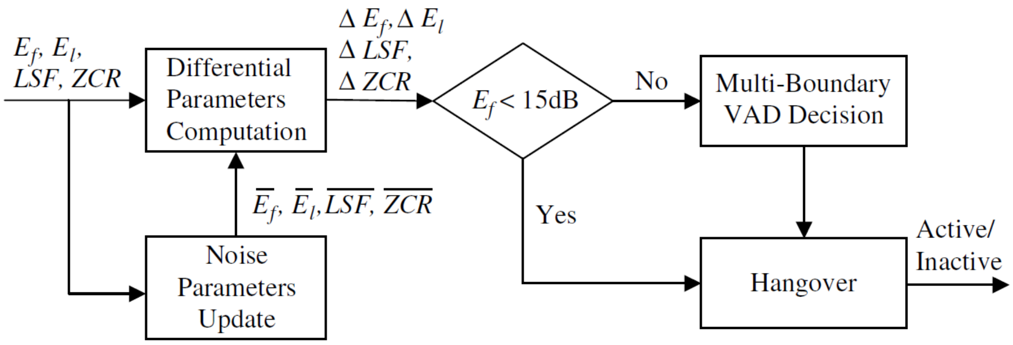

FIGURE 2.1: Block diagram of the ITU-T G.729 Annex B VAD [1]

The *instantaneous parameters* are then differenced with their most recent average noise-only counterparts in order to derive an additional set of so called *difference parameters* which are used for speech/non-speech classification. The set of all possible *difference parameters* describes a four dimensional Euclidean space in which a specific region contains the speech frames while another region describes the noise-only frames. The current vector of parameters is compared against the pre-computed regions in order to classify the current frame. The two regions are initially identified by visual inspection of the points' distribution over a large set of clean and noisy recordings. An energy threshold of $E_f < 15$ dB is applied before the multi-boundary classification in order to minimise short glitches on low-energy frames.

ITU-T G.729 Annex B uses an additional four-step heuristic-based smoothing scheme after the initial multi-boundary classification:

1. An active voice decision is extended to the current frame if its energy is above a certain threshold
2. An active voice decision is extended to the current frame if the previous two frames were speech and the absolute energy difference between the current and previous frames' is under a certain threshold
3. An inactive voice decision is extended to the current frame if the previous 10 frames were noise-only and the absolute energy difference between current and previous frames' is under a certain threshold
4. The active voice frame is labelled as inactive if the current frame energy is below a noise floor by a certain threshold

The main VAD algorithm also performs updating of the noise parameters ($\overline{LSF}$, $\overline{E_f}$, $\overline{E_l}$, $\overline{ZCR}$) by a secondary VAD decision which does not need to be as robust as the primary one since it is used only for estimation of the noise parameters.

2.1.2 ETSI AMR1 and AMR2

ETSI proposed two VAD alternatives for use in the Adaptive Multi-Rate speech traffic channels. In both algorithms, the decision is primarily based on the energy of the signal across different frequency bands.

Block diagram of the AMR Option 1 VAD is presented in Figure 2.2. The original algorithm includes additional processing steps to those depicted in the Figure in order to determine whether the incoming signal, if not noise-only, contains speech, special information tone (STI) or other (e.g. music), however this details are are omitted in this description.

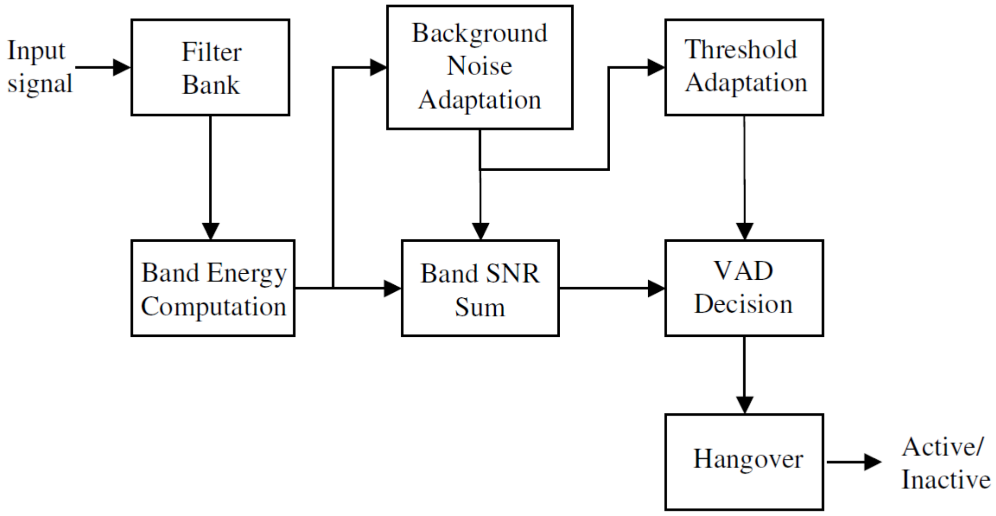

Figure 2.2: Block diagram of the ETSI AMR Option 1 VAD [1]

The input signal is first passed through a series of nine band-pass filters which split the time-domain signal into different frequency bands based on the Table 2.1. The signal level $level[n]$ is calculated at the output of each filter as a sum of the absolute values of all samples in the current frame. The VAD feature is then computed according to the Equation 2.1

$$\mathrm{SNR} = \sum_{n=1}^{9} \max(1.0, \frac{level[n]}{bckr\_est[n]})^2 \tag{2.1}$$

where $bckr\_est[n]$ is the estimated level of noise at frequency band n. The VAD feature from the above equation is compared to a threshold in order to classify the current frame. The threshold is determined based on the estimated average background noise level which is the sum of $bckr\_est[n]$ for all n. As a final processing step, AMR Option 1 VAD includes a hang-over scheme in order to detect the low-energy endings of speech bursts.

Block diagram of ETSI AMR Option 2 VAD is presented in Figure 2.3. The concept is similar to Option 1 VAD, however the incoming signal is split into different frequencies not by time-domain band-pass filtering, but by first computing the Discrete Fourier Transform (DFT) of the signal and performing further analysis in the frequency domain.

	Frequencies
Filter 1	0 - 250 Hz
Filter 2	250 - 500 Hz
Filter 3	500 - 750 Hz
Filter 4	750 - 1000 Hz
Filter 5	1000 - 1500 Hz
Filter 6	1500 - 2000 Hz
Filter 7	2000 - 2500 Hz
Filter 8	2500 - 3000 Hz
Filter 9	3000 - 4000 Hz

TABLE 2.1: Cut-off frequencies for the ETSI AMR1 band-pass filters [7]

The frequencies are clustered into bands (channels) and the energy of each channel is calculated [35]. In the next processing steps, SNR of each channel is calculated and transformed to a *voice metric* by a specific function which results in the final VAD feature to be classified by using a threshold. An additional part of the system performs updates of the noise statistics based on the spectral deviation estimate.

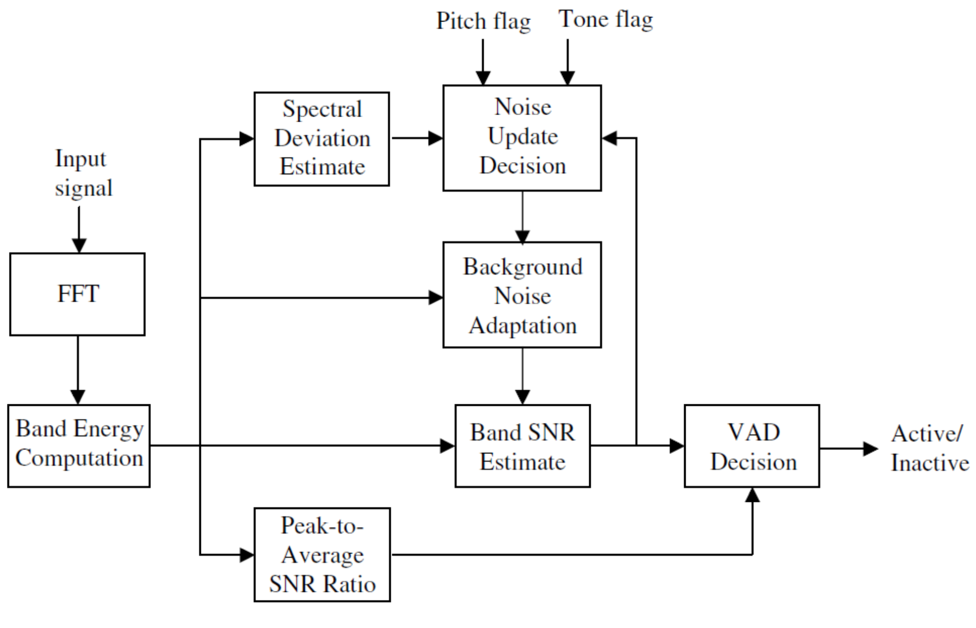

FIGURE 2.3: Block diagram of the ETSI AMR Option 2 VAD [1]

2.1.3 TIA/EIA IS-733

TIA/EIA IS-733 is a speech coder in which the signal might be encoded at four different rates (1, 1/2, 1/4, 1/8 of the base rate) depending on the characteristics of the currently transmitted frame. Rate 1 is used for low quality signals where additional reduction might compromise the already low intelligibility. Rate 1/2 is used for good quality stationary and periodic frames. Rate 1/4 is used for unvoiced speech and rate 1/8 for speech inactive frames. A VAD algorithm is used to determine the rate at which the current frame should be encoded and transmitted.

Block diagram of TIA/EIA IS-733 VAD is presented in Figure 2.4. The algorithm starts with computing the energy of the input signal across two different frequency bands (0.3 - 2.0 kHz and 2.0 - 4.0 kHz) and subsequently the SNR based on the estimated noise energy. The VAD decision is based on two adaptive thresholds, which depend on the level of the estimated background noise, one for each frequency band. If both low and high band SNRs are higher than the threshold, rate 1 is selected. Only one SNR being above its threshold causes the signal to be encoded at rate 1/2. Both SNRs below the threshold indicate noise-only frame, encoded at rate 1/8.

