Far-Field Voice Activity Detection and Its Applications in Adverse Acoustic Environments

Petsatodis, Theodoros

Publication date:
2012

Document Version
Early version, also known as pre-print

Far-Field Voice Activity Detection and Its Applications in Adverse Acoustic Environments

Petsatodis Theodoros
Department of Electronic Systems
Aalborg University

A Dissertation Submitted to the Department of Electronic Systems and the Committee on Graduate Studies of Aalborg University in Partial Fulfilment of the Requirements for the Philosophy Doctor (PhD) in Wireless Communications

Denmark, 2012

Academic Supervisor: Professor Dr. Ramjee Prasad, Aalborg University

Co-Supervisors: Professor Dr. Fotios Talantzis, Athens Information Technology
Professor Dr. Christos Boukis, Accenture Interactive Greece
Professor Dr. Zheng-Hua Tan, Aalborg University

Day of the defense: November $30^{th}, 2012$

Assessment Committee: Professor Sergios Theodoridis,
University of Athens, Greece

Associate Professor Gerasimos Potamianos,
University of Thessaly, Greece

Associate Professor Ove Andersen, (Chairman)
Aalborg University, Denmark

This dissertation is a result of the collaboration between
CTiF, Aalborg University (AAU) and Athens Information Technology (AIT).

Abstract

Voice Activity Detection (VAD), being in the focus of speech processing research for many years, is nowadays a mature technology with application in several sectors. Embedded VAD components in telecommunications systems (like in cellular telephony) attempt to reduce power consumption of transmitters and bandwidth utilization. VAD technology is also integrated in speech-processing systems, such as Speaker Identification, Automatic Event Detection, and Automatic Speech Recognition, to prevent their operation in the absence of speech, and thus reduce the error rates of each of these systems.

The performance of VAD systems depends strongly on various factors, including the discriminative ability of the classification criterion employed, the dynamics of the additive noise and the signal to noise ratio. Speech signals transmitted within reverberant enclosures and captured using far-field microphones are subject to reverberation effects, competitive sound sources, and speaker movement. Furthermore, speech distribution varies with time and can be affected by several unpredictable factors including speaker's temper, mood, gender, age, and more. Thus, during the design phase of a VAD, special considerations have to be taken in order to build a robust system able to operate under variable and adverse conditions.

Given that for most of speech processing systems it is of crucial importance to have a reasonable approximation for the probability density function (pdf) of speech, understanding the properties of speech distribution plays a very important role in the design of speech processing systems. Within the framework of this work, variability of speech distribution, when using far-field microphones, is analysed under the presence of noise and reverberation.

Observations of how speech distribution is shaped by external interferences are then used as the basis to develop an adaptive unsupervised VAD scheme. This VAD, in contrary to other approaches employing fixed distribution assumptions, relies on effectively modelling the distribution of speech as convex combination of a Gaussian, a Laplacian, and a two-sided Gamma distribution. The increased adaptability of the system along with the encapsulated adaptive threshold allows the system to perform remarkably under adverse complex phenomena.

Following recent technological trends, of incorporating microphone arrays in numerous commercial applications (eg. mobile phones, VOIP terminals) and research environments (smart rooms), a multiple microphone VAD is also considered. The system processes signals captured by far-field sensors in order to integrate spatial information in addition to the frequency content available at a single sensor. The core of the system resides on the modification of a multiple observation hypothesis, testing at each sensor the probability of having an active speaker and then fusing the decisions. The VAD operates without the need of Direction-of-Arrival (DOA) estimation and eliminates additional delay imposed by previous multi-microphone VAD technologies.

The system developed for the multi-microphone VAD serves as the platform to merge VAD with a very powerful analysis framework namely, the Empirical Mode Decomposition (EMD). This highly efficient method relies on local characteristics of time scale of the data to analyse and decompose non-stationary signals into a set of so called intrinsic mode functions (IMF). These functions are injected to the multiple microphone VAD scheme in order to decide upon speech presence or absence. The outcome of this procedure demonstrates significantly enhanced performance compared to single microphone approaches.

Speech distribution information is also encapsulated in a supervised VAD scheme. Operating in the far-field, the core of the system employs Hidden Markov Models the states of which are modelled using Gaussian Mixture Models to cater for the dynamics of captured speech. Given the bi-modality of speech production, a simple visual-VAD is also developed to examine performance enhancement when fusing audio and video information.

In the final part of the work, applications of VAD in the context of integration with other signal processing systems are also considered. Performance benefits of combining the multi-microphone VAD with DOA estimation are demonstrated. Optimization through adaptation of speech shape characteristics in the embedded Time Delay Estimation (TDE) scheme is also considered, the same way that was beneficial for the convex combination based VAD. Towards this direction, the underlying assumption of Gaussian distributed source is replaced by that of Generalized Gaussian distribution that allows the evaluation of the problem under a larger set of speech-shaped distributions, ranging from Gaussian to Laplacian and Gamma. The analysis performed, revealed a significant research outcome.

Furthermore, performance enhancement when using VAD in combination with noise reduction systems is also discussed in terms of residual suppression within silence intervals. For this scope, a noise reduction architecture has been developed based on cascading an one-pass scheme.

The final application of VAD, examined in the thesis, is in the area of biomedical signal processing. A modification of one of the VAD systems developed, is employed to provide preliminary detection of one of the major breathing-related sleep disorders, apnea. The idea behind the development of this system is the capability of unobtrusively monitoring patients at home, improving the reliability of detection of sleep disorders in home environments, offering comfort and time saving to patients.

English-Danish Short Summary

This thesis considers far-field Voice Activity Detection (VAD) and its applications in adverse acoustic environments. In the first part of the thesis, speech distribution variability under external interferences is investigated. The study forms the basis for the development of an unsupervised VAD based on the convex combination of a set of prime distributions, able to robustly operate under adverse conditions. The work continues with the design of a multi-microphone VAD that encapsulates spatial, apart from frequency and time, information. The multi-microphone VAD is later used in combination with a powerful data analysis method towards achieving optimal performance. The speech distribution information is also encapsulated in a supervised VAD scheme employing Hidden Markov Models, the states of which are modelled using Gaussian Mixture Models to cater for the dynamics of the captured speech. Additionally, a visual-VAD is developed and fused with the audio-based one. In the last part of the thesis, applications of VAD are discussed. Time Delay Estimation is investigated in depth towards encapsulating speech distribution information. Noise reduction in combination with VAD is also discussed. Finally, a modified VAD is employed, in the field of biomedical signal processing, to provide preliminary detection of apnea.

Denne afhandling omhandler detektering af stemmeaktivitet (Voice Activity Detection (VAD)) i fjernfeltet og dens anvendelse i ugunstige akustiske miljøer. I den første del af afhandlingen undersøges talefordelings variabilitet under eksterne interferenser. Undersøgelsen danner basis for udvikling af en ikke-overvåget VAD, der er baseret på den konvekse kombination af et sæt distributioner og i stand til robust at operere under ugunstige betingelser. Arbejdet fortsætter med designET af et multi-mikrofon VAD-system, der indkapsler rumlig information bortset fra frekvensen og tiden. Multimikrofonsystem VADen anvendes senere i kombination med en kraftfuld dataanalysemetode til at opnå optimal ydelse. Talefordelings oplysninger er også indkapslet i en overvåget VAD anordning, der anvender Hidden Markov Modeller, hvis tilstande er modelleret ved hjælp af Gaussiske miksturmodeller for at tage højde for dynamikken i den optagede tale. Endvidere er en visuel-VAD udviklet og fusioneret med den lydbaserede VAD. I den sidste del af afhandlingen er anvendelser af VAD diskuteret. Estimering af tidsforsinkelser undersøges i dybden i forbindelse med at indkapsle distributionsoplysningerne i talen. Støjreduktion i kombination med VAD er også drøftet. Endelig er en modificeret VAD anvendt inden for medicinsk signalbehandling for at give en foreløbig detektering af apnø.

To my family who has been always supporting my choices
and to Dora for her understanding

Acknowledgements

A great number of people without whom this dissertation would not have been successfully accomplished have to be acknowledged. First of all, I would like to thank Professor Ramjee Prasad for giving me the opportunity to pursue my Ph.D. at Aalborg University under his supervision. Professor R. Prasad has always been supportive and encouraging throughout my studies. Although, I'm mostly grateful to him for his intolerance to modesty that made me work even harder.

I would also like to express my deepest gratitude to my two co-supervisors, Professors Fotios Talantzis and Christos Boukis for their constant support, encouragement and for believing in me even in times I was about to give up. In particular, I need to thank Fotios for inspiring me through his unlimited ideas and our endless discussions on the research field, but also for showing me, in every step I made, the big picture of my achievements. Especially, I have to thank him for occasionally breaking his mobile phone just to make sure that I always evolve my hardware skills in parallel to my studies.

I'm grateful to Christos for introducing me to the field of audio processing and research in general, and for providing me with the foundations upon which my work has been developed. I thank him because he never ceased to care about my progress despite the difficulties he faced and his career change. He made sure I started this work the best way I could, as Fotis made sure I continued and finished it the same way. They were both heavily involved in what I consider the most significant outcome of my work, our friendship.

Additionally, I wish to thank Professor Zheng-Hua Tan for his assistance, understanding and advising towards the successful completion of this work. I consider myself privileged to have him as my advisor especially for making sure I always stick to the study plan.

I wish to thank Prof. Aristodemos Pnevmatikakis, the most dedicated scientist I've ever met and the highly inspiring Prof. Lazaros Polymenakos, for their support and time invested to provide suggestions, comments and guidance throughout our collaboration.

I would be ungrateful if I haven't mention my colleagues Andreas, Elpiniki, Kostas, Menios, Nikos, Nikos, Osama, Sakis, Stavros and Vasilis for their assistance and support all this time. Moreover, I need to thank Dr. Charalampos Doukas and Prof. Ilias Maglogiannis for our collaboration in the field of apnea detection.

Furthermore, many thanks are due to my friends Giorgos, Charalampos and Lefteris for their understanding and for making sure, that every time we went cycling I suffered such injuries, the recovery periods of which were long enough to help me concentrate in my studies.

Finally, I would like to thank Athens Information Technology (AIT) and in particular Didoe Prevedourou for providing me the facilities and such a supportive working environment.

Contents

List of Figures

LIST OF FIGURES

List of Tables

Glossary

AWGN: Additive White Gaussian Noise

AED: Acoustic Event Detection

CDF: Cumulative Distribution Function

CLT: Central limit theorem

CNG: Comfort Noise Generator

CCMSM: Convex combination of multiple statistical models

CT: Close-talking

DCT: Discrete Cosine Transform

DD: Decision Directed

DFT: Discrete Fourier Transform

DOA: Direction-of-Arrival

DTX: Discontinuous Transmission

EEG: Electroencephalography

EM: Expectation Maximization

EMD: Empirical Mode Decomposition

FF: Far-Field

FFT: Fast Fourier Transform

FN: False Negatives

FP: False Positives

ΓD: Gamma Distribution

GCC: Generalized cross-correlation

GGD: Generalized Gaussian Distribution

GMM: Gaussian Mixture Models

HOS: Higher-order Statistics

HMM: Hidden Markov Models

HHT: Huang Hilbert Transform

IMF: Intrinsic Mode Function

KLT: Karhunen Loeve Transform

KS: Kolmogorov-Smirnov

LD: Laplacian Distribution

LPC: Linear Prediction Coding

LRT: Likelihood Ratio Test

MCCC: Multichannel cross-correlation coefficient

MFCC: Mel-Frequency Cepstral Coefficient

MGGD: Multivariate Generalized Gaussian distribution

MI: Mutual Information

MM-LRT: Multiple Microphone likelihood ratio test

MMMO-LRT: Multiple Microphone Multiple Observation likelihood ratio test

MO-LRT: Multiple Observation likelihood ratio test

MSM: Multiple Statistical Models

NN: Neural Networks

OSA: Obstructive Sleep Apnea

PD: Predicted Estimation

pdf: probability density function

PHAT: Phase Transform

PSG: Polysomnography

P_c**:** Speech Detection Error Rate

P_e**:** Average Detection Error Rate

P_f**:** Nonspeech Detection Error Rate

P_{fe}**:** Front End Clipping Error Rate

P_m**:** Mid Clipping Error Rate

P_{ov}**:** Late Nonspeech Detection Error Rate

ROI: Region Of Interest

SENS: Sensitivity

SL: Single Linkage

SM-LRT: Single Microphone likelihood ratio test

SNR: Signal-to-Noise Ratio

SpDEP: Speaker Dependent

SpIND: Speaker Independent

SPCF: Specificity

STFT: Short-Time Fourier Transform

SVM: Suport Vector Machines

SWMSM: Switching of Multiple Statistical Models

TDE: Time Delay Estimation

TN: True Negatives

TP: True Positives

TΓD: Two-sided Gamma Distribution

VAD: Voice Activity Detection

VOIP: Voice over Internet Protocol

List of Publications

This Thesis is a monograph, which contains some unpublished material, but is mainly based on the following publications.

