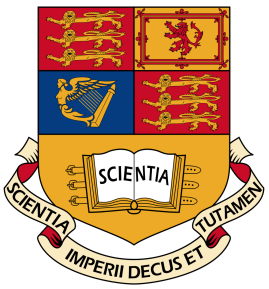

Imperial College London

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

December 2013

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

Declaration of Authorship i

Abstract ii

Acknowledgements iii

Contents iv

List of Figures v

List of Tables vi

1 Introduction 1
- 1.1 Voice Activity Detection . . . 1
- 1.2 Applications of VAD . . . 2
  - 1.2.1 Automatic Speech Recognition . . . 3
  - 1.2.2 Speech Coding and Transmission . . . 3
  - 1.2.3 Noise Estimation and Speech Enhancement . . . 4
- 1.3 Structure of a typical VAD system . . . 5
  - 1.3.1 Pre-processing . . . 6
  - 1.3.2 Feature Extraction . . . 6
  - 1.3.3 Classification . . . 6
  - 1.3.4 Post-processing . . . 6
- 1.4 Thesis organisation . . . 7

2 Literature Survey of the VAD algorithms 8
- 2.1 Standard VAD algorithms . . . 8
  - 2.1.1 ITU-T G.729 Annex B . . . 9
  - 2.1.2 ETSI AMR1 and AMR2 . . . 10
  - 2.1.3 TIA/EIA IS-733 . . . 11
- 2.2 Other VAD algorithms . . . 11

Bibliography 12

List of Figures

1.1 A sample utterance corrupted by -5 dB car noise 2
1.2 Automatic Speech Recognition system with Voice Activity Detection module 3
1.3 Dual-mode transmission system with Voice Activity Detection module . . 4
1.4 Block diagram of a typical Voice Activity Detection system 5

2.1 Block diagram of the ITU-T G.729 Annex B Voice Activity Detector . . . 9
2.2 Block diagram of the ETSI AMR Option 1 Voice Activity Detector 11

List of Tables

2.1 Cut-off frequencies for the ETSI AMR1 band-pass filters 11

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 that we are not interested in. Additionally, VAD decision is especially difficult for the unvoiced phonemes [1] whose spectrum contains no periodicity and is often similar to the one of white noise [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. This 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 Voice over IP, 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 picked-up 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.

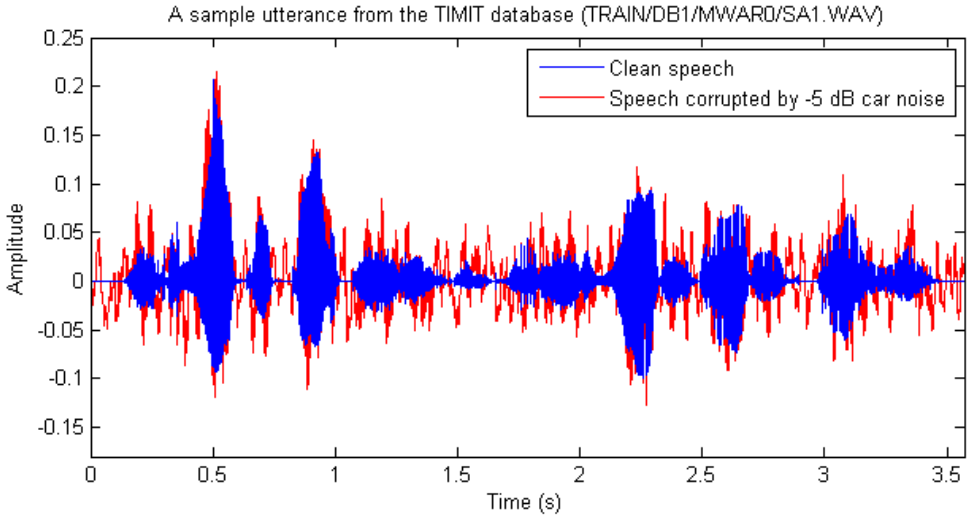

FIGURE 1.1: A sample utterance from the TIMIT speech corpus corrupted by -5 dB car noise from the NOISEX-92 database

Recently, numerous VAD approaches have been proposed, based on various features such as ....

1.2 Applications of VAD

VAD is often the first step in many signal processing applications including speech recognition [4–8], speech coding and transmission [4, 9–12], speech enhancement [4, 13, 14], noise estimation [4] or speaker recognition [15]. 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 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 [4]. 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 ASR 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.