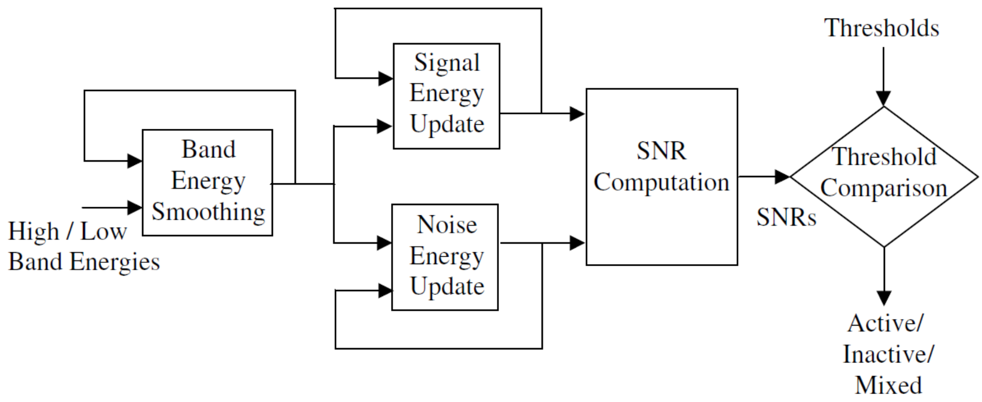

FIGURE 2.4: Block diagram of the TIA/EIA IS-733 VAD [1]

2.1.4 Summary

Some VAD algorithms have been standardised by various telecommunication standards institutes (ITU-T, ETSI, TIA/EIA). Those algorithms are predominantly based on simple features such as the energy of the signal or the zero-crossing rate, sometimes across different frequency ranges. While such measures are often sufficient to meet the requirements of the high SNR transmission found in telecommunications, their performance drops significantly with the increased power of noise. Nevertheless, the standard algorithms serve as a convenient benchmark for the more recently proposed VAD features, described in section 2.2 which are aimed to be more noise-robust and useful in other speech processing tasks.

2.2 Noise-robust VAD algorithms

Apart from standard VAD algorithms described in section 2.1, many independent researchers have made numerous attempts to develop novel noise-robust Voice Activity Detection methods. Most of these research results either in invention of the new features or identification of ways in which the existing ones might be improved. Ideally, the most robust feature should have no common values for the noise and speech frames. Figure 2.5 shows the distribution of values for some of the features used by algorithms described in this section for the clean speech from the TIMIT [4] database and a variety of noise types from the NOISEX-92 [5] database. It is clear, that in both cases the values for the clean speech are mostly distinct from the ones for noise frames, however there is still much overlap between them which indicates a room for improvement.

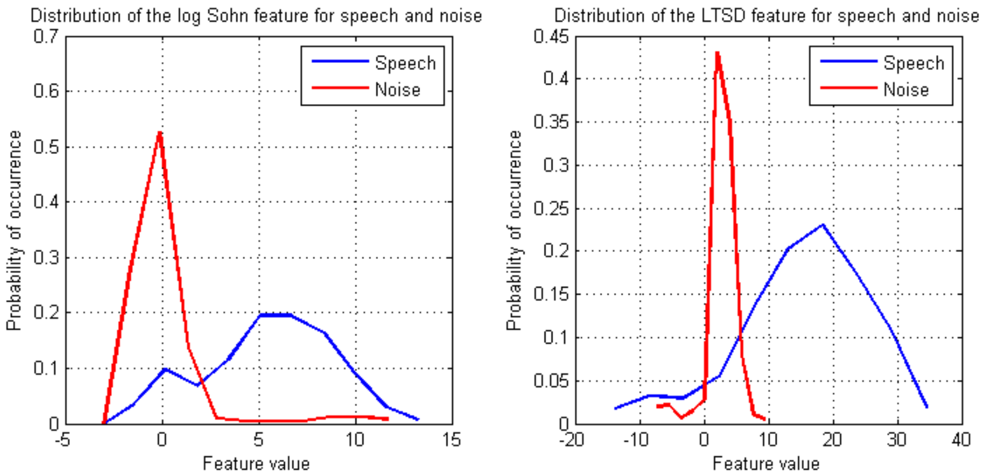

FIGURE 2.5: Distribution of Sohn's [29] and LTSD [9] features for speech and noise

2.2.1 Entropy-based VAD

In contrast to the standardised energy-based methods, some researchers investigated the idea of using entropy for Voice Activity Detection. Entropy, as originally defined by Shannon [36], is a measure of uncertainty in a random variable, given by the equation:

$$E = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{2.2}$$

where p_i is the probability of the random variable having a value of i among N distinct values which it might take. In case of VAD, p_i often relates to either a single bin in a histogram of the amplitudes of a signal or a single frequency in the magnitude or power spectrum.

The purely time domain approach to VAD using entropy has been explored by Weaver *et al.* in [25]. Block diagram of the key parts of the algorithm is presented in Figure 2.6. Authors propose to first calculate the histogram of the amplitudes of the signal and then assign an entropy measure to each frame, as defined in Equation 2.2, where p_i is the normalised (such that $\sum_{i=1}^{N} p_i = 1$) mass of the i-th bin of the histogram.

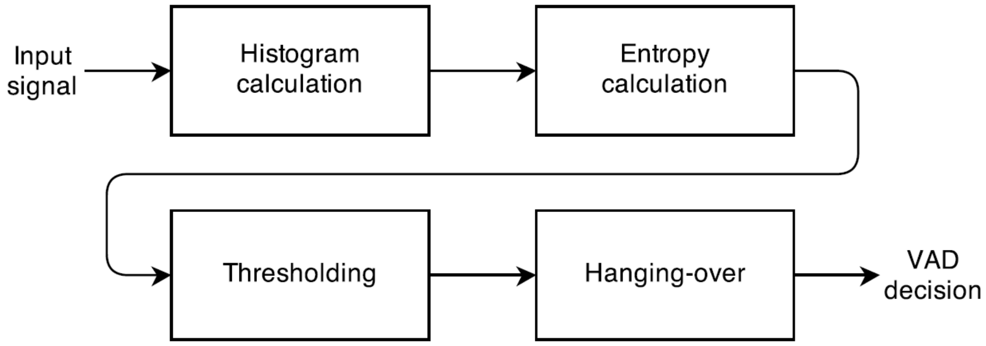

FIGURE 2.6: Block diagram of the time-domain entropy-based VAD [25]

While this approach is more noise-robust than the simple energy-based methods, especially for the stationary narrowband noise types, its performance is significantly affected by the various coloured noises. In order to mitigate this weakness, authors propose to use a weighting filter in order to enhance the typical speech frequencies, however it needs to be kept in mind that if we are faced with a noise with spectral characteristics very similar to speech (especially in case of the *babble noise*), the filter would become essentially useless as it would emphasise the noise as well.

A frequency domain approach to using entropy for VAD has been considered by Renevey *et al.* in [30]. Instead of first computing the histogram of the amplitudes of the input signal, the frequency-domain algorithm starts with calculating the power spectrum of each frame. Entropy of the spectrum is then calculated by means of the following equation:

$$E(|Y(\omega,t)|^2) = -\sum_{i=1}^{L} P(|Y(\omega_i,t)|^2)\log\left(P(|Y(\omega_i,t)|^2)\right) \tag{2.3}$$

where $P(|Y(\omega_i,t)|^2)$ denotes the fraction of the sum of all harmonics which is attributable to the current frequency i.

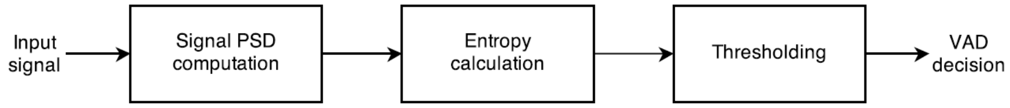

FIGURE 2.7: Block diagram of the frequency-domain entropy-based VAD [30]

All entropy-based approaches to VAD are most effective for the white noise and alike, since the distribution of entropy for speech is much different from the one for noise. The completely unpredictable nature of white noise yields high entropy values, while the more organised and predictable clean speech signals present a naturally lower entropy. Both the time and frequency domain algorithms' performance drops significantly for a variety of coloured noise types. In order to improve the performance for in such environments, authors of [30] propose using a *whitening* filter which divides the spectrum of the current frame by an average of all frames.

2.2.2 Likelihood Ratio Test VAD

'A Statistical Model-Based Voice Activity Detector' proposed by Sohn *et al.* [18] is one of the most widely cited VAD algorithms due to its ease of implementation, robustness and extensibility. The idea of using a Likelihood Ratio Test (LRT) has been considered by many other researchers who tried to improve on the original approach [17, 37]. The initial algorithm has been developed in [29] and extended to improve noise-robustness in [18]. Block diagram of Sohn's VAD is presented in Figure 2.8.

The algorithm is based on a LRT to discriminate between two hypotheses:

H_0 - speech absent

H_1 - speech present

The measure which is used for the preliminary VAD decision (i.e. before the hang-over scheme) is a combination of the likelihood ratios from each frequency bin k:

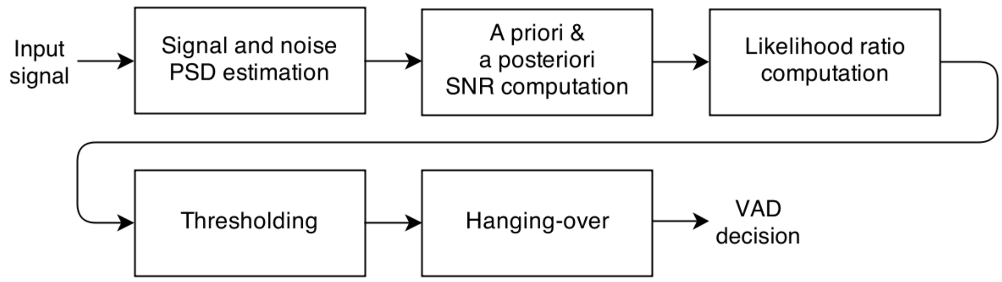

FIGURE 2.8: Block diagram of the Statistical Model-Based VAD [18]

$$\log \Lambda = \frac{1}{L} \sum_{k=0}^{L-1} \log \frac{1}{1+\xi_k} e^{\frac{\gamma_k \xi_k}{1+\xi_k}} \tag{2.4}$$

where L is the number of samples in each frame, ξ_k is the *a priori* SNR and γ_k is the *a posteriori* SNR. In order for the algorithm to work, one needs to obtain the values for $\xi_k = \frac{|S_k|^2}{|N_k|^2}$ and $\gamma_k = \frac{|X_k|^2}{|N_k|^2}$ where $|X_k|^2, |S_k|^2, |N_k|^2$ are the PSDs of the noisy speech, clean speech and noise respectively. $|N_k|^2$ can be obtained by means of any noise estimation procedure (e.g. [38] or [39]), whose accuracy influences the noise robustness of the algorithm. The VAD feature is thresholded by an empirically determined constant for the preliminary decision. Finally, authors proposed a Hidden Markov model (HMM) hang-over scheme in order to improve accuracy of the algorithm for the low-energy beginnings and endings of the speech utterances.