Journals

(1) Petsatodis T., Boukis C., Talantzis F., Z.-H. Tan, R. Prasad, **"Convex Combination of Multiple Statistical Models with application to VAD"** , *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, Volume: PP , Issue: 99, 10.1109/TASL.2011.2131131.

(2) Doukas, C. Petsatodis, T. Boukis, C. Maglogiannis, I. **"Automated sleep breath disorders detection utilizing patient sound analysis"**, *Biomedical Signal Processing and Control, Elsevier*, Volume 7, Issue 3, May 2012, Pages 256-264.

(3) Petsatodis T., Talantzis F., Boukis C., Z.-H. Tan, R. Prasad, **"Exploring Super-Gaussianity towards robust information-theoretical time delay estimation"** *Journal of the Acoustical Society of America (JASA)*, 2012, (to appear).

(4) Petsatodis T., Boukis C., Talantzis F., Z.-H. Tan, R. Prasad, **"EMD-based Multiple Input Likelihood Ratio VAD"**, *Journal on Audio, Speech, and Music Processing, EURASIP*, 2012, (submitted).

Conferences

(5) Petsatodis T., Talantzis F., Boukis C., Z.-H. Tan, R. Prasad, **"Multi-Sensor Voice Activity Detection based on Multiple Observation Hypothesis Testing"**, *INTERSPEECH*, 2011, Florence, Italy.

(6) Petsatodis T., Boukis C., **Efficient Voice Activity Detection in Enclosures using Far-Field Microphones**. *IEEE Digital Signal Processing*, 2009, Santorini, Greece.

(7) Petsatodis T., Boukis C., Pnevmatikakis A. , **Robust Multimodal Voice Activity Detection**. *IEEE Digital Signal Processing*, 2009, Santorini, Greece.

Chapter 1

Introduction

1.1 Voice Activity Detection

Environmental characteristics, such as noise and reverberation, inhibit technical barriers to most speech processing systems towards achieving ideal performance rates. In order to overcome such adversities, numerous noise reduction and speech enhancement techniques have been developed, operating on the basis of accurately estimating noise and speech statistics, obtained usually by means of a Voice Activity Detector (VAD).

The VAD problem considers detecting the presence of speech in a noisy signal. VAD decision, Fig.1.1, is normally based on a feature vector extracted on a short-time frame-by-frame basis to allow operation in real-time. Those must be discriminative speech features suitable for detection. The VAD module has to decide up two hypotheses; speech presence and speech absence, relying on some rule or method for assigning current frame to a class. A decision smoothing algorithm is usually employed, as a post decision filtering stage, in order to improve the robustness against the noise by reducing the risk of the low-energy portion of speech at the end of an utterance being falsely rejected or by rejecting false alarms.

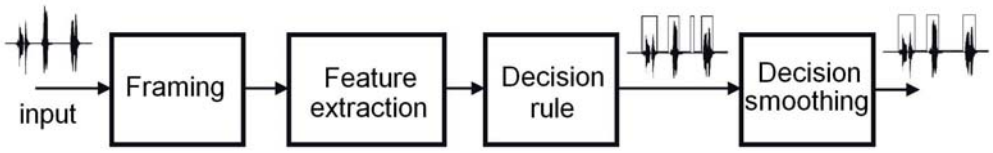

Figure 1.1: Typical block diagram of VAD processes.

VAD accuracy is of paramount importance in modern telecommunication systems, where in conjunction with comfort noise generator (CNG) and discontinuous transmission (DTX) modules, play a critical role in enhancing the system performance. Actual speech activities normally occupy 40-60% of the time of a regular conversation in a telecommunication system and it has been estimated that DTX with VAD decision could approximately double the transmission channels capacity

1. INTRODUCTION

(1, 2, 3). Voice activity detection (VAD) enables reallocating resources during the periods of speech absence. Embedded VAD components in telecommunications systems (like in cellular telephony) attempt to reduce power consumption of transmitters and bandwidth utilization (4). VAD technology is also integrated in other speech-processing systems, such as Speaker Identification, Automatic Event Detection, and Automatic Speech Recognition, to prevent their operation in the absence of speech, thus reducing the error rate (5). In the field of Wireless Sensor Networks, that has been traditionally focused on low duty-cycle applications, there is a growing need to support real-time streaming of audio (using FF microphones) and/or low-rate video for use in emergency situations and short-term intruder detection (surveillance) (6, 7). In these systems VAD is a fundamental component which is used, to reduce transmission rate during the silent periods of the input signal thus increasing the system capacity, reducing the co-channel interference and the transmitter power consumption.

During the last decade, numerous researchers have developed different strategies for detecting speech on a noisy signal. The different VAD systems developed can be classified in to the two major categories of unsupervised and supervised methods, depending whether *a priori* information is used in order to train system parameters. VAD approaches of the former class typically rely on the continuous observation of a specific metric to decide on the content of an audio signal. Such metrics are energy levels, zero-crossing rate, periodicity, linear prediction coding (LPC) parameters, Higher-order Statistics (HOS) of the LPC residuals of the signal, power envelope dynamics, fractals and mutual information (4, 8, 9, 10, 11, 12, 13). Recently introduced statistical VAD algorithms attempt to mathematically formulate the problem, by employing the Likelihood Ratio Test (LRT) as a decision criterion (14, 15, 16, 17, 18). The drawback of most of these solutions is that the statistical characteristics of the employed models of speech and noise are assumed to be fixed and *a priori* defined.

Supervised VAD systems like Hidden Markov Models (HMM), Neural Networks (NN) and Suport Vector Machines (SVM) have been also proposed, that require a vast amount of training data in order to optimise their parameters (1, 18, 19, 20).

The speech/non-speech classification task is not as trivial as it appears, and most of the VAD algorithms fail when far field (FF), (instead of close talking) microphones are employed and/or the level of background noise or reverberation increases. Furthermore, speech characteristics vary with time and can be affected by several unpredictable factors including speaker's temper, mood, gender, age, and more and environmental characteristics. Thus, despite the fact that VAD has been in the focus of speech processing research for many years, speech/non-speech detection is still an open problem affecting numerous applications in several sectors. This broad application spectrum of VAD and its critical role in system robustness makes the need for development of

accurate algorithms that can perform adequately under highly varying conditions and reverberant environments imperative.

1.2 Performance Metrics

A VAD distinguishes speech from non-speech frames in the presence of background noise. In general, VAD errors can be categorized into two main types of errors, notably clipping errors and false detection errors. Clipping errors occur when speech frames are misclassified as noise frames, which is intolerable in speech encoders due to its effect on speech intelligibility, while false detection errors are due to misclassifying noise frames as speech frames.

The performance evaluation metrics are summarized in the following list (14, 21):

- *Speech Detection Error Rate (P_c)* : the ratio of the incorrect decisions at speech segments over the total time of speech segments (voice clipping).
- *Non-speech Detection Error Rate (P_f)* : the ratio of the incorrect decisions at non-speech segments over the total time of non-speech segments (false alarm).
- *Average Detection Error Rate (P_e)* : the average error rate estimated as the mean of P_c and P_f.

The above metrics can be decomposed in a set of more specialized ones (22), although, they are not very often used in bibliography:

- *Front End Clipping Error Rate (P_{fe})* : the ratio of the incorrect decisions at speech segments over the total time of speech segments introduced in passing from noise to speech activity.
- *Mid Clipping Error Rate (P_m)* : the ratio of the incorrect decisions at speech segments over the total time of speech segments $P_m = P_c - P_{fe}$.
- *Late Non-speech Detection Error Rate (P_{ov})* : the ratio of the incorrect decisions at non-speech segments over the total time of non-speech segments due to the VAD flag remaining active in passing from speech activity to noise.

1.3 Motivation and Research Objectives

The performance of a VAD, like for most of speech processing systems, is significantly downgraded when far-field (FF) microphones are used, instead of the conventional close-talking (CT), due to reverberation effects, competitive sound sources, and speaker movement that can significantly alter

the statistical characteristics of captured speech. Thus, the primary objective of this work focuses on the development of robust VAD systems and algorithms, able to operate with one or more FF microphones within adverse environments. These algorithms should be able to cater for the varying energy of speech signals captured with FF sensors, and ought to tackle the obstacles introduced by the inherent reflections and the highly-intensive interfering noises.

The development will be based on evolving currently employed frameworks and newer techniques. Time-frequency analysis will be considered, ranging from Fourier analysis to more recent techniques like the Empirical Mode Decomposition. Statistical metrics like the Bayesian likelihood ratio test, the zero-cross rate and the periodicity will be re-evaluated in this context and novel criteria will be constructed.

Towards achieving VAD robustness, the effects on the characteristics of captured speech, imposed by phenomena such as noise, reverberation and speaker movement, have to be carefully studied. Outcomes of this evaluation will serve as the basis for the development of VAD systems able of adaptively adjusting internal modelling parameters to match the characteristics of the instantaneous input.

Moreover, part of the research will focus on microphone arrays and especially on the utilization of spatial information from multiple and independent microphone in contrary to single microphone based approach, which can only utilize time and/or frequency information. Direction-of-Arrival estimation and its performance dependence on reverberation will be also reconsidered in the context of VAD.

Also supervised classification methods like hidden Markov models (HMM), Gaussian Mixture Models (GMMs) and other will be considered and assessed in this framework. In parallel, the bi-modality of speech generation, in terms of audio and video information, extracted from a talking person, will be investigated with the scope of merging the two fields towards achieving enhanced performance.

The effect of the developed VAD algorithms on other speech processing applications fields will be examined. Those will include speaker tracking, noise reduction and acoustic event detection. Furthermore, within the framework of of this work, an other objective will be also to investigate the possibility of encapsulating outcomes and techniques that demonstrated better performance for VAD in other speech processing systems in order to improve their performance.

The algorithms developed, as part of this the work, will be designed in a real-time frame-by-frame processing basis to allow for their integration in modern telecommunication systems, smart rooms and other technologies. Thus, their performance will have to be always compared with that of existing real-time technologies.

1.4 Contributions

The efforts towards meeting the goals and research objectives, that were set in the beginning of this work, resulted in a series of outcomes in the area of VAD and its applications. The basis of the work was an extensive study on the effects of noise and reverberation on speech distribution at various intensity levels and conditions. We demonstrate with simulations and experiments that when far-field microphones are used instead of the conventional close-talking, reverberation effects, competitive sound sources, and speaker movement can significantly alter the distribution of captured speech. In contrary to previous studies, we depicted that captured speech with FF microphones is not solely Gaussian, Laplacian, or Gamma distributed, given its non-stationarity in time and its dependence on external interferences. This way we justified why speech processing systems relying solely on the Gaussian or other fixed assumptions are expected not to perform adequately under varying conditions. Fixed distribution assumptions can be accurate only under specific conditions of reverberation and noise. Those outcomes, actually directed the whole research effort into the development dynamically adaptive systems able to overcome such environmental adversities and speech dynamics. Hopefully, this can inspire similar work from other researchers on related fields.