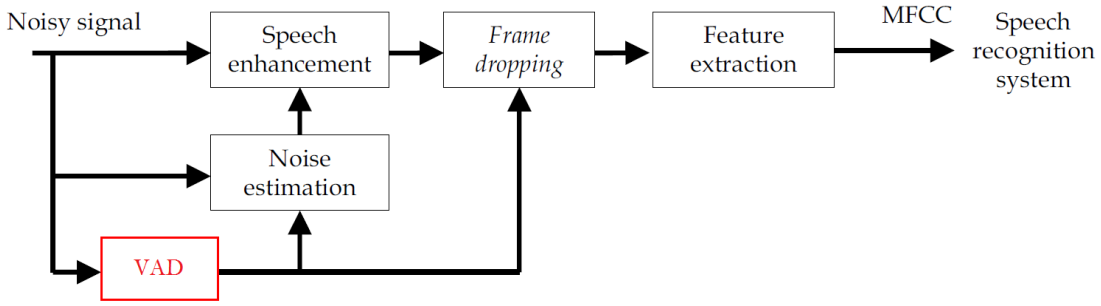

FIGURE 1.2: Block diagram of an Automatic Speech Recognition system with Voice Activity Detection module [4]

1.2.2 Speech Coding and Transmission

A typical phone conversation involves each person speaking on average no more than 50% of the time [12]. Using this fact it can be concluded that 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 Detection 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 [11] 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 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* [4, 11, 12] 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.

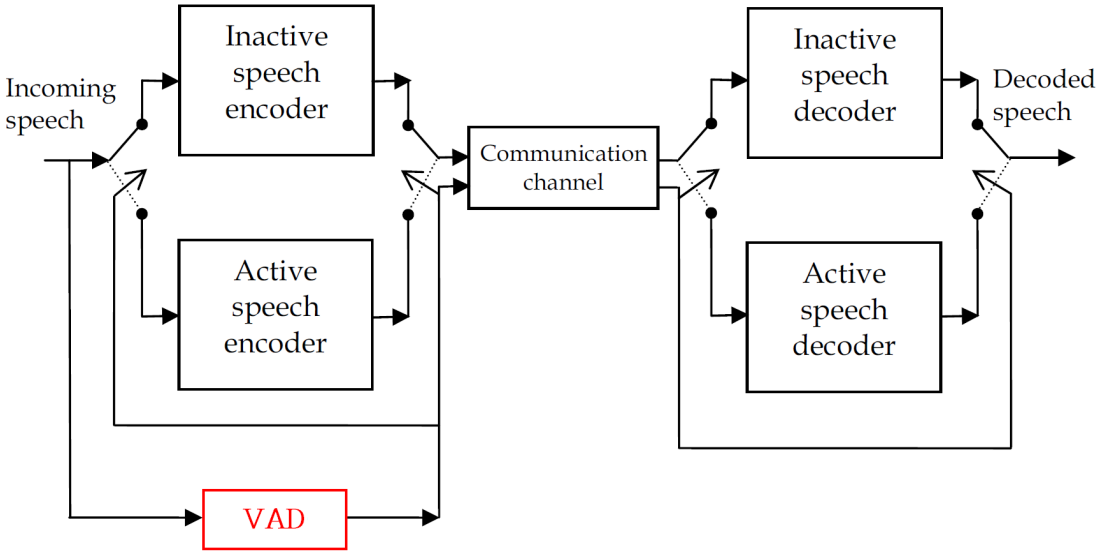

FIGURE 1.3: Block diagram of an dual-mode transmission system with Voice Activity Detection module [11]

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, 4]. 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, the noise needs to be additive, stationary and uncorrelated with the clean speech signal. Additionally, 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 nontrivial problem. 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.

1.3 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.

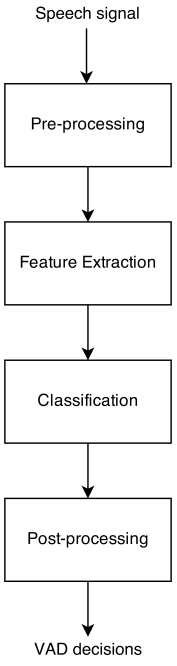

FIGURE 1.4: Block diagram of a typical Voice Activity Detection system

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. The module might perform noise estimation and suppression in the signal in order to improve performance of the VAD. Additionally, during pre-processing the input signal is often split into frames which are typically 10-50 ms long.

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 for the speech/non-speech classification. These can be time-domain [ADD REFERENCE], frequency-domain [ADD REFERENCE], cepstral-domain [ADD REFERENCE] features or other, depending on the specific VAD algorithm. Ideally, in order to acheve 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 variety of decision rules, ranging from simple thresholding [ADD REFERENCE] to more advanced methods such as statistical likelihood ratio tests [ADD REFERENCE] or machine learning [ADD REFERENCE]. 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 [ADD REFERENCE] 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 in order to reduce the number of false positives and false negatives. For example, 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.