While the original idea to the derivation of the unknown *a priori* SNR ξ_k involved the Maximum Likelihood estimator $\xi_k^{ML} = \gamma_k - 1$, in [18] a limitation of the procedure has been identified which makes it biased towards H_1. In an effort to improve the algorithm, authors proposed a decision-directed (DD) estimation procedure which reduces the fluctuation of the likelihood ratios by using a MMSE short-time spectral amplitude estimator [40] and a first-order low pass filter.

In an effort to further improve the LRT-based algorithm, Cho *et al.* [37] investigated the DD approach with particular interest in the detection errors occurring at the endings of speech utterances. It was determined that the frequent misdetections are due to the delay in the DD *a priori* SNR estimator, which prevents the estimated value to drop quick enough for the likelihood ratio to stay above the threshold during the short, low-energy speech offset regions. In order to alleviate this problem, authors proposed a smoothed likelihood ratio (SRT), which delays the sudden drops in the LR at the speech offset regions due to the constant $\alpha \approx 0.9$. The SRT is defined in Equation 2.5 where n relates to the frame number while k is the frequency bin. The final VAD decision is calculated by taking a geometric mean of the SRTs from all frequency bins.

$$\Phi(n,k) = \exp\left\{\alpha \log\left(\Phi(n-1,k)\right) + (1-\alpha)\log\left(\Lambda(n,k)\right)\right\} \tag{2.5}$$

Both [18] and [37] VAD methods consider only a single frame when making a speech/non-speech decision and hence they are likely to misclassify the low-energy frames for which the short-time SNR is much lower than the average SNR of the entire signal. To aid the proper detection of such frames, in [28] Ramirez *et al.* proposed a multiple observation vector which considers M frames before and ahead of the current frame in formulating the likelihood ratio. The main rationale behind this idea is that the weaker speech frames are often surrounded by the stronger parts, and their inclusion in the LRT might boost its value above the threshold. While this approach results in somewhat improved performance, it also introduces the delay of M frames to the algorithm, which might prevent it from being used in some real-time applications.

2.2.3 Long-Term Spectral Divergence VAD

Another popular and widely cited algorithm is the Ramirez *et al.* [9] VAD based on the long-term speech information. The main assumption of the algorithm is that the most discriminative speech/non-speech information lies on the shape of the magnitude spectrum of the analysed signal. However, instead of considering each frame independently, the algorithm also includes the information contained in the neighbouring frames. The reason behind that is to boost the detection of the low-energy unvoiced phonemes which are typically surrounded by the high-energy voiced ones. Therefore, it can be said that the LTSD algorithm uses an implicit hang-over scheme incorporated directly in the voicing feature.

The algorithm starts by computing the so-called long-term spectral envelope (LTSE) which uses information contained in the current frame as well as N preceding and succeeding frames i.e. $2N + 1$ frames in total for every calculation. Based on the LTSE, the long-term spectral divergence (LTSD) is calculated which serves as the VAD decision rule. In Ramirez's study, it has been established that the best performance is achieved for $N = 6$ however this value is likely to be dependent on the particular application as well as the level of noise.

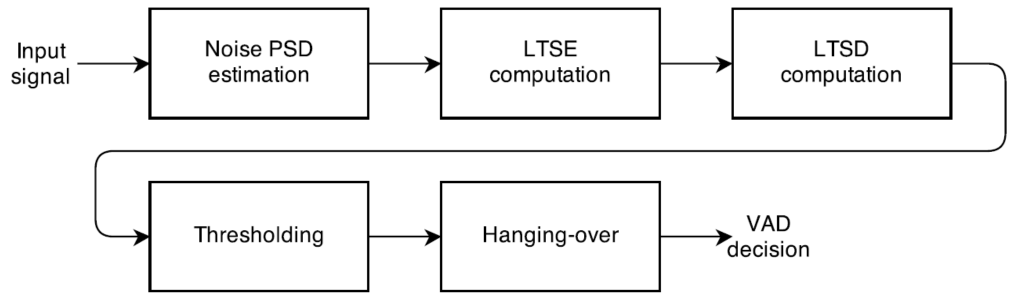

FIGURE 2.9: Block diagram of the Long-Term Spectral Divergence VAD [9]

Block diagram of the LTSD-based VAD is presented in Figure 2.9. The algorithm starts with an assumption that the first N frames of each utterance do not contain any speech and that the average magnitude spectrum of the noise can be estimated from them. After that, the LTSE for each frame is computed according to Equation 2.6 where $X(k,l)$ is the amplitude spectrum at frequency k for frame l.

$$\mathrm{LTSE}(k,l) = \max\{X(k,l-N),\ldots,X(k,l),\ldots,X(k,l+N)\} \tag{2.6}$$

The LTSD is obtained from Equation 2.7 where M is the number of frequency bins in the DFT and $N(k)$ is the average noise amplitude spectrum at frequency k as estimated before. Essentially what the equation describes is the average deviation of the LTSE from the noise statistics at each frequency bin. In other words, this measure might be interpreted as a variation of the estimated *a posteriori* signal-to-noise ratio, which is an idea exploited in many VAD algorithms.

$$\mathrm{LTSD}(l) = 10\log_{10}\left(\frac{1}{M}\sum_{k=0}^{M-1}\frac{\mathrm{LTSE}^2(k,l)}{N^2(k)}\right) \tag{2.7}$$

Eventually, the LTSD feature is thresholded to form a preliminary VAD decision which might be further revised by a separate hang-over scheme.

2.2.4 Pitch and fundamental frequency based VAD

A rather unique feature of voiced speech is its spectral harmonicity. The magnitude spectrum of voiced phonemes contains clearly visible peaks at equal intervals corresponding to the fundamental frequency F_0 or pitch, terms which in speech processing context are often used interchangeably. A spectrogram of a sample utterance corrupted by 0 dB car noise is presented in Figure 2.10. Although the energy of the noise is high (making the speech detection a challenge for the energy-based algorithms), its PSD occupies mostly the low frequencies (yellow box) causing the harmonic peaks (red boxes) to remain undistorted. Even in the presence of 0 dB white noise (Figure 2.11), which is much richer in frequency components than car noise, the harmonic peaks are preserved, although to a much smaller extent. While it is clearly possible to use the harmonicity features for the detection of the voiced parts of speech utterances, the unvoiced phonemes' spectrum does not contain harmonic peaks. Therefore, detection of the unvoiced phonemes remains difficult and often requires an additional technique or a specialised hang-over scheme.

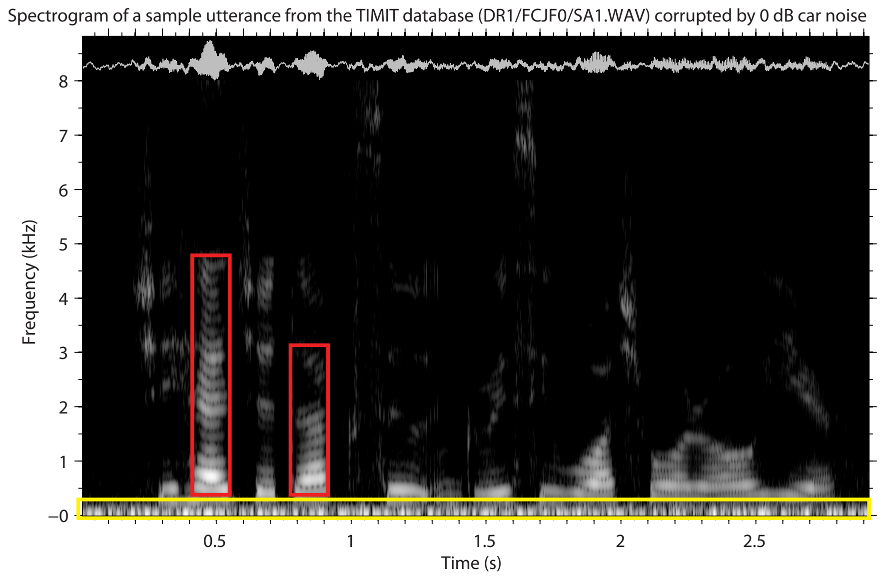

FIGURE 2.10: Spectrogram of a sample utterance corrupted by 0 dB car noise

In [8] Ishizuka *et al.* proposed a VAD (dubbed PARADE) based on the ratio of the powers of periodic to aperiodic components of the signal. Block diagram of the algorithm is presented in 2.12. The VAD decision is based on the likelihood ratio defined in Equation 2.8 where $\phi(i)$ equals $\frac{\hat{\lambda}_p(i)}{\hat{\lambda}_a(i)}$ - the ratio of the average power per frequency bin of the periodic to aperiodic components of the signal in frame i.

$$\Lambda(i) = \frac{1}{\phi(i)} \exp\left\{\frac{1}{2}\left(\phi(i)^2 - \frac{1}{\phi(i)^2}\right)\right\} \tag{2.8}$$

The authors propose to approximate the average powers from Equations 2.9 and 2.10 where ϑ is the number of harmonics in the current frame and η is a specific constant for