The first offspring of this study is a statistical voice activity detector, which relies on the modelling of the distribution of speech as a convex combination of a Gaussian, a Laplacian, and a two-sided Gamma distribution. The decision criterion of the proposed algorithm is the weighted sum of three likelihood ratio tests, each one corresponding to one of the fundamental core distributions. The computation of the corresponding weights has been based on the statistical distances of the instantaneous input samples from the Gaussian, the Laplacian, and the two-sided Gamma distribution, estimated using the Kolmogorov-Smirnov test. Experiments performed using artificially reverberated and contaminated with additive noise anechoic audio data revealed that the specific voice activity detector outperforms the existing systems in terms of error rate and that it produces reliable results even under adverse noise conditions and reverberation effects. The result justified our initial hypothesis, that speech distribution can be better modelled as linear combination of a set of primal distributions rather than any other single distribution approach.

In the next step, we considered the encapsulation of spatial information, embedded in signals captured by far-field microphone arrays, in a VAD scheme. The developed scheme is taking advantage of the spatial information provided by multiple sensors without the need of knowledge of direction-of-arrival estimates like previous approaches. Simulations performed demonstrated that the proposed system remains more robust than a set of related counterparts without imposing additional delay in the system or being subject to reverberation.

The multi-microphone VAD served as the platform to merge our results with a very powerful analysis framework namely Empirical Mode Decomposition (EMD). This highly efficient signal decomposition method significantly enhanced the performance of VAD acting as a speech enhancement technique prior to voice detection. The outcome of this procedure demonstrates significantly enhanced performance compared to single microphone approaches.

In the area of supervised voice activity detection, a system based on the modelling capabilities of hidden Markov models has been developed. Gaussian Mixtures modelling has been employed per model state to cater for the variable distribution of speech in respect to the outcomes of the research on speech distribution. Given the bi-modality of speech generation process, conveying both audio and visual information, an Audio-Visual VAD that combines the advantages of both modalities has also been considered. Although the developed system wasn't based on two optimal modalities of video and audio, fusing of different VAD schemes showed that there is a noticeable increment in performance even under extremely adverse conditions.

Following the study plan we then concentrated on exploring applications of VAD. The performance benefits of combining the developed multi-microphone VAD with a direction-of-arrival (DOA) estimation scheme were demonstrated on the basis that speech emission is a discontinuous sound source. This has been done in order to overcome instabilities of audio tracking based on DOA when the speaker of interest is not emitting speech. The employed DOA system was based on information theoretical TDE system. The optimization of this TDE scheme was also considered in the context of encapsulating speech shaped distributions in the underlying assumption of the speech model employed. Thus, we investigated how the performance of a robust information-theoretical TDE algorithm, changes as we switch between different underlying assumptions for the distribution of speech in respect to the instant input. The analysis performed, revealed a significant research outcome. The employed marginal MI criterion based TDE is not depended on the underlying assumption of the distribution of speech when that belongs to the family of Generalized Gaussian distribution, exploiting the invariance property of MI. To support the analysis, closed forms of the multivariate and univariate differential entropies for the Generalized Gaussian distribution were derived, that encapsulate the entropies of other well known distributions like Gaussian, Laplacian and Gamma.

Additionally, performance enhancement when using VAD in combination with noise reduction systems has been also documented in terms of residual suppression within silence intervals. For this scope an efficient noise reduction architecture has been developed based on cascading an one-pass scheme.

Finally, we steered our focus in acoustic event detection. More specifically an automated Sleep Apnea detection system utilizing snore sound analysis has been presented. The core of the pro-

cessing algorithm employed was based partially on the convex combination of multiple statistical models VAD aiming to a computationally lightweight system that could be potentially used to monitor patients at home. This way prognosis, treatment procedure and offering the maximum comfort to patients is improved. Conducted experiments using the system in various conditions have indicated increased accuracy in detecting snoring against background noise and indication of apneic events compared to other obtrusive methodologies.

1.5 Dissertation Outline

Following the research contributions agenda, in the first part of the thesis, Chapter 2 deals with the investigation of speech distribution variability under external interferences. The study forms the basis for the development of an unsupervised adaptive VAD based in the convex combination of a set of prime distributions detailed in Chapter 3. Following the schema of Fig.1.2, in Chapter 4 the work continues with the design of a multi-microphone VAD that incorporates spatial information. The multi-microphone VAD is later used in combination with Empirical Mode Decomposition in Chapter 5. Speech distribution information is also encapsulated in a supervised VAD scheme employing Hidden Markov Models and Gaussian Mixture Models to cater for the dynamics of captured speech distribution in Chapter 6. A visual-VAD is developed in the context of fusing the two modalities of speech generation. In the last part of the thesis, applications of VAD are discussed. In Chapter 7, Time Delay Estimation is investigated in depth, towards encapsulating speech distribution information. In Chapter 8, noise reduction in combination with VAD is also discussed. Finally, in Chapter 9 a modified version of the VAD presented in Chapter 3 is employed in the field of biomedical signal processing to provide preliminary detection of apnea.

More specifically, in Chapter 2, we review variability of speech distribution, when using far-field microphones, under the presence of noise and reverberation. Given that for most of speech processing systems it is of crucial importance to have a reasonable approximation for the probability density function (pdf) of speech, understanding the properties of speech distribution plays a very important role in the design of speech processing systems.

Chapter 3 the utilization of Likelihood Ratio Test (LRT) in Voice Activity Detection is explored. The basic LRT VAD algorithm under the assumption of Gaussian distributed noise and speech is presented. Problems emerging from the Gaussian assumption will be discussed in respect to the observations and conclusions inferred in Chapter 2. In addition, a robust VAD scheme based on these observations will be developed. The proposed VAD employs a convex combination scheme comprising three statistical distributions - a Gaussian, a Laplacian, and a two-sided Gamma - to effectively model captured speech. The mechanism according which the contribution of each

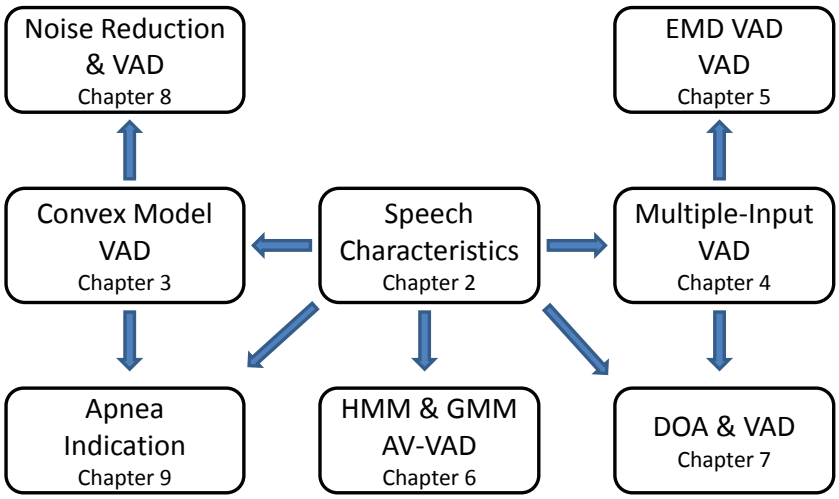

Figure 1.2: Dissertation technology connectivity.

distribution to this convex combination is automatically adjusted based on the statistical characteristics of the instantaneous audio input is also indicated. To further improve the performance of the system, an adaptive threshold is introduced, while a decision-smoothing scheme caters to the intra-frame correlation of speech signals. Extensive experiments under realistic scenarios are presented to support the proposed approach of combining several models for increased adaptation and performance.

Chapters 4 considers employing microphone arrays in VAD. The system developed processes signals captured by far-field sensors in order to integrate spatial information in addition to the frequency content available at a single channel sound capturing. The core of the system resides on the modification of a multiple observation hypothesis, testing at each sensor the probability of having an active speaker and then fusing the decisions. The VAD operates without the need of DOA estimation and eliminates additional delay imposed by previous multi-microphone VAD technologies.

In Chapter 5 the system developed as a multi-microphone VAD serves as the platform to merge VAD with a very powerful analysis framework the Empirical Mode Decomposition (EMD). This highly efficient method relies on local characteristics of time scale of the data to analyse and decompose non-stationary signals into a set of so called intrinsic mode functions (IMF). These functions are injected to the multiple microphone VAD scheme in order to decide upon speech presence or absence. The outcome of this procedure demonstrates significantly enhanced performance compared to single microphone approaches.

Chapter 6 presents a supervised VAD system that operates in the far-field. The core of this system consists of two left-right Hidden Markov Models that operate in the feature domain. Taking into consideration the observations emerged from Chapter 2 for the distribution of speech the observation probability distribution function of each state is modelled using Gaussian Mixture Models. An adaptive threshold is derived, that allows for optimum performance even in the case of varying noise statistics. Furthermore, to cater for the inter-frame correlation, especially in the case of speech presence, a hang-over scheme is employed. Furthermore, given speech generation is a bi-modal process conveying both, audio and visual information, an audiovisual VAD that combines the advantages of both modalities is considered.

In the final part of the work, applications of VAD are also considered. Chapter 7 illustrates performance benefits of combining the multi-microphone VAD with Direction-of-Arrival (DOA) estimation. Optimization through adaptation of speech shape characteristics in the embedded Time Delay Estimation (TDE) scheme is also considered, the same way that was beneficial for the convex combination based VAD. Towards this direction, the underlying assumption of Gaussian distributed source is replaced by that of Generalized Gaussian distribution that allows the evaluation of the problem under a larger set of speech-shaped distributions, ranging from Gaussian to Laplacian and Gamma.

In Chapter 8, performance enhancement when using VAD in combination with noise reduction systems is also discussed in terms of residual suppression within silence intervals. For this scope an efficient noise reduction architecture has been developed based on cascading an one-pass scheme.

The last application considered, is related to Acoustic Event Detection in the field biomedical signal processing. In Chapter 9 a modification of the VAD developed in Chapter 3 , is employed to provide preliminary detection of one of the major breathing-related sleep disorders, apnea.

Finally, Chapter 10 summarizes the thesis and concludes with some future research directions in the field of Voice Activity Detection.

Part I
Speech Characteristics

Chapter 2

Speech Characteristics

2.1 Introduction

In this Chapter we discuss the variability of speech distribution, under the presence of noise and reverberation. Understanding the properties of speech distribution plays a very important role in the design of speech processing systems. For most of such systems it is of crucial importance to have a reasonable approximation for the probability density function (pdf) of speech itself and the noise present.

The assumption of gaussianity of the speech source is a very common practice, despite the fact that has already been shown (23, 24) that in several feature domains including time, Fourier Transform (FT), Discrete Cosine Transform (DCT), and Karhunen Loeve Transform (KLT), distributions of clean speech, with SNR down to 20dB, are very well approximated by Laplacian (LD) and two-sided Gamma distributions (TTD). Furthermore, it has been also shown that the distribution of speech, captured with far-field microphones, is highly varying, depending on the noise and reverberation conditions.

Thus, the performance of systems relying on fixed distribution assumptions is expected to fluctuate depending on the specific underling assumption for the speech distribution. In addition, speech distribution varies with time and can be affected by several unpredictable factors including speaker's temper, mood, gender, age and environmental characteristics.

2.2 System Model

Speech signals captured within reverberant enclosures using FF microphones are subject to superposition of reflected versions of the source signal. The captured speech signal can also be affected by competitive sound sources and background noise. Source movement also affects the characteristics of the captured signal. Assuming a single speaker, the speech signal captured by a distant

microphone at time t is given by

$$x(t) = h(t) * s(t) + n(t) \tag{2.1}$$

where $s(t)$ denotes the source speech signal at time t, $h(t)$ the corresponding acoustic impulse response, $n(t)$ the additive noise, and $*$ denotes convolution (Fig. 2.1). The impulse response, $h(t)$, is the recording of the reverberation that is caused by an acoustic enclosure when an impulse is played, characteristic for every different receiver location (Fig. 2.2).