1.4 Thesis organisation

The rest of this document is organised as follows:

- Chapter 2 contains a comprehensive 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 presents a thorough evaluation of selected VAD algorithms from Chapter 2 on the TIMIT speech corpus corrupted by various noise types from the NOISEX-92 database
-

Chapter 2

Literature Survey of the 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). It is important to note that most of 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.

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

- ITU-T G.729 Annex B [11] 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 [16] for application to the Global System for Mobile Communications (GSM)
- TIA/EIA IS-733 [17] 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 [18] 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)

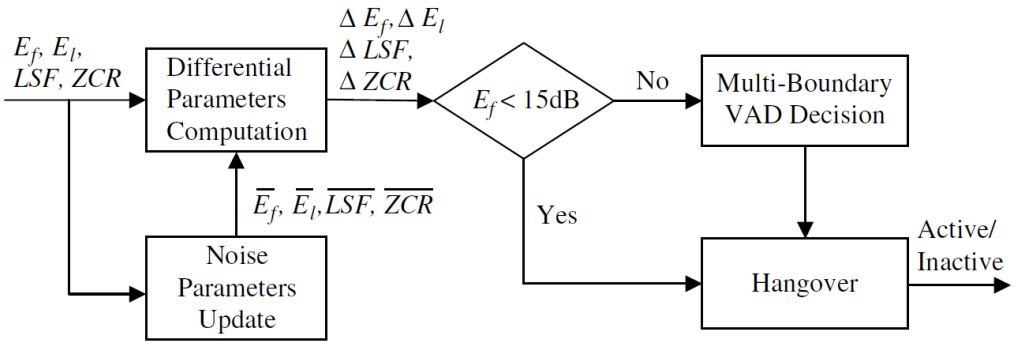

FIGURE 2.1: Block diagram of the ITU-T G.729 Annex B Voice Activity Detector [1]

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 $E_f < 15dB$ 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 VAD algorithm also performs updates 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 extremely robust since it is used only for noise parameters estimation.

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 to be classified across different frequency bands.

The block diagram of the AMR Option 1 VAD is presented in Figure 2.2. The input signal is first passed through a series of band-pass filters which split the time-domain signal into different frequency bands based on the Table 2.1.

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

TABLE 2.1: Cut-off frequencies for the ETSI AMR1 band-pass filters [16]

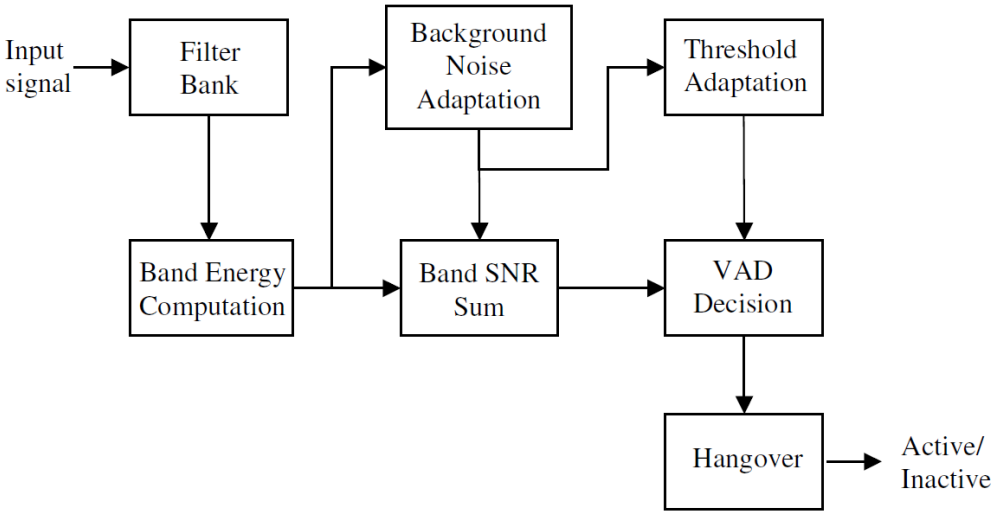

FIGURE 2.2: Block diagram of the ETSI AMR Option 1 Voice Activity Detector [1]

2.1.3 TIA/EIA IS-733

2.2 Other VAD algorithms

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

[5] S. Kuroiwa, M. Naito, S. Yamamoto, and N. Higuchi. Robust Speech Detection Method for Telephone Speech Recognition System. *Speech Communication*, 1999.

[6] 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*, 2006.

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

[8] 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*, 2010.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] 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.

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

[18] 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.