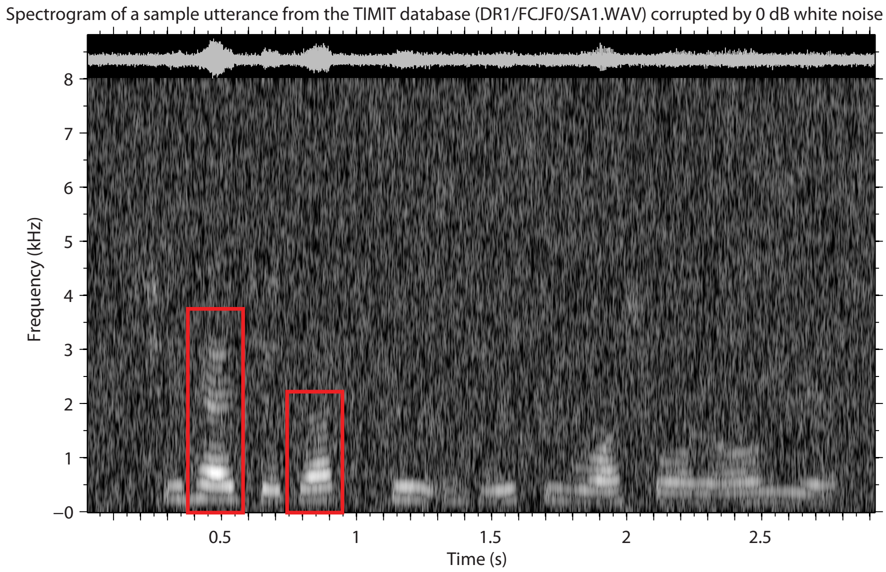

FIGURE 2.11: Spectrogram of a sample utterance corrupted by 0 dB white noise

power estimation.

$$\hat{\lambda}_a = \frac{\lambda - \eta \sum_{m=1}^{\vartheta} |X(mf_0)|^2}{1 - \eta\vartheta} \tag{2.9}$$

$$\hat{\lambda}_p = \lambda - \hat{\lambda}_a \tag{2.10}$$

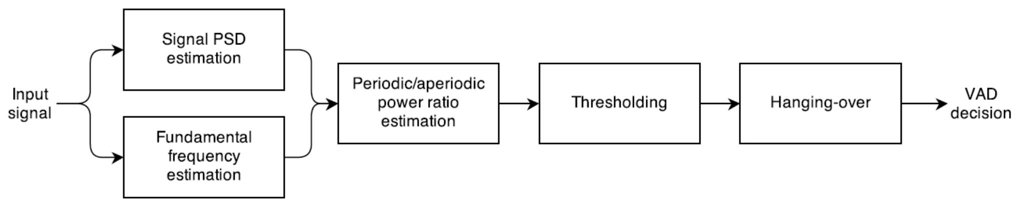

FIGURE 2.12: Block diagram of the periodic/aperiodoc component ratio VAD [8]

While this concept is likely to be robust to various non-stationary noise types, by definition it cannot cope with the unvoiced parts of speech, detection of which has to be performed by a hang-over scheme. In the original paper, authors did not specify how to estimate the number of harmonics ϑ for each frame, which is crucial to the proper working of the algorithm. A potential improvement might also come from an improved pitch detection method.

Another approach to using harmonic frequency components[1] for VAD has been investigated by Tan *et al.* in [27]. The authors claim, that under low SNR the approach from [28] is likely to fail since the LRs from the high-energy frames will not be strong enough to aid the proper classification of the low-energy frames. Therefore, they propose a new way of calculating the LRs for the voiced frames, defined in Equation 2.11 where ϑ is the number of harmonics within the resolution of the DFT and f_0 denotes the fundamental frequency. The algorithm calculates the LR only using the frequency bins which are multiples of the fundamental frequency.

$$\log \Lambda_v = \frac{1}{\vartheta} \sum_{m=1}^{\vartheta} \log \Lambda(m f_0) \tag{2.11}$$

Voiced frames are pre-identified by the pitch determination algorithm, and all others are considered as unvoiced. The LR for the unvoiced frames is calculated in a standard way from [29] i.e. by considering all frequency bins.

2.2.5 Summary

Voice Activity Detection has been studied by numerous researchers over the recent years. Some of the most simple features proposed in the literature, apart from energy, are based on the entropy of the signal, either in time or frequency domain.

The popular LRT based approach, initially proposed by Sohn *et al.*, has served as a basis for many researchers who tried to improve the original algorithm. While Sohn *et al.* employed the LRT to the SNR, the idea has also been applied to other features, such as the periodic to aperiodic component ratio. Nevertheless, many VAD algorithms still utilise the estimated SNR as a voicing feature. Ramirez *et al.* proposed one such VAD, where the SNR is calculated for a given frame including information from the neighbouring frames, rather than considering each frame independently.

The most noise-robust VAD algorithms are likely to be the ones which are based on the harmonicity of the speech signal. Since voiced speech typically does not present high energy in all frequency bins, it is beneficial to first identify the fundamental frequency of a signal and consider the power around the multiples of it. While this approach inherently cannot detect the aperiodic unvoiced phonemes, a clever hang-over scheme can

[1]This algorithm also builds on the idea of LRT described in section 2.2.2. The main improvement comes however from utilising the spectral harmonicity therefore it has been included in this section

help in their proper classification, by extending the initial VAD decisions to include such misdetected speech segments.

2.3 Conclusion

This chapter presented a literature survey of the most commonly encountered approaches to Voice Activity Detection. In section 2.1 some standard algorithms have been described while section 2.2 reviewed the recent research efforts in the VAD area.

In terms of noise-robustness, the standard algorithms are unlikely to achieve satisfactory performance under low SNR conditions since they are primarily based on features such as the energy or the zero-crossing rate which are easily degradable by high background noise levels. In such conditions, performance of the recently proposed VAD algorithms is expected to be much better.

Chapter 3

Experimental set-up

3.1 Motivation

A large number of VAD algorithms proposed in the literature and a lack of standard methods for their evaluation combined with an abundance of speech and noise corpora makes it difficult to compare the results reported in research papers. Therefore, identification of an algorithm whose performance is objectively best is not possible without their proper benchmarking. Another problem emerges from a variety of different hang-over schemes used in experiments from the literature. These schemes can greatly affect the final VAD decisions and their use is especially important with the harmonicity based features which rely exclusively on the hang-over scheme for detection of the unvoiced phonemes. In order to unify the effect of hanging-over, the same scheme should be applied to all VAD features. Finally, some algorithms require estimation of the same parameters before the actual decision feature can be computed. For the algorithms chosen in this evaluation, these include noise power spectrum and pitch estimation. In order to reduce the performance bias coming from this estimates, the same procedures should be implemented in all algorithms.

The first aim of this chapter is to describe the implementation details of the selected VAD algorithms so that the differences in their workings can be easily identified[1]. Secondly, the speech and noise recordings as well as the corresponding ground truth labels are described so that the experiments can be repeated. For speech, a subset of the TIMIT [4] database

[1]Ideally, they should differ only in the voicing features used for classification.

has been used. Noise recordings have been taken from the NOISEX-92 [5] database and added to the clean speech at various power levels.

3.2 VAD features

Due to a large number of VAD features proposed in the literature, implementation and evaluation of all of them is impractical. However, a small number of features which appear to be both noise-robust and based on different ideas can be selected for comparison. In this evaluation, the features from the following VAD algorithms have been implemented:

- A Statistical Model-Based Voice Activity Detector [18] (hereinafter referred to as 'Sohn')
- Efficient Voice Activity Detection Algorithms Using Long-term Speech Information [9] (hereinafter referred to as 'LTSD')
- Entropy Based Voice Activity Detection in Very Noisy Conditions [30] (hereinafter referred to as 'Entropy')
- Noise Robust Voice Activity Detection Based on Periodic to Aperiodic Component Ratio [8] (hereinafter referred to as 'PARADE')
- Voice Activity Detection using Harmonic Frequency Components in Likelihood Ratio Test [27] (hereinafter referred to as 'Harmfreq')

In particular, the LTSD feature has been chosen because it is calculated based on multiple frames surrounding the one for which the decision is being made. Harmfreq is similar the Sohn feature, but calculates the likelihood ratio for the voiced phonemes only at the multiples of the fundamental frequency. PARADE, on the other hand, is supposed to be robust to a variety of non-stationary noise types.

3.2.1 Shared implementation parts

Calculation of a number of VAD features requires prior estimation of some signal characteristics. In particular, Sohn, LTSD and Harmfreq need a noise power spectrum for the SNR estimate. Similarly, PARADE and Harmfreq require a pitch estimate in order to calculate the power of the signal at its multiples. In the original research papers either no procedure has been recommended (i.e. in Sohn [18] for noise PSD estimation) or different procedures have been suggested for estimation of the same property (i.e. pitch estimation in PARADE [8] and Harmfreq [27]). In order to reduce the bias from errors in the estimates, the same noise and pitch estimation procedures have been applied to all algorithms. For the noise PSD, an implementation of the MMSE noise power estimator [39] provided by VOICEBOX [41] has been used. Similarly, the fundamental frequency in PARADE and Harmfreq has been estimated using PEFAC [42]. Finally, before any processing, the incoming speech signal has been divided into 50 ms-long non-overlapping frames and windowed with a periodic Hanning window.

3.2.2 Algorithm-specific implementation parts

Apart from the same implementation parts used by different algorithms, there are parameters which are specific to some VAD features. Summary of the algorithm-specific implementation details is presented below:

- Sohn - the constant for decision-directed a priori SNR estimation is $\alpha = 0.95$
- LTSD - the LTSE lookup into neighbouring frames is $N = 3$. Note that it is a smaller value than $N = 6$ recommended in [9] due to a longer frame length and a lack of the overlap factor
- PARADE - the number of harmonics is $\vartheta = 10$ or the maximum that fit in the resolution of the DFT
- Harmfreq - $\alpha = 0.95$ (same as in Sohn). $\vartheta = 10$ or the maximum that fit in the resolution of the DFT (same as in PARADE). A frame is treated as voiced if the probability returned by PEFAC is greater than 0.5

3.3 Hang-over scheme

In order to reduce the differences in VAD decisions coming from different hang-over schemes, the same method has been applied to all evaluated VAD features, pseudo code of which is presented in Algorithm 1. This is a modified version of the scheme originally published by ETSI in [43]. The main idea behind the scheme is to initialise the so-called *hang-over timer* whenever the maximum number of consecutive speech decisions in a buffer of B frames is greater than S_p (speech possible) or S_l (speech likely). The proposed value for B by ETSI is 7 and the same has been used in this implementation. However, the values for parameters S_p, S_l, L_s and L_m were made smaller than the ones outlined in the standard, since the length of the processed signal frames is longer (i.e. 50 ms) and there is no overlap between them.