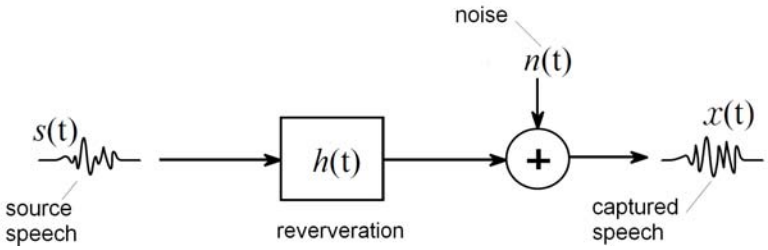

Figure 2.1: System model.

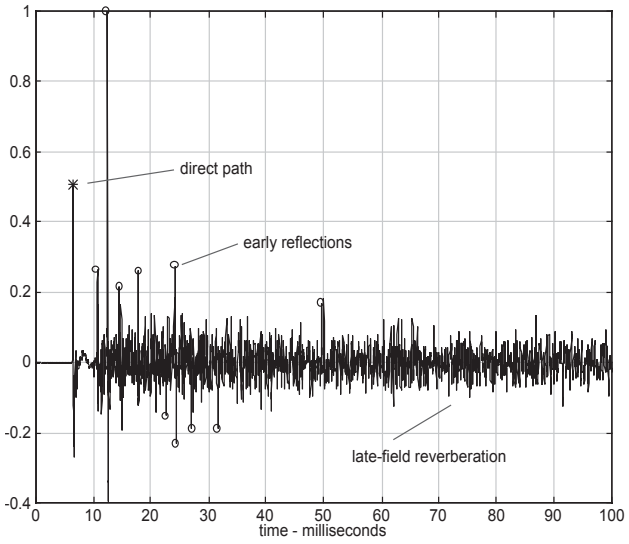

Figure 2.2: Example of an office room's impulse response.

In most speech-processing systems, it is important to have a reasonable approximation for the probability density function (pdf) of speech and noise. It has been shown that the pdf of source speech samples in the time domain and short-time frequency domain (when using frame lengths

close to 32 msec) is much better modelled by a Laplacian or a Gamma density function than a Gaussian one (23).

To evaluate the distribution of captured speech under diverse conditions when using distant microphones, a series of recordings were performed using a close-talking microphone in the small Anechoic Chamber of AAU (25) to collect clean speech samples. For the anechoic data collection, 13 participants (7 males and 6 females) were recorded at 16kHz, speaking at their mother-languages for approximately 15 min each, reading sentences and words presented to them with random pause intervals. Eight different languages appear in the data set (Arabic, Bulgarian, Chinese, Greek, Italian, Portuguese, Urdu, and Turkish). The participants were also recorded speaking English for 15 additional minutes under the same pattern. Speech intervals occupy half of the recording time. The recordings have been annotated manually.

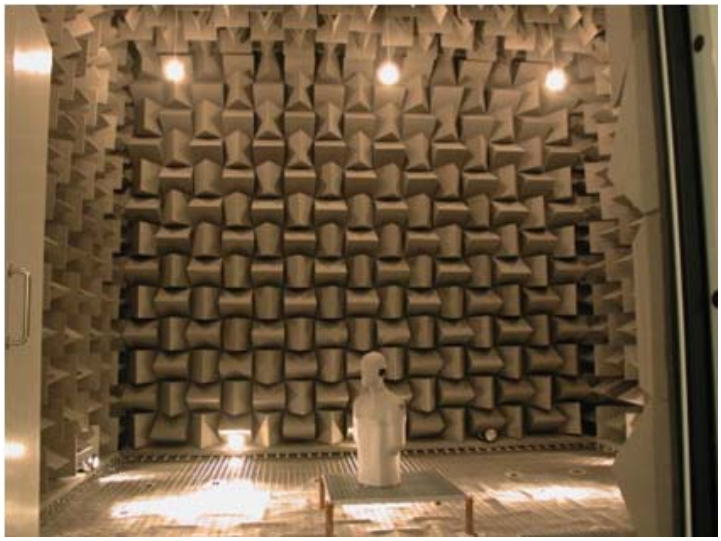

Figure 2.3: Inside the small anechoic chamber at Aalborg University

2.3 Source Speech Distribution

The distribution of the source speech (assumed to be anechoic and noiseless) varies depending on the duration of the time window used in its analysis. More specifically, when long-time window is used, the histogram of source speech amplitude in the time domain is approximated by a TΓD. As the duration of time window is further reduced, the distribution of source speech is better approximated by LD (23, 24). Furthermore, when segments of source speech shorter than 10 msec are analysed, the distribution that best fits the histogram of speech will be subject to the pronounced vowels, consonants and phonemes.

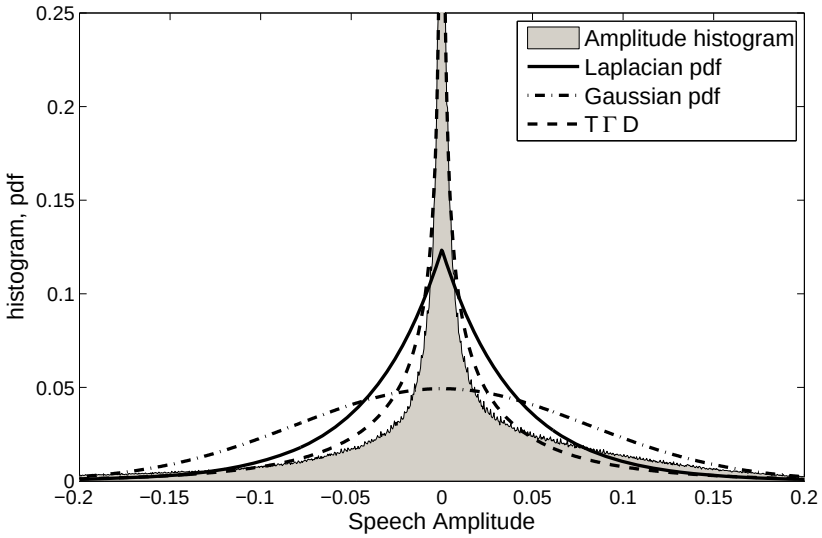

Figure 2.4: Source Speech Amplitude distribution and the three theoretical pdfs with the same mean and variance.

The histogram of the amplitude of source speech in the time and frequency domain is presented in Figs. 2.4 and 2.5, respectively. These have been derived from a source speech signal segment approximately 3 min long. For the derivation of the STFT, source speech has been segmented using a time window of 40 msec and 75% overlap.

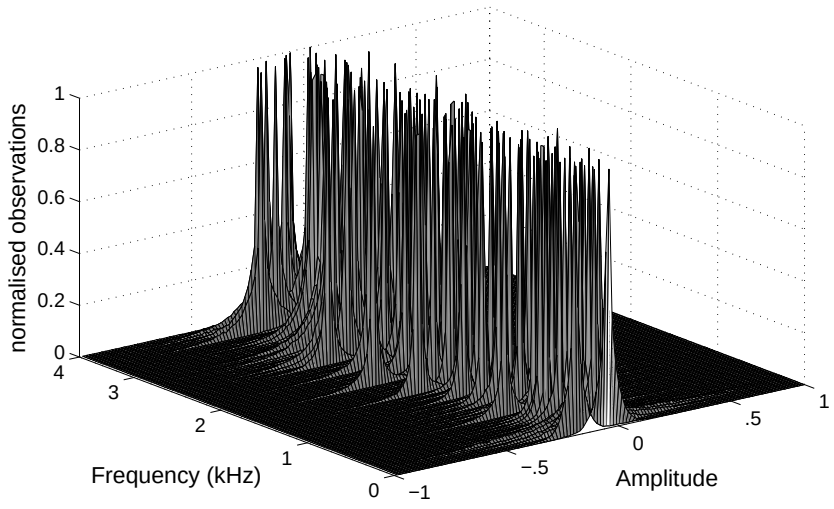

Figure 2.5: Source Speech Amplitude Distribution of frequencies. Histograms have been normalized to their maximum value per frequency bin.

The distribution of source speech in the frequency domain is approximated by a TΓD pdf for most of the frequency bins (Fig. 2.5). Apparently, there is a connection between the distribution of speech in time domain and in the frequency domain.

2.4 Effect of Noise

To assess the effect of additive noise on the statistics of captured speech, Additive White Gaussian Noise (AWGN) was artificially added to the source speech signals. The distribution of noisy speech was evaluated for several SNR values and the results are presented in Figs. 2.6 and 2.7 for the time and frequency domain, respectively.

From Fig. 2.6, it is observed that in the presence of 0 dB of AWGN, the distribution of captured speech becomes GD. The effect of AWGN on the statistics of the amplitude of captured speech in the frequency domain is similar: the more the SNR increases, the more their distribution becomes GD-shaped. Indeed, for 20 dB of AWGN, the distributions of higher-frequency amplitudes of captured speech are being shaped to LD (Fig. 2.7) whereas for 15 dB, this is also obvious for amplitude of mid-frequency components. Further reduction of SNR leads to transformation of low- and mid-frequency amplitudes. Transition from TΓD to LD and GD is finalized for SNR of 0 dB or less.

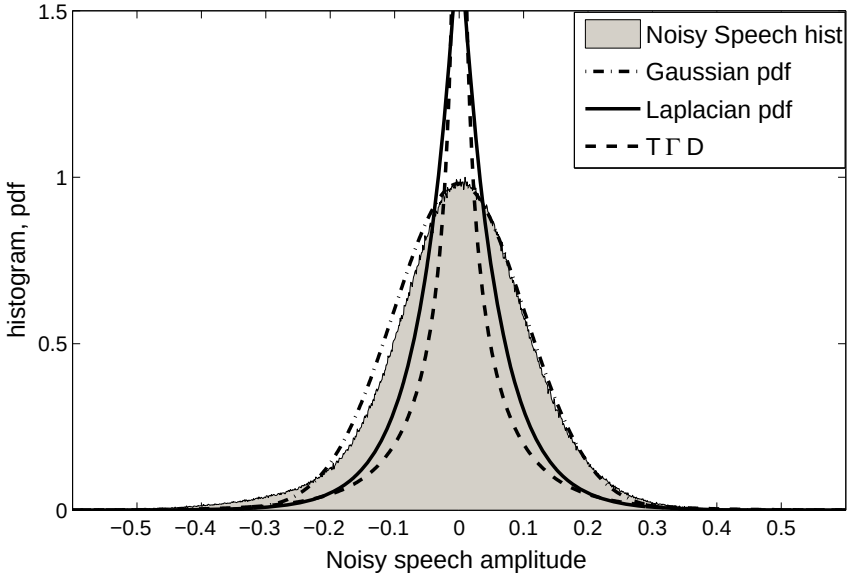

Figure 2.6: Noisy Speech Amplitude Distribution for SNR of 0 dB and three theoretical pdfs sharing the same mean and variance.

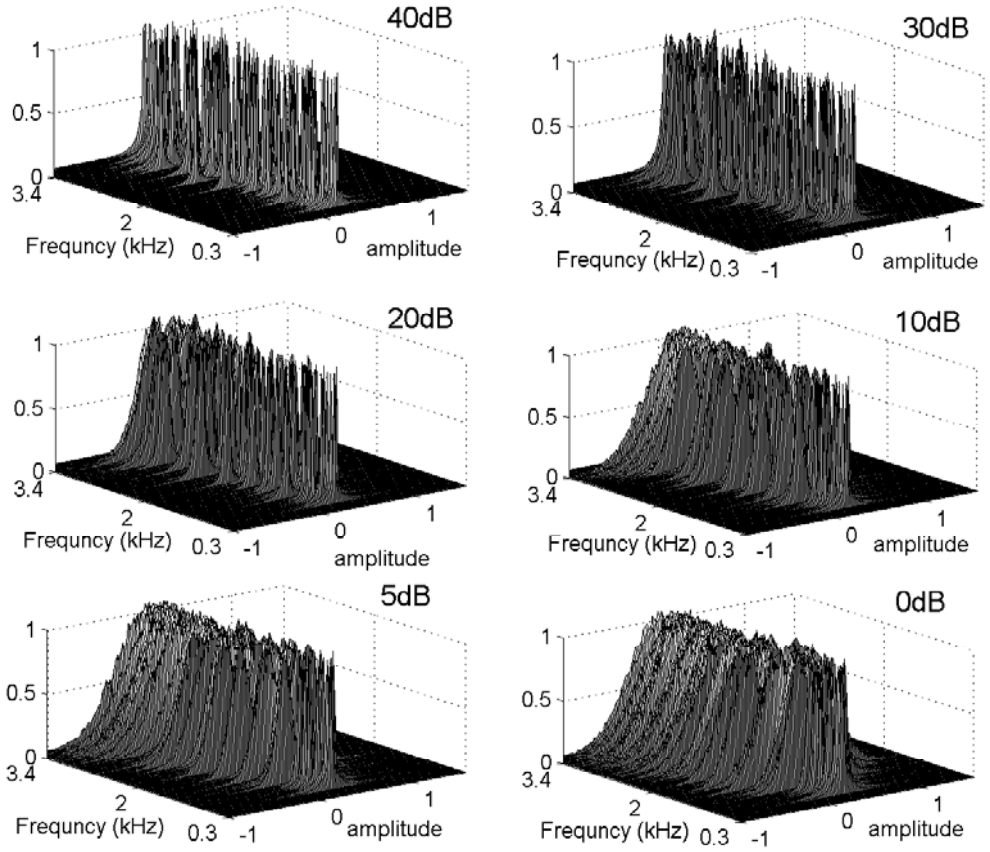

Figure 2.7: Amplitude distribution of captured noisy speech in frequency domain. Histograms have been normalized to their maximum value. The distributions of frequencies tend to be better approximated by GD as noise intensity increases. Higher frequencies are affected more than the lower ones due to less energy.

To quantify the effect of AWGN on the pdf of captured speech, the distance of its time amplitude histogram for the TTD, the LD, and the GD was estimated using the KS distance test. KS-test is based on estimating the maximum distance T between the empirical cumulative distribution function (cdf) F_X and the theoretical distribution F evaluated at the instant sample points X (26):

$$T(\mathbf{X}) = \max|F_X - F| \tag{2.2}$$

The most attractive feature of this test is that it does not depend on the underlying cdf being tested. In addition, it is an exact test, contrary to other (like chi-square) tests that depend on an adequate sample size for the approximations to be valid. Fig. 2.8 is derived by employing the

KS-test on the current captured speech input in time domain to estimate against the TΓD, LD, and GD for various SNR values.

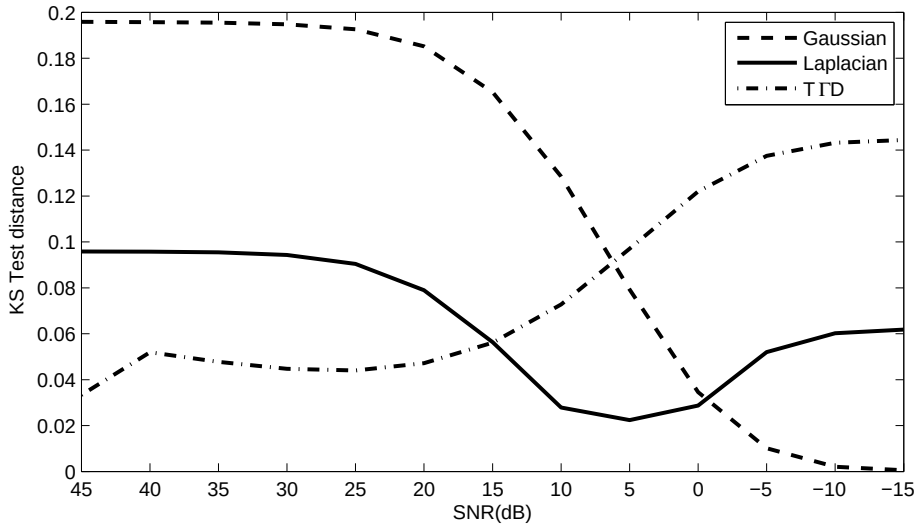

Figure 2.8: KS-test distance estimation of amplitude distribution from the three theoretical pdfs in time domain as SNR drops.

From Fig. 2.8 it is observed that the statistics of the captured speech tend to be better approximated by GD as SNR decreases. TΓD is closer to the distribution of captured speech only for high SNR values. As the SNR drops below 15dB the LD is closer to the statistics of the noisy signal. For SNR values below 0dB the distribution of the captured speech is again better approximated by GD.

The distribution of frequency components is also affected by additive noise in a similar manner, which is justified by the fact that time and spectrum are linearly dependent. Figs. 2.9, 2.10 and 2.11 present the effect of additive noise on the statistics of spectral amplitudes for noisy speech. These graphs show that the distribution of spectral amplitude of captured speech is closer to TΓD for high SNR; as the SNR decreases below a frequency-dependent threshold which is close to 20 dB, it is better approximated by LD. Further SNR reduction leads to Gaussian-shaped distribution of captured speech frequency amplitudes.

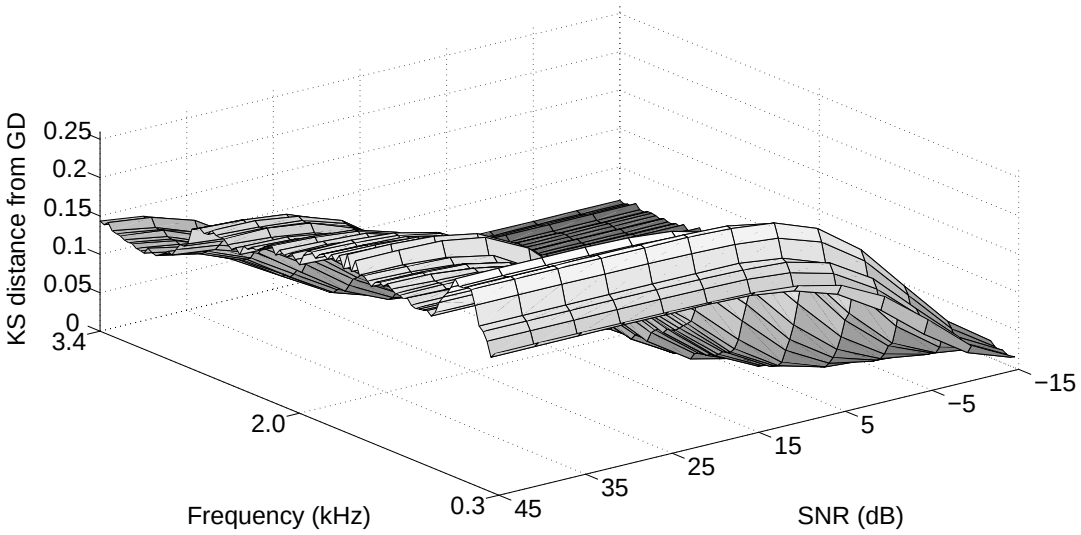

Figure 2.9: Gaussian KS-test distance estimation in frequency domain vs. SNR.

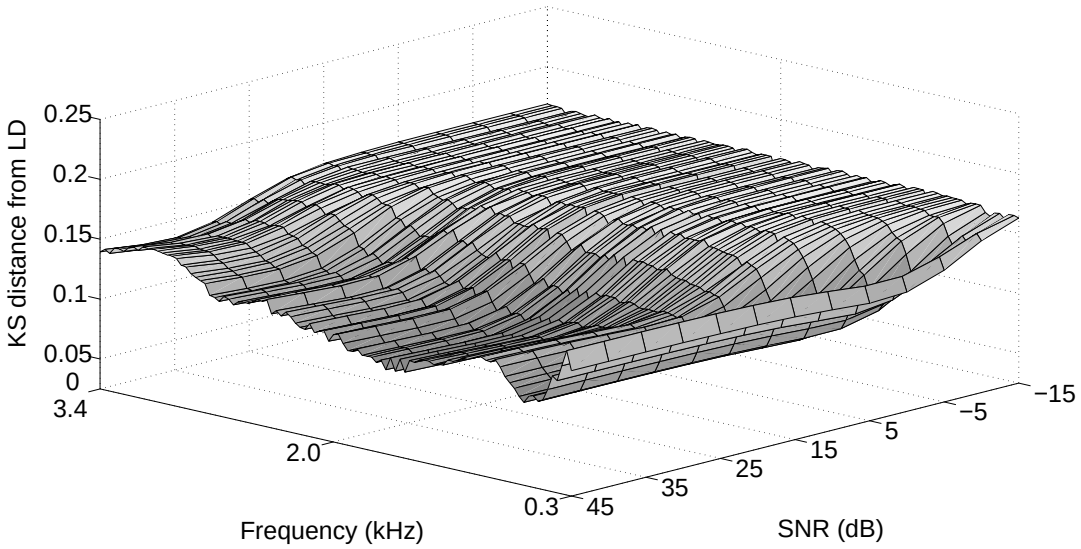

Figure 2.10: Laplacian KS-test distance estimation in frequency domain vs. SNR.

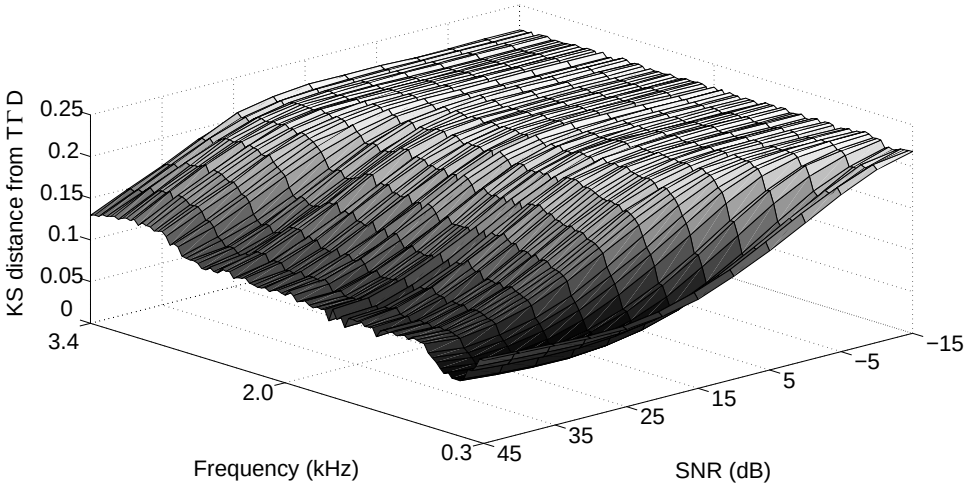

Figure 2.11: TΓD KS-test distance estimation in frequency domain vs. SNR.

2.5 Effect of Reverberation

To evaluate the effect of reverberation on the statistics of captured speech, the *Image Model for Small room Acoustics* (27) has been employed for increasing rate of reverberation time T_{60} on source speech signal (28) for [4.4, 5.8, 2.6]m room dimensions. The distribution of amplitude for reverberant speech in time and frequency domain is depicted in Figs. 2.12 and 2.13, respectively.