Algorithm 1 Hang-over scheme used in all VAD algorithms

INPUT: F - number of frames
INPUT: V - VAD decisions prior to the hang-over scheme
OUTPUT: hV - VAD decisions after the hang-over scheme

$B \leftarrow 7$ {buffer length}
$S_p \leftarrow 2$ {speech possible}
$S_l \leftarrow 3$ {speech likely}
$L_s \leftarrow 5$ {short hang-over time}
$L_m \leftarrow 8$ {medium hang-over time}

$T \leftarrow 0$ {hang-over timer}
for $i = 1$ to $F - B + 1$ **do**
  $CB \leftarrow V(i : i + B - 1)$ {current buffer of VAD decisions}
  $M \leftarrow$ maximum consecutive speech-present decisions in CB
  if $M \geqslant S_l$ **then**
    $T \leftarrow L_m$
  else if $M \geqslant S_p$ **and** $T < L_s$ **then**
    $T = L_s$
  else if $M < S_p$ **and** $T > 0$ **then**
    $T \leftarrow T - 1$
  end if
  if $T > 0$ **then**
    $hV(i) \leftarrow 1$
  else
    $hV(i) \leftarrow 0$
  end if
end for
$hV(F - B + 2 : F) \leftarrow V(F - B + 2 : F)$ {assign the pre hang-over decisions to the last B-1 frames}
return hV

3.4 Speech corpus

There exists a number of speech corpora for evaluation of VAD and ASR. For the English language the TIMIT speech corpus [4] seems to be the most widely used and hence has been selected as a source of utterances for this evaluation. There are a number of issues with the original recordings which need to be resolved with so that the evaluation can be more precise. There are:

1. Single utterances are on average no longer than 4 seconds and hence too short for proper VAD evaluation
2. The recordings contain a very small number of non-speech segments
3. There are differences in average power between the utterances

In order to deal with the first two issues, a number of randomly chosen utterances from every dialect (i.e. DR1 to DR8) has been concatenated into a single speech recording, adding 2.5 seconds of silence at the beginning, ending and between the utterances. In an attempt to equalise the energy, the amplitude of each utterance has been normalised such that the average power per sample of all of them is equal. This is important because if the concatenated utterances differed in the average power level, after adding the noise, the one with initially higher power could be easily detected by almost any VAD procedure, while the detection of the other one would be very difficult due to the much lower short-time SNR.

After concatenation, the ground truth speech/non-speech labels were obtained by a combination of a the output from a simple energy based VAD with hand labelling the small number of misclassified segments. Figure ... shows an example of the energy VAD and the final VAD labels modified by hand.

3.5 Noise corpus and SNR

Chapter 4

Evaluation of VAD algorithms

4.1 Main Section 1

4.2 Main Section 2

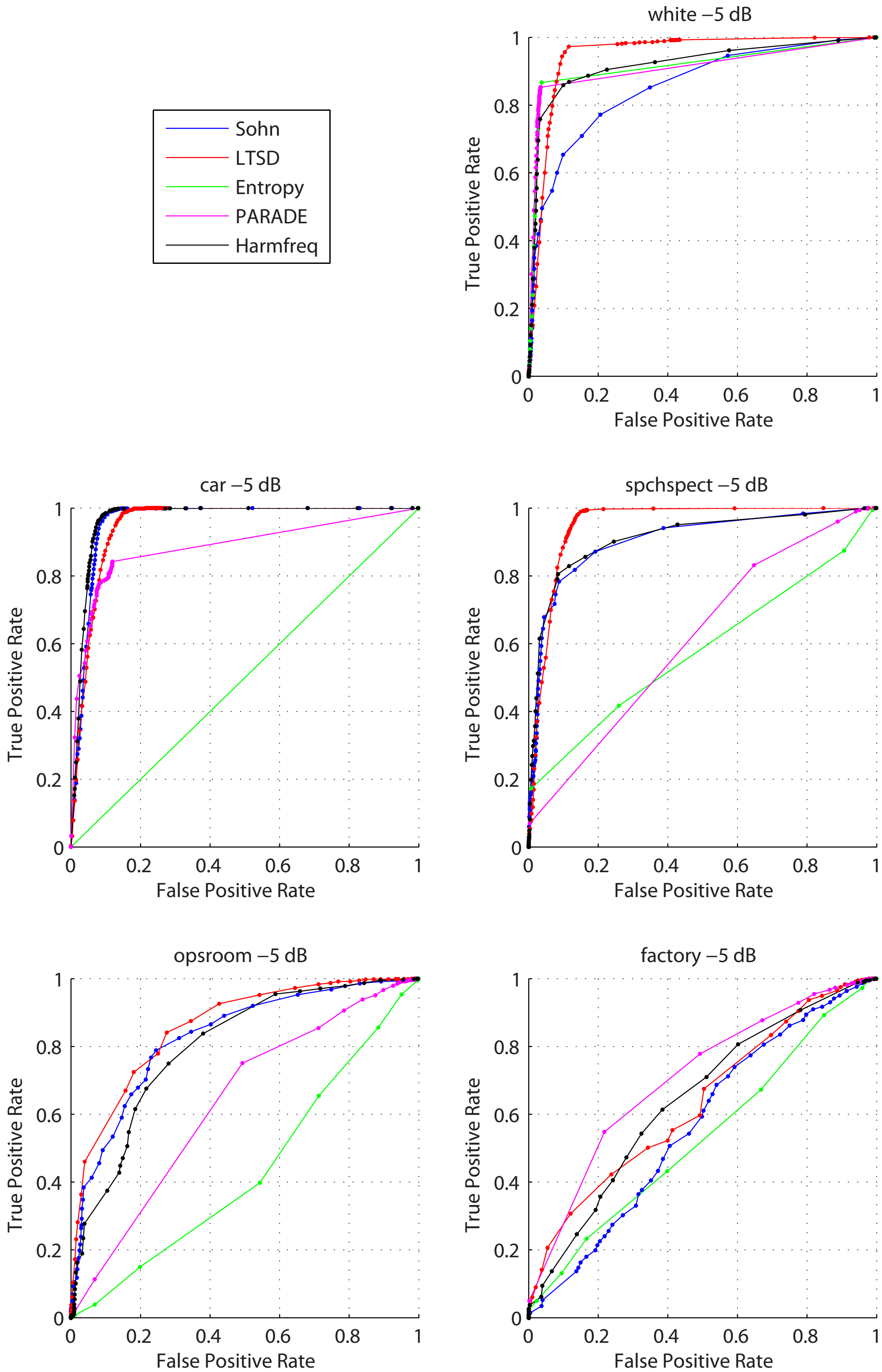

FIGURE 4.1: ROC curves of the evaluated VAD algorithms *with* hang-over under -5 dB SNR

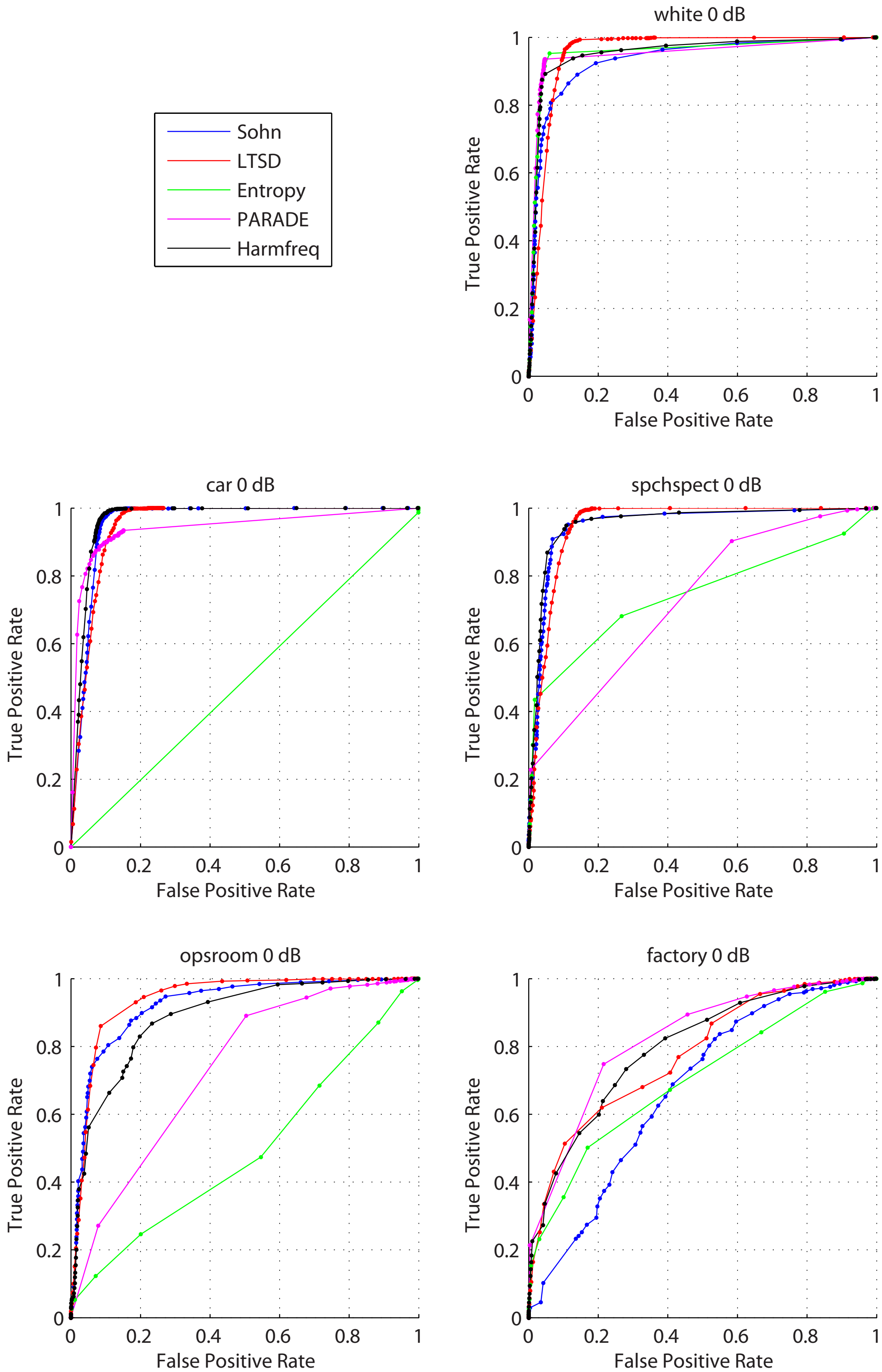

Figure 4.2: ROC curves of the evaluated VAD algorithms *with* hang-over under 0 dB SNR