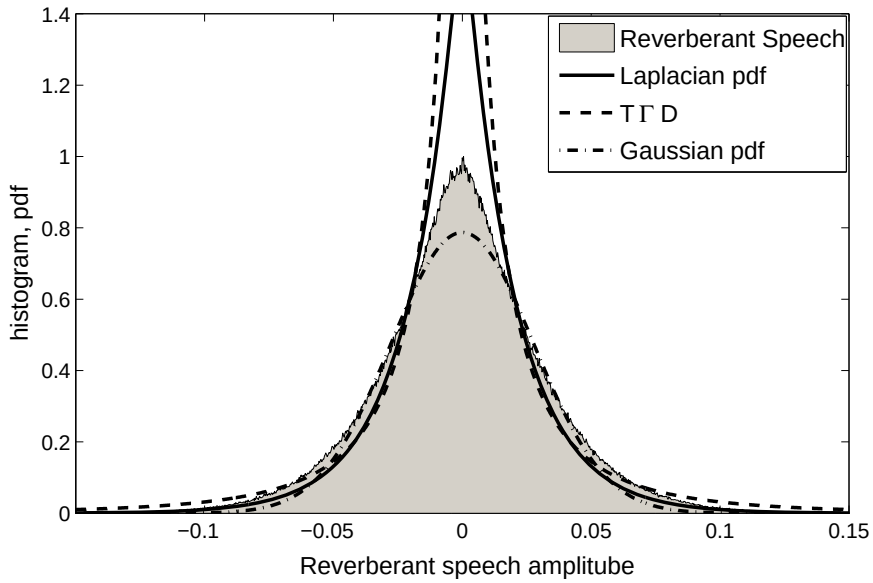

Figure 2.12: Reverberant Speech Amplitude Distribution $T_{60} = 2.2$ sec and the three theoretical distributions with same mean and variance.

As the reverberation time increases over $T_{60} > 0.3sec$, the distribution of time amplitude of captured speech becomes LD. Similarly, the distribution of frequency amplitudes is also affected by reverberation. Contrary to the case of AWGN though, the effect of reverberation on captured speech is evident only for high values of T_{60}. Only for reverberation times greater than 0.7 sec, the distribution of most of the captured reverberant speech frequency amplitudes is better modelled by LD.

Only for values of T_{60} greater than 2.2 sec, the distribution of spectral amplitude becomes GD, which is a scenario not frequently met in real life. Figure 2.14 shows that for low reverberation times, the distribution of captured speech is better approximated by TTD. As the T_{60} increases above 0.1 sec, the LD is closer to the distribution of captured speech.

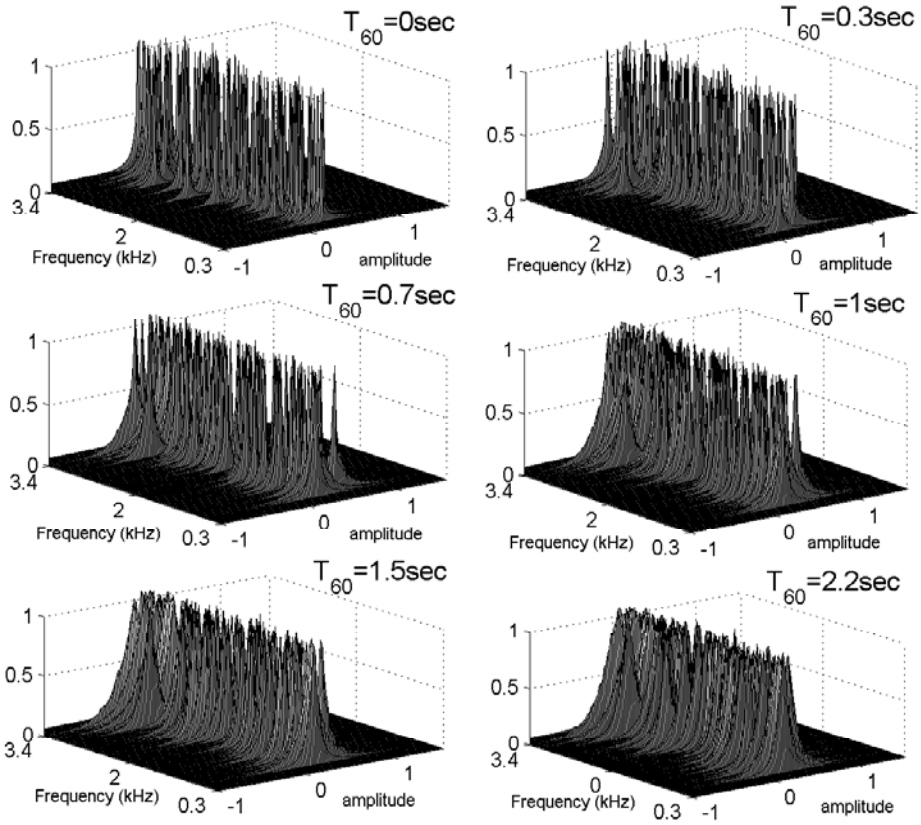

Figure 2.13: Effect of reverberation on the spectral amplitude of captured speech for various values of T_{60}. Histograms have been normalized to their maximum value.

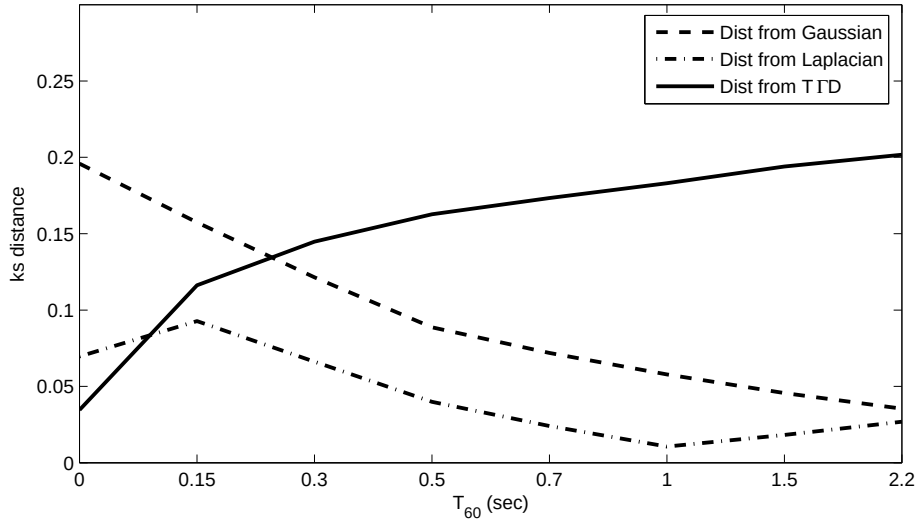

Figure 2.14: KS-test distance estimation vs. T_{60} in time domain

The effect of reverberation on the distribution of the spectral amplitude of captured speech is similar, as can be observed from Figs. 2.15-2.17.

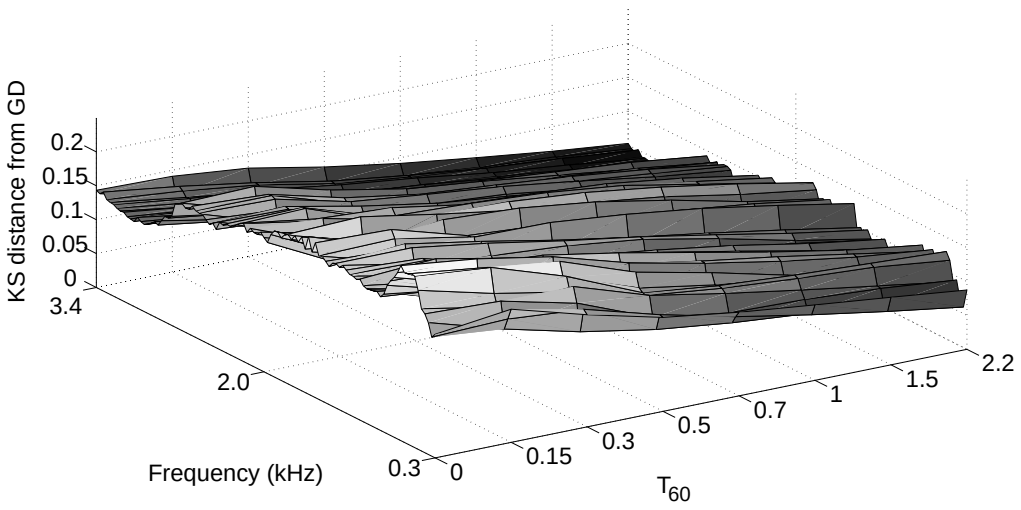

Figure 2.15: Gaussian KS-test distance of frequency distribution vs. T_{60}

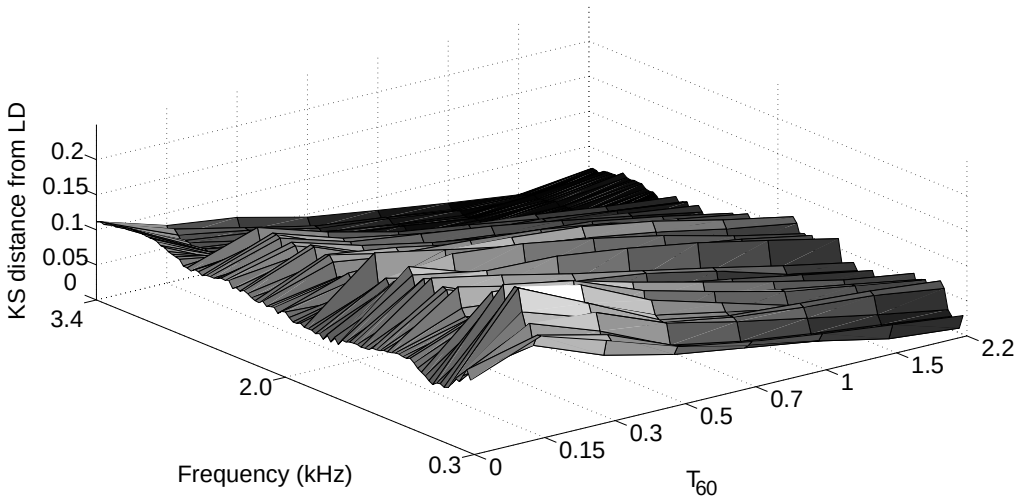

Figure 2.16: Laplacian KS-test distance of frequency distribution vs. T_{60}

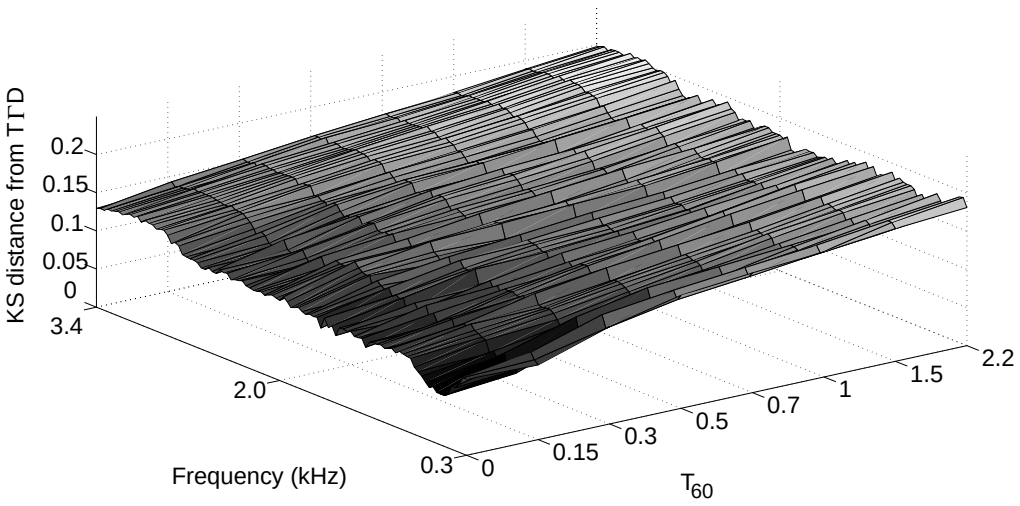

Figure 2.17: TΓD KS-test distance of frequency distribution vs. T_{60}

Notice that extremely long reverberation times are required for the distribution of captured speech to become GD. This can be justified by central limit theorem (CLT); a large number of echoes with varying delays and attenuation should be present for the source signal to become GD at the receiver side as the continuous addition of data of different distributions will eventually lead into a GD shape distribution. In theory, this can be observed for unrealistic large values of T_{60}.