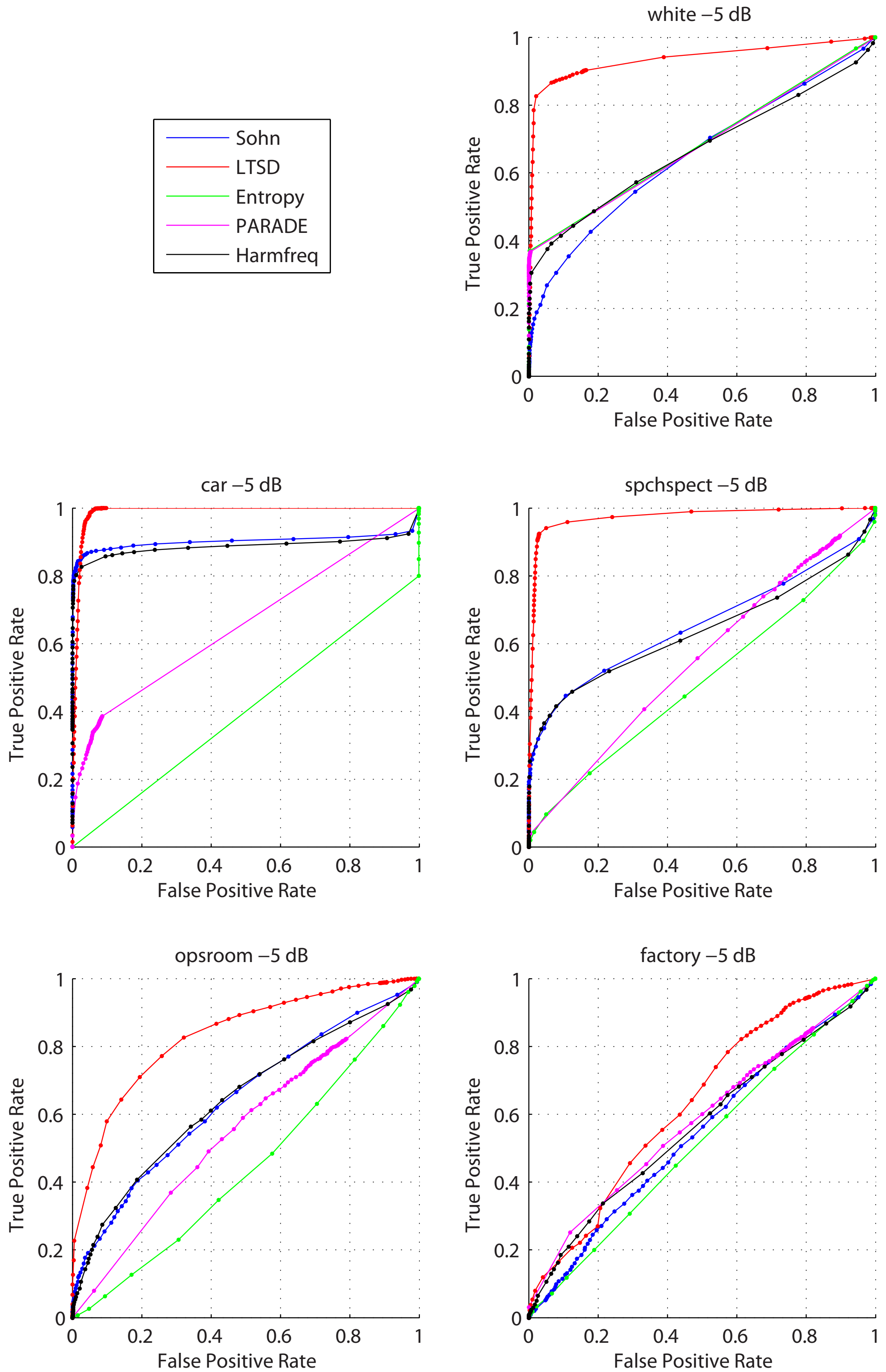

Figure 4.3: ROC curves of the evaluated VAD algorithms *without* hang-over under -5 dB SNR

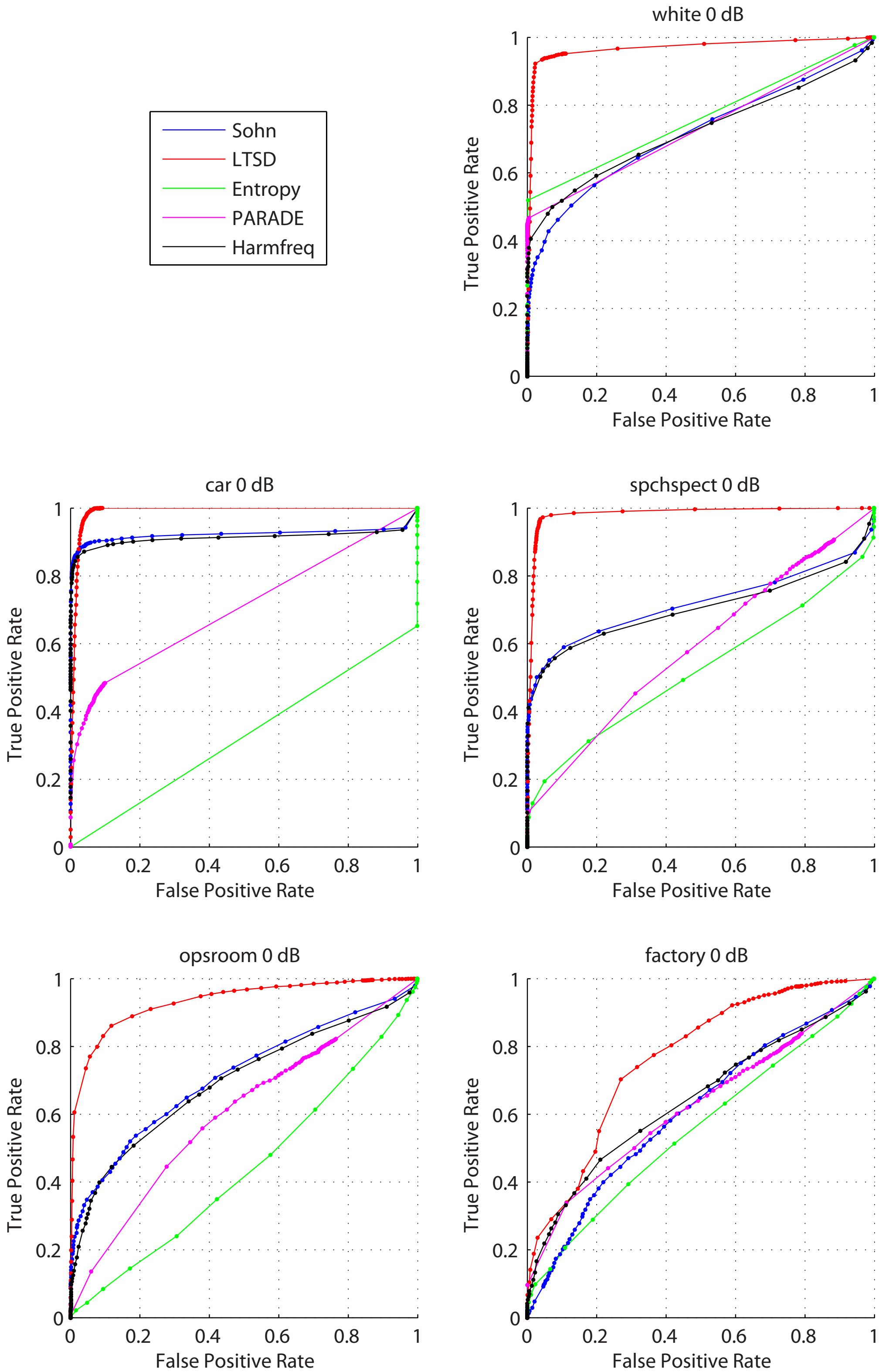

FIGURE 4.4: ROC curves of the evaluated VAD algorithms *without* hang-over under 0 dB SNR

Appendix A

Additional evaluation results

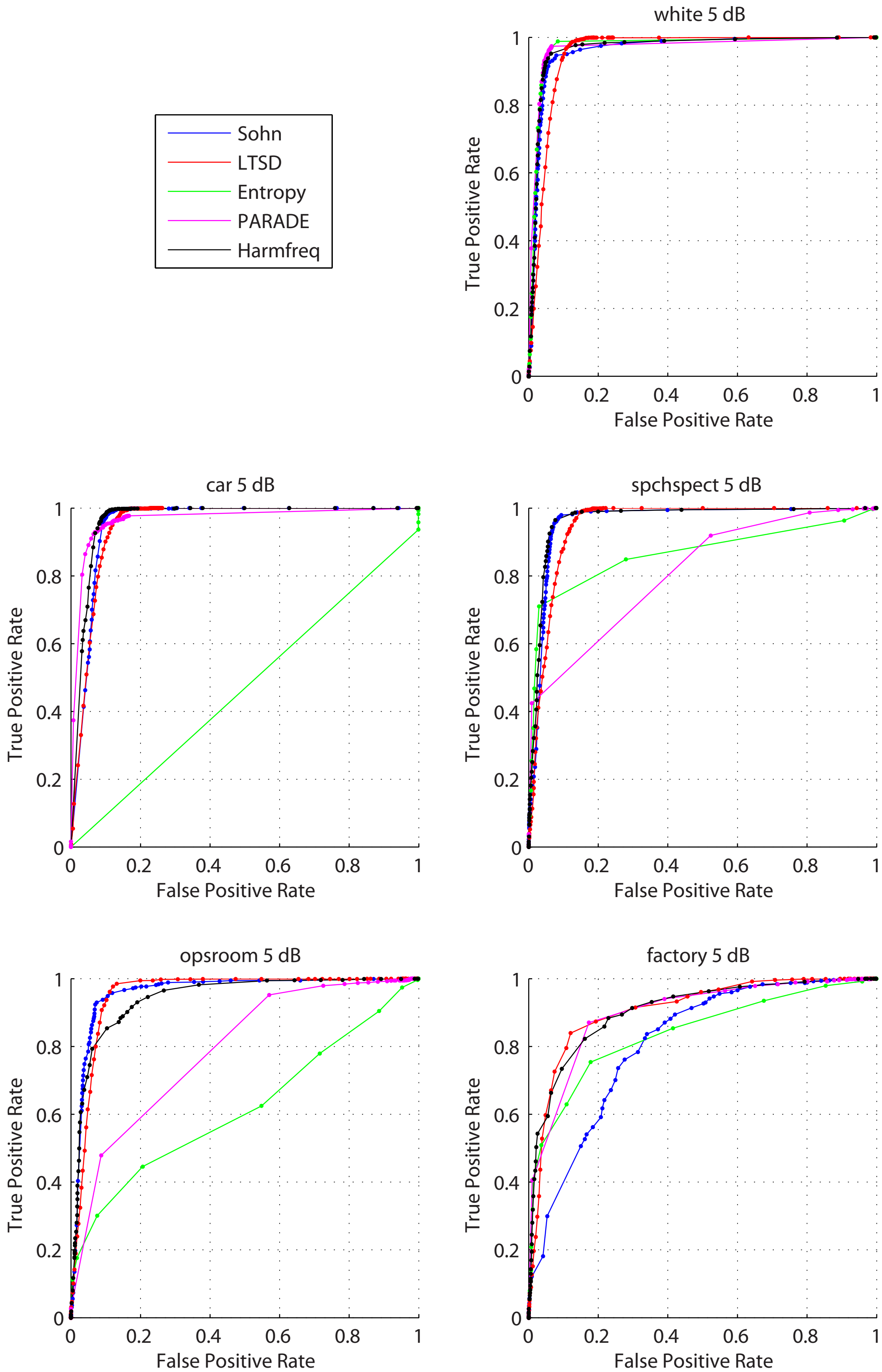

FIGURE A.1: ROC curves of the evaluated VAD algorithms *with* hang-over under 5 dB SNR

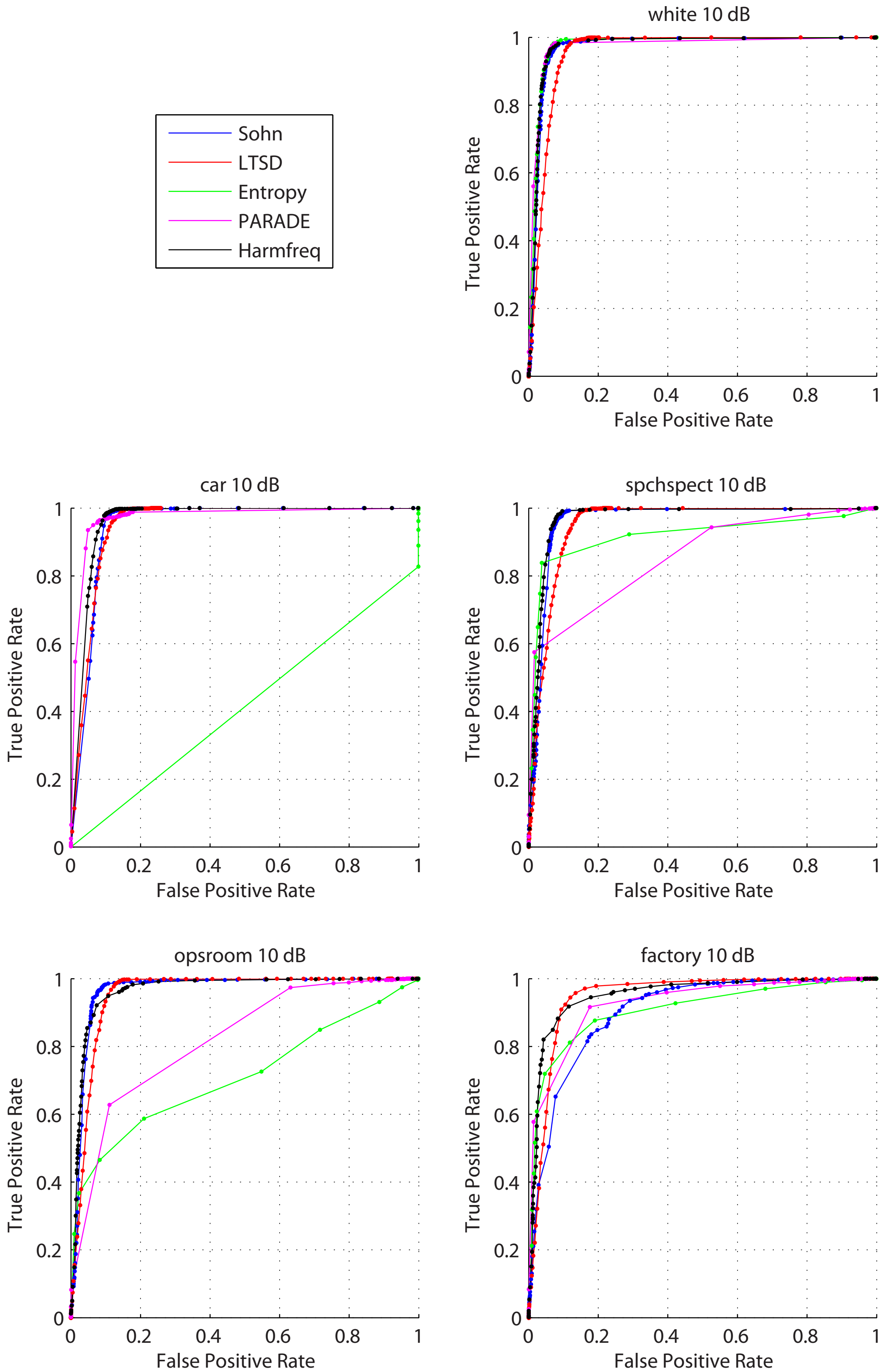

Figure A.2: ROC curves of the evaluated VAD algorithms *with* hang-over under 10 dB SNR

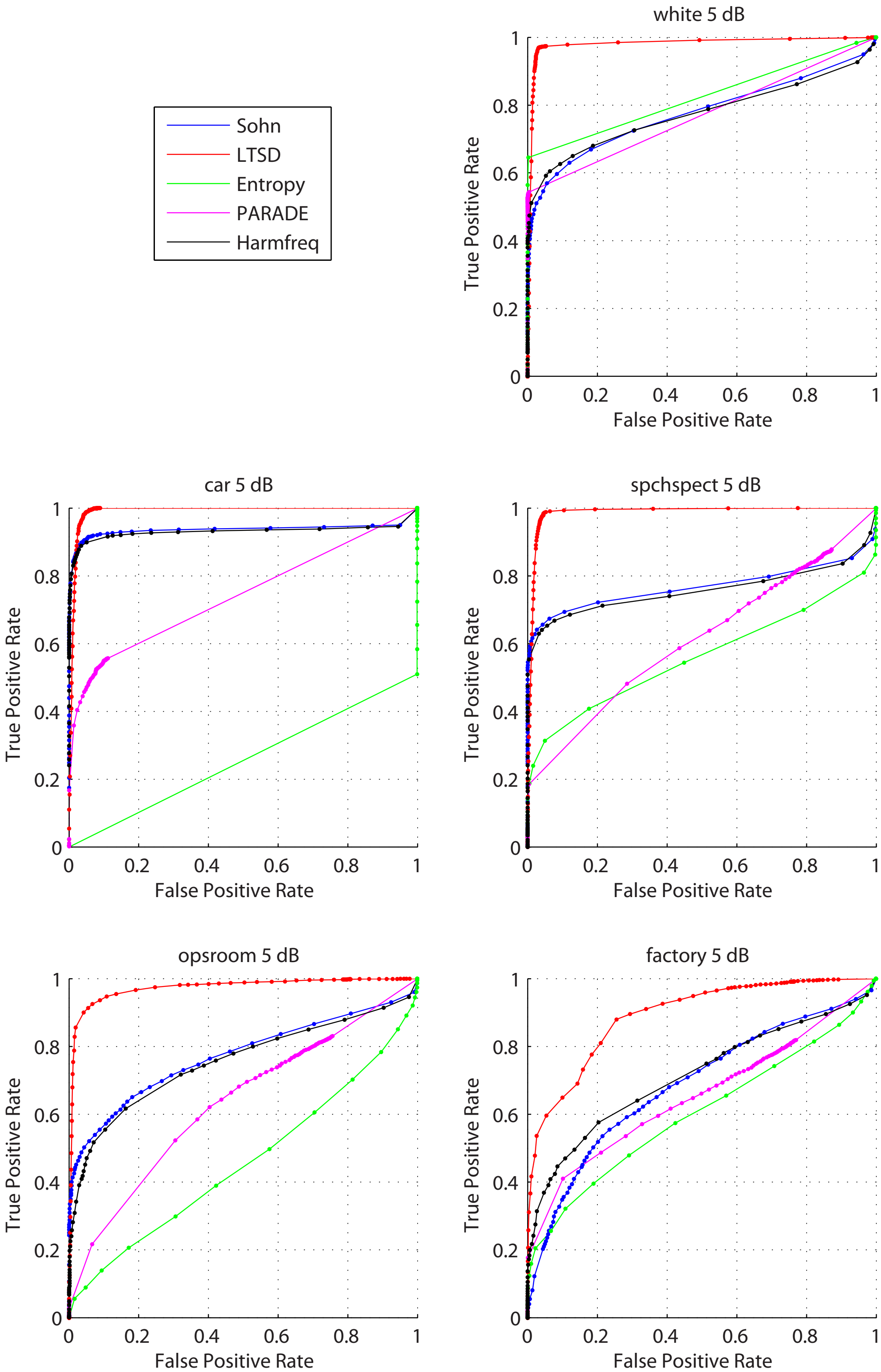

FIGURE A.3: ROC curves of the evaluated VAD algorithms *without* hang-over under 5 dB SNR