2.6 Effect of Simultaneous Noise and Reverberation

Under a realistic situation, reverberation coexists with additive noise sources in the same enclosure making the detection of speech signals using far-field microphones even more difficult. Even low reverberation times (T_{60}) when combined with AWGN can cause severe effects on the performance of VAD systems. This is illustrated in Fig. 2.18 where the distance of the captured speech distribution from the GD was estimated for different values of SNR and T_{60}. The experiment has been performed by initially convolving source speech with the room impulse response and then adding noise as dictated by eqn. (1).

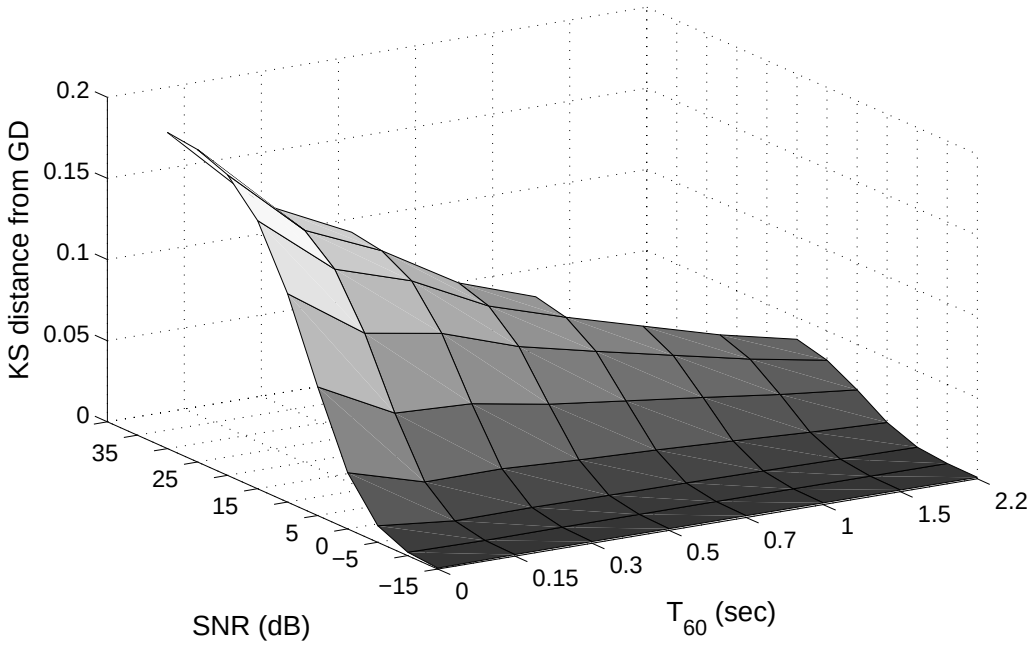

Figure 2.18: Gaussian KS-test distance vs. T_{60} vs. SNR

The simulation performed, the results of which are depicted in Fig. 2.18, reveals that when combined phenomena occur the distribution of speech approaches GD even faster. The height of ks-distance curves for different values of SNR reduces with the increasing reverberation time. The same attribute holds for reverberation curves as the noise increases. As expected, the effect of additive noise on the distribution of captured speech is more profound than the effect of reverberation, as the slope of the plot is steeper towards T_{60} axis than towards SNR.

2.7 Distance Effect

When working with unobtrusive far-field microphones, mounted on fixed positions on the wall, even the slightest movement of the speaker can affect the shape of the distribution of captured speech. Fig. 2.19 shows how this distribution approaches Gaussian as the speaker moves away from the microphones. Such a movement causes the SNR of the direct speech signal to drop (due to air attenuation) and the distribution of the reflected speech replicas to be altered. The situation is even more adverse when the fixed microphones are placed near the modes of the enclosure.

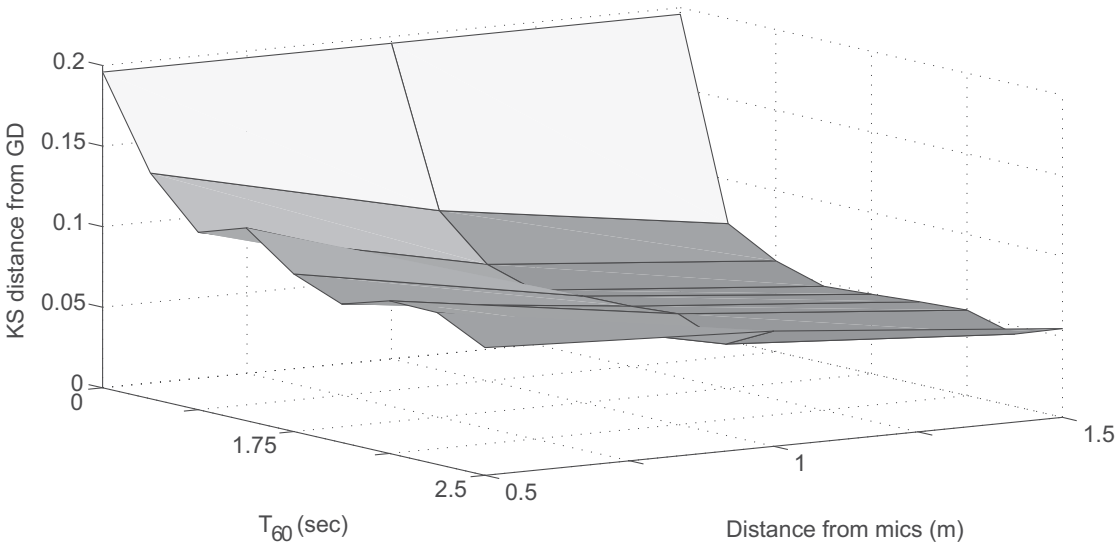

Figure 2.19: Gaussian KS-test distance vs. T_{60} vs. distance from mics

2.8 Conclusions

In this Chapter the effects of noise and reverberation on speech distribution have been explored for various intensity levels. We demonstrated that when far-field microphones are used instead of the conventional close-talking, reverberation effects, competitive sound sources, and speaker movement can significantly alter the distribution of captured speech. It becomes apparent that captured speech is not solely GD, LD, or ΓD distributed, given its non-stationarity in time and its dependence on external interferences. Speech processing systems relying solely on the Gaussian or other assumptions are expected not to perform adequately under varying conditions. Fixed distribution assumptions can be accurate only under specific conditions of reverberation and noise. Thus, in order for a system to be able to operate under variable conditions, encapsulating speech

distribution dynamics is of critical importance. Outcomes of studying how the distribution of speech is shaped by factors such as those examined here, are used later on as the design basis of systems described in detail in the following Chapters.

Part II

Unsupervised Voice Activity Detection

Chapter 3

Employing Likelihood Ratio Test for Voice Activity Detection

3.1 Introduction

In this chapter the utilization of Likelihood Ratio Test (LRT) in Voice Activity Detection will be explored. The basic LRT VAD algorithm under the assumption of Gaussian distributed noise and speech is presented. Problems emerging from the Gaussian assumption will be discussed in respect to the observations and conclusions inferred in Chapter 2. In addition, a robust VAD scheme based on these observations will be developed. The proposed VAD employs a convex combination scheme comprising three statistical distributions - a Gaussian, a Laplacian, and a two-sided Gamma - to effectively model captured speech. The mechanism according to which the contribution of each distribution to this convex combination is automatically adjusted based on the statistical characteristics of the instantaneous audio input is also presented. To further improve the performance of the system, an adaptive threshold is introduced, while a decision-smoothing scheme caters to the intra-frame correlation of speech signals. Extensive experiments under realistic scenarios are presented to support the proposed approach of combining several models for increased adaptation and performance. Finally, an important observation regarding the behaviour of LRT-based VAD algorithms in the presence of reverberation is also presented.

3.2 Binary Hypothesis VAD

Assuming a single speaker, the speech signal captured by a distant microphone at time t is given by

$$x(\tau) = h(\tau) * s(\tau) + n(\tau) \tag{3.1}$$

where $s(\tau)$ denotes the source speech signal at time τ, $h(\tau)$ the corresponding acoustic impulse response, $n(\tau)$ the additive noise, and $*$ denotes convolution.

3. EMPLOYING LIKELIHOOD RATIO TEST FOR VOICE ACTIVITY DETECTION

Voice activity detection can be expressed as the likelihood ratio of two hypotheses stating speech presence and absence (17). Assuming additive noise the two hypotheses H_1 and H_0 that indicate speech presence and speech absence are accordingly:

$$H_0 : \text{speech absence} : \mathbf{X}(t) = \mathbf{N}(t) \tag{3.2}$$

$$H_1 : \text{speech presence} : \mathbf{X}(t) = \mathbf{S}(t) + \mathbf{N}(t) \tag{3.3}$$

where $\mathbf{X}(t) = [X_0(t), X_1(t), ..., X_{K-1}(t)]^T$, $\mathbf{S}(t) = \left[S_0(t), S_1(t), ..., S_{K-1}(t)\right]^T$, $\mathbf{N}(t) = [N_0(t), N_1(t),$ $..., N_{K-1}(t)]^T$ are the noisy captured speech, reverberated speech, and noise frequency components. Here $\mathbf{N}(t)$ is assumed to encapsulate the reverberation effects.

Real and imaginary parts of noise and speech frequency spectrum are in general assumed to be zero mean Gaussian distributed. The probability densities for the noise and speech components with k denoting the frequency bin are given by

$$f_n^G(\mathbf{N}_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{\mathbf{N}_k(t)^2}{2\sigma_{n,k}^2}} \tag{3.4}$$

$$f_s^G(\mathbf{S}_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{s,k}^2}} e^{-\frac{\mathbf{S}_k(t)^2}{2\sigma_{s,k}^2}} \tag{3.5}$$

where $\sigma_{n,k}^2$, $\sigma_{s,k}^2$ the slowly varying variances of the Gaussian distributed noise and speech respectively. The probability density functions conditioned on H_0 and H_1 are given by

$$p(X(t)|H_0) = \prod_{k=0}^{K-1} \frac{1}{\pi\lambda_{n,k}} \exp\left\{-\frac{|X_k(t)|^2}{\lambda_{n,k}}\right\} \tag{3.6}$$

$$p(X(t)|H_1) = \prod_{k=0}^{K-1} \frac{1}{\pi\left[\lambda_{n,k} + \lambda_{s,k}\right]} \exp\left\{-\frac{|X_k(t)|^2}{\lambda_{n,k} + \lambda_{s,k}}\right\} \tag{3.7}$$

where $\lambda_{n,k}$ and $\lambda_{s,k}$ denote the variances of N_k, S_k respectively.

3.2.1 Employing Likelihood Ratio Test (LRT) in VAD

In the case of single microphone VAD scheme the likelihood ratio for the kth frequency bin is defined as

$$\Lambda_k \equiv \frac{p(X_k(t)|H_1)}{p(X_k(t)|H_0)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\} \tag{3.8}$$

where $\xi_k \equiv \lambda_{s,k}/\lambda_{n,k}$ and $\gamma_k \equiv |X_k(t)|^2/\lambda_{n,k}$ the *a priori* and *a posteriori* signal to noise ratios (29).

The decision criteria is based on evaluating the geometric mean of the likelihood ratios for the individual frequencies and is given by

$$\log \Lambda = \frac{1}{K} \sum_{k=0}^{K-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.9}$$

and elaborating on eqn.3.8

$$\log \Lambda = \frac{1}{K} \sum_{k=0}^{K-1} \{\gamma_k - \log(\gamma_k) - 1\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.10}$$

where η denotes the threshold of decision.