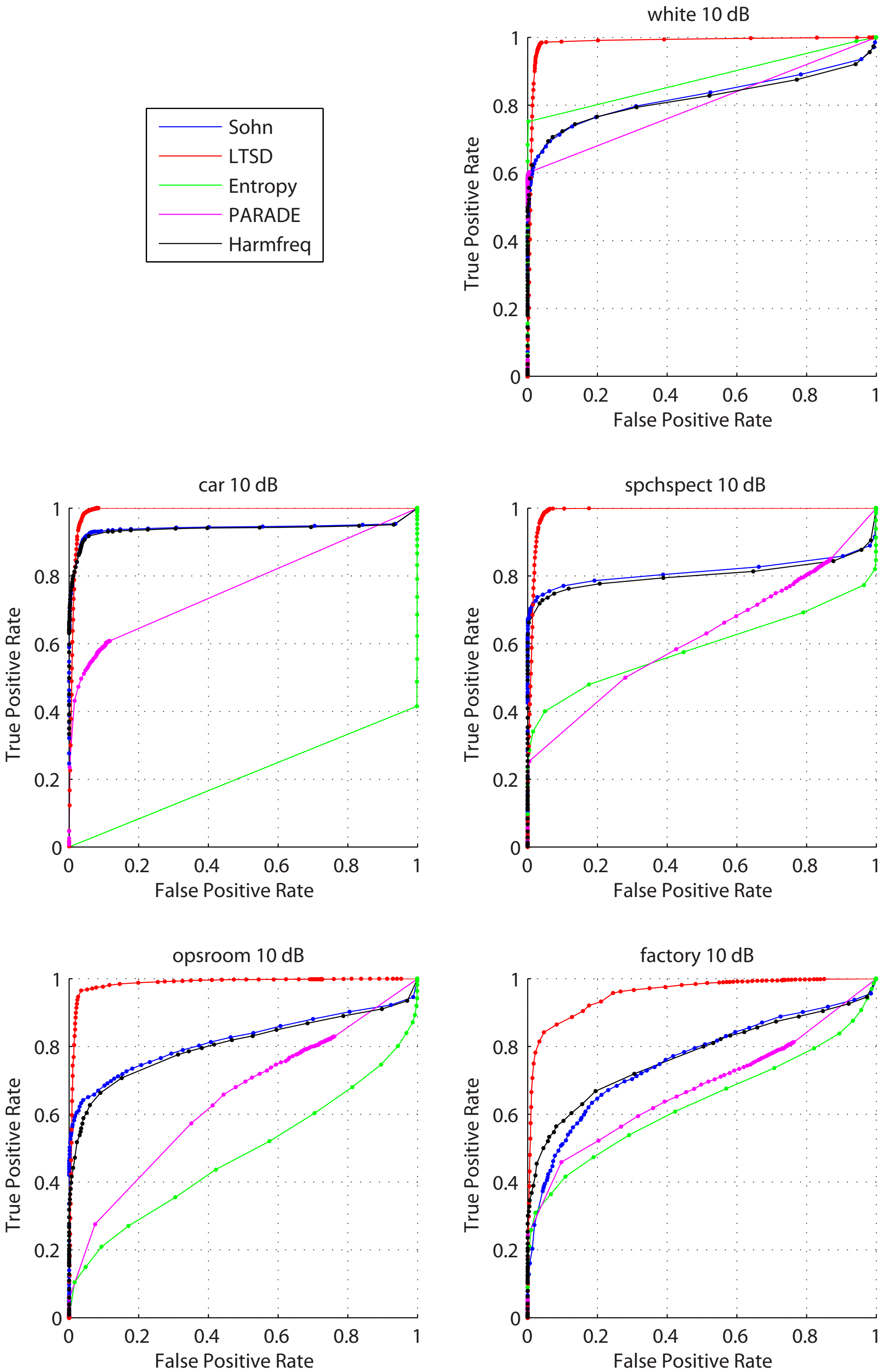

FIGURE A.4: ROC curves of the evaluated VAD algorithms *without* hang-over under 10 dB SNR

Bibliography

[1] A. M. Kondoz. *Digital Speech. Coding for Low Bit Rate Communication Systems.* John Wiley & Sons, 2004.

[2] P. R. Michaelis. Human Speech Digitization and Compression. In W. Karwowski, editor, *International Encyclopedia of Ergonomics and Human Factors.* CRC Press, 2006.

[3] L. R. Rabiner and M. R. Sambur. An Algorithm for Determining the Endpoints of Isolated Utterances. *The Bell System Technical Journal*, February 1975.

[4] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993.

[5] A. Varga and H. 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 Volume 12 Issue 3*, 1993.

[6] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. 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*, 1997.

[7] European Telecommunications Standards Institute. Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels; General Description (GSM 06.94 version 7.1.0 Release 1998), 1999.

[8] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki. Noise robust voice activity detection based on periodic to aperiodic component ratio. *Speech Communication Volume 52 Issue 1*, 2010.

[9] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication Volume 42 Issues 3-4*, 2004.

[10] Telecommunication Standardization Sector International Telecommunication Union. ITU-T Recommendation G.729 - coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP), 2012.

[11] K. Li, M. N. S. Swamy, and M. O. Ahmad. An Improved Voice Activity Detection Using Higher Order Statistics. *IEEE Transactions On Speech And Audio Processing Volume 13 Issue 5*, 2005.

[12] R. Tucker. Voice activity detection using a periodicity measure. *IEE Proceedings on Communications, Speech and Vision Volume 139 Issue 4*, 1992.

[13] J. Ramirez, J. M. Gorriz, and J. C. Segura. *Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding.* InTech, 2007.

[14] S. Kuroiwa, M. Naito, S. Yamamoto, and N. Higuchi. Robust Speech Detection Method for Telephone Speech Recognition System. *Speech Communication Volume 27 Issue 2*, 1999.

[15] A. Martin and L. Mauuary. Robust Speech/Non-Speech Detection Based on LDA-Derived Parameter and Voicing Parameter for Speech Recognition in Noisy Environments. *Speech Communication Volume 48 Issue 2*, 2006.

[16] I. Shafran and R. Rose. Robust Speech Detection and Segmentation for Real-Time ASR Applications. *International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[17] J. M. Gorriz, J. Ramirez, E. W. Lang, G. C. Puntonet, and I. Turias. Improved Likelihood Ratio Test Based Voice Activity Detector Applied to Speech Recognition. *Speech Communication Volume 52 Issues 7-8*, 2010.

[18] J. Sohn, N. S. Kim, and W. Sung. A Statistical Model-Based Voice Activity Detection. *IEEE Signal Processing Letters*, January 1999.

[19] R. Venkatesha Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav. Comparison of Voice Activity Detection Algorithms for VoIP. *Seventh International Symposium on Computers and Communications*, 2002.

[20] C. B. Southcott, D. Freeman, G. Cosier, D. Sereno, A. van der Krogt, A. Gilloire, and H. J. Braun. Voice Control of the Pan-European Digital Mobile Radio System. *Global Telecommunications Conference and Exhibition 'Communications Technology for the 1990s and Beyond' (GLOBECOM)*, 1989.

[21] Y. Park and S. Lee. Speech enhancement through voice activity detection using speech absence probability based on Teager energy, 2013.

[22] K. R. Borisagar, D. G. Kamdar, B. S. Sedani, and G. R. Kulkarni. Speech Enhancement in Noisy Environment Using Voice Activity Detection and Wavelet Thresholding. *IEEE International Conference on Computational Intelligence and Computing Research*, 2010.

[23] M. Sahidullah and G. Saha. Comparison of Speech Activity Detection Techniques for Speaker Recognition, 2012.

[24] Y. Kida and T. Kawahara. Voice Activity Detection based on Optimally Weighted Combination of Multiple Features. *9th European Conference on Speech Communication and Technology INTERSPEECH 2005*, 2005.

[25] K. Weaver, K. Waheen, and F. M. Salem. An Entropy based Robust Speech Boundary Detection Algorithm for Realistic Noisy Environments. *Proceedings of the International Joint Conference on Neural Networks, Volume 1*, 2003.

[26] Z. Tuske, P. Mihajlik, Z. Tobler, and T. Fegyo. Robust Voice Activity Detection Based on the Entropy of Noise-Suppressed Spectrum. *9th European Conference on Speech Communication and Technology INTERSPEECH 2005*, 2005.

[27] L. N. Tan, B. J. Borgstrom, and A. Alwan. Voice Activity Detection using Harmonic Frequency Components in Likelihood Ratio Test. *International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[28] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio. Statistical Voice Activity Detection Using a Multiple Observation Likelihood Ratio Test. *IEEE Signal Processing Letters Volume 12*, 2005.

[29] J. Sohn and W. Sung. A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation. *International Conference on Acoustics, Speech and Signal Processing*, 1998.

[30] P. Renevey and A. Drygajlo. Entropy Based Voice Activity Detection in Very Noisy Conditions. *Eurospeech 2001*, 2001.

[31] R. Kotcher and K. Monteith. Noise-resilient speech segmentation using the Voting Experts algorithm, 2013.

[32] X-L. Zhang and J. Wu. Denoising Deep Neural Networks based Voice Activity Detection, 2013.

[33] M. Stadtschnitzer, T. V. Pham, and T. T. Chien. Reliable Voice Activity Detection Algorithms Under Adverse Environments, 2008.

[34] Telecommunications Industry Association/Electronic Industries Alliance. High Rate Speech Service Option 17 for Wideband Spread Spectrum Communication Systems, 1997.

[35] E. Cornu, H. Sheikhzadeh, R. L. Brennan, H. R. Abutalebi, E. C. Y. Tam, P. Iles, and K. W. Wong. ETSI AMR-2 VAD: Evaluation And Ultra Low-Resource Implementation. *International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[36] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal Volume 27*, 1948.

[37] Y. D. Cho, K. Al-Naimi, and A. Kondoz. Improved voice activity detection based on a smoothed statistical likelihood ratio. *International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[38] M. Rainer. Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics. *IEEE Transactions on Speech and Audio Processing Volume 9 Issue 5*, 2001.

[39] T. Gerkmann and R. C. Hendriks. Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay. *IEEE Transactions on Audio, Speech and Language Processing Volume 20 Issue 4*, 2012.

[40] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1984.

[41] M. Brookes. VOICEBOX: Speech Processing Toolbox for MATLAB, 2014. URL `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[42] S. Gonzalez and M. Brookes. PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing Volume 22 Issue 2*, 2014.

[43] European Telecommunications Standards Institute. ETSI ES 202 050 v 1.1.3 Speech Processing, Transmission and Quality Aspects; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, 2003.