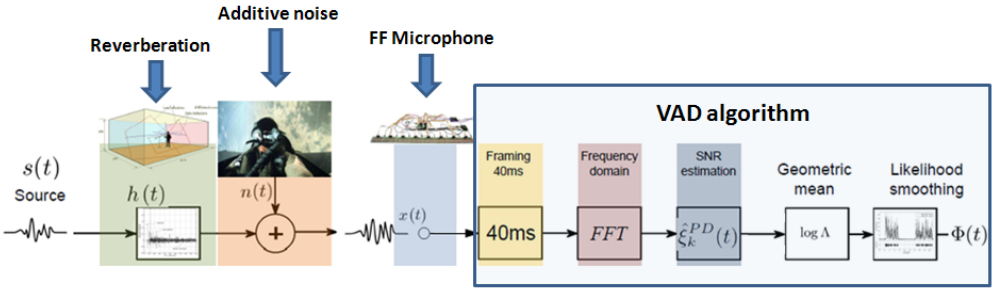

Figure 3.1: Likelihood Ratio Test based VAD system model.

The performance of such VAD systems is significantly downgraded if far-field microphones are used instead of the conventional close-talking ones due to reverberation effects, competitive sound sources, and speaker movement that can alter the shape of speech distribution as shown in Chapter 2. Furthermore, speech distribution varies with time and can be affected by several unpredictable factors including speaker's temper, mood, gender, age, and more.

Towards the direction of encapsulating speech shaped distribution in a VAD scheme Gazor (24) implemented an LRT-based VAD that combined a Laplacian speech model along with a Gaussian noise model. Furthermore Shin et al. extended the work on speech statistical modelling to a VAD that is based on the GΓD model (30, 31). The core of the system was based on the LRT of two GΓD modelling speech and noise. The parameters of each distribution were tuned based on the statistics of the input. Authors in (32) enriched this work using it in conjunction with the Bayesian Information Criterion. Nevertheless, the core of the system in (30) was not completely autonomous since a set of parameters had to be tuned manually for different noise conditions.

Chang et al. (14) extended the concept by introducing a VAD able to switch between a set of statistical models. The switching mechanism (that relies on the evaluation of Kolmogorov-Smirnov (KS) test of three theoretical distributions against the distribution of the current input) enabled the adaptation to environmental changes by modifying the employed LRT.

Although, based on the observations of Chapter 2 it becomes apparent that speech distribution is not solely Gaussian, Laplacian, or Gamma, given its nonstationarity in time and its dependence on external interferences. A weighted sum of a set of distributions would be more accurate in representing the instant distribution of captured speech. Such an approach allows better modelling of speech distribution in frequency and time as environmental characteristics evolve.

3.3 Convex Combination of Multiple Statistical Models for VAD

To overcome problems of adaptability to the non-stationarity of speech a VAD system able to adapt to varying conditions of noise and reverberation has been developed. The speech modelling is supported study on statistical speech characteristics and their dependence on external noise and reverberation when using far-field sensors. The core of the system is a convex combination of three distributions, a Gaussian (GD), a Laplacian (LD), and a two-sided Gamma Distribution (TΓD) (23, 24). Each distribution is selected to model captured speech under different scenarios and the assumption of GD noise. The participating pdfs are presented below.

3.3.1 Probability Distribution of Noise

Following earlier approaches, it is assumed that both the real and the imaginary parts of noise frequency components are zero mean following GD. The pdf of $\mathbf{N}_k(t)$ for the case of noise with k denoting the frequency bin is given by

$$f_n^G(N_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{N_k(t)^2}{2\sigma_{n,k}^2}} \tag{3.11}$$

where $\sigma_{n,k}^2$ is slowly varying with time variance factor of the Gaussian assumed distributed noise for the k^{th} frequency component. The imaginary part follows a similar distribution.

3.3.2 Probability Distribution of Speech

Embarking on the outcomes of the analysis of *Section II*, it is assumed that both the real and the imaginary parts of the frequency distribution of captured speech are better modelled as a mixture of a TΓD, a LD, and a GD. These pdfs are given by

$$\text{GD} : f_s^G(S_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{s,k}^2}} e^{-\frac{S_k(t)^2}{2\sigma_{s,k}^2}} \tag{3.12}$$

$$\text{LD} : f_s^L(S_k(t)) = \frac{1}{2\alpha_{s,k}} e^{-\frac{|S_k(t)|}{\alpha_{s,k}}} = \frac{1}{\sqrt{2}\sigma_{s,k}} e^{-\frac{\sqrt{2}|S_k(t)|}{\sigma_{s,k}}} \tag{3.13}$$

$$\text{T}\Gamma\text{D} : f_s^\Gamma \frac{\sqrt[4]{3}}{2\sqrt{\pi\sigma_{s,k}}\sqrt[4]{2}} |S_k(t)|^{-\frac{1}{2}} e^{-\frac{\sqrt{3}|S_k(t)|}{\sqrt{2}\sigma_{s,k}}} \tag{3.14}$$

for the k^{th} frequency component, where $\sigma_{s,k}^2$ and $a_{s,k}$ are the slowly varying variance and scale factors of the Gaussian- and Laplacian-distributed speech, respectively. The scale factor $a_{s,k}$ is related to variance $\sigma_{s,k}^2$ through

$$\sigma_{s,k}^2 = 2a_{s,k}^2 \tag{3.15}$$

3.3.3 Conditional Distributions

Using the predefined statistical models of voice and assuming Gaussian noise, the conditional pdfs of speech absence can be expressed as

$$H_0 : f_{X|H_0}(X_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{\mathbf{X}_k(t)^2}{2\sigma_{n,k}^2}} \tag{3.16}$$

Given the three different models of speech, a set of speech presence hypotheses is derived

<u>*Case 1*</u>: H_1 for *Gaussian* speech model

$$\begin{aligned} H_1 : f_{X|H_1}^{(G)}(X_k(t)) = & \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_{s,k}\sigma_{n,k}} e^{-\frac{\mathbf{S}_k(t)^2}{2\sigma_{s,k}^2} - \frac{(\mathbf{X}_k(t)-\mathbf{S}_k(t))^2}{2\sigma_{n,k}^2}} dS_k(t) \\ &= \frac{1}{\sqrt{2\pi}\sigma_{s,k}\sigma_{n,k}\sqrt{\frac{1}{{\sigma_{n,k}}^2} + \frac{1}{{\sigma_{n,k}}^2}}} e^{-\frac{X_k(t)^2}{2({\sigma_{n,k}}^2 + {\sigma_{s,k}}^2)}} \end{aligned} \tag{3.17}$$

Case 2: H_1 for *Laplacian* speech model

$$
\begin{aligned}
H_1 : f^{(L)}_{X|H_1}(X_k(t)) = & \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2}\sigma_{s,k}\sqrt{2\pi\sigma^2_{n,k}}} e^{-\frac{\sqrt{2}|S_k(t)|}{\sigma_{s,k}} - \frac{(X_k(t)-S_k(t))^2}{2\sigma^2_{n,k}}} dS_k(t) \\
&= \int_{-\infty}^{0} \frac{1}{\sqrt{2}\sigma_{s,k}\sqrt{2\pi\sigma^2_{n,k}}} e^{\frac{\sqrt{2}S_k(t)}{\sigma_{s,k}} - \frac{(X_k(t)-S_k(t))^2}{2\sigma^2_{n,k}}} dS_k(t) \\
&\quad + \int_{0}^{\infty} \frac{1}{\sqrt{2}\sigma_{s,k}\sqrt{2\pi\sigma^2_{n,k}}} e^{\frac{-\sqrt{2}S_k(t)}{\sigma_{s,k}} - \frac{(X_k(t)-S_k(t))^2}{2\sigma^2_{n,k}}} dS_k(t) \qquad (3.18) \\
&= \frac{1}{2\sqrt{2}\sigma_{s,k}} e^{\frac{{\sigma_{n,k}}^2}{{\sigma_{n,k}}^2}} \times \\
&\qquad \left[e^{\frac{\sqrt{2}X_k(t)}{\sigma_{s,k}}} \operatorname{erfc}\left(\frac{\sqrt{2}{\sigma_{n,k}}^2 + \sigma_{s,k}X_k(t)}{\sqrt{2}\sigma_{n,k}\sigma_{s,k}} \right) \right. \\
&\qquad \left. + e^{\frac{-\sqrt{2}X_k(t)}{\sigma_{s,k}}} \operatorname{erfc}\left(\frac{\sqrt{2}{\sigma_{n,k}}^2 - \sigma_{s,k}X_k(t)}{\sqrt{2}\sigma_{n,k}\sigma_{s,k}} \right) \right]
\end{aligned}
$$

Case 3: H_1 for TΓD speech model

$$
\begin{aligned}
H_1 : f^{(\Gamma)}_{X|H_1}(X_k(t)) = & \\
&= \int_{-\infty}^{\infty} \frac{\sqrt[4]{3}}{4\pi\sqrt[4]{2}\sqrt{\sigma_{s,k}}\sigma_{n,k}} |S_k(t)|^{-\frac{1}{2}} \\
&\quad \times e^{-\frac{\sqrt{3}|S_k(t)|}{\sqrt{2}\sigma_{s,k}} - \frac{(X_k(t)-S_k(t))^2}{2\sigma^2_{n,k}}} dS_k(t) \\
&= \frac{\sqrt[4]{3}}{4\pi\sqrt[4]{2}\sqrt{\sigma_{s,k}}\sigma_{n,k}} e^{-\frac{X_k(t)^2}{2\sigma^2_{n,k}}} \times \qquad (3.19) \\
&\qquad \left[\int_{-\infty}^{0} \frac{e^{\left(\frac{X_k(t)S_k(t)-2S_k(t)^2}{{\sigma_{n,k}}^2} + \frac{\sqrt{3}S_k(t)}{2\sigma_{s,k}} \right)}}{\sqrt{-S_k(t)}} dS_k(t) \right. \\
&\qquad \left. + \int_{0}^{\infty} \frac{e^{\left(\frac{X_k(t)S_k(t)-2S_k(t)^2}{{\sigma_{n,k}}^2} - \frac{\sqrt{3}S_k(t)}{2\sigma_{s,k}} \right)}}{\sqrt{S_k(t)}} dS_k(t) \right]
\end{aligned}
$$

The value of the first integral of eqn. (3.19) when $\frac{\sqrt{3}\sigma_{n,k}^2}{\sigma_{s,k}} > 2X_k(t)$ becomes:

$$\int_{-\infty}^{0} \frac{e^{\left(\frac{X_k(t)S_k(t)-2S_k(t)^2}{{\sigma_{n,k}}^2}+\frac{\sqrt{3}S_k(t)}{2\sigma_{s,k}}\right)}}{\sqrt{-S_k(t)}} dS_k(t) = \\ = \frac{\pi\sqrt{-\sigma_{s,k}N^{(+)}}}{2\sqrt{2}\sigma_{s,k}} e^{\frac{N^{(+)^2}}{D}} \\ \times \left(\jmath_a\left[-\frac{1}{4}, \frac{N^{(+)^2}}{D}\right] + \jmath_a\left[\frac{1}{4}, \frac{N^{(+)^2}}{D}\right]\right) \tag{3.20}$$

where $N^{(\pm)} = \sqrt{3}{\sigma_{n,k}}^2 \pm 2\sigma_{s,k}X_k(t)$, $D = (4\sigma_{n,k}\sigma_{s,k})^2$ and $\jmath_a$, the modified Bessel function of the first kind.

If $\frac{\sqrt{3}{\sigma_{n,k}}^2}{\sigma_{s,k}} \leq 2X_k(t)$, the value of the first integral of eqn(3.19) becomes:

$$\int_{-\infty}^{0} \frac{e^{\left(\frac{X_k(t)S_k(t)-2S_k(t)^2}{{\sigma_{n,k}}^2}+\frac{\sqrt{3}S_k(t)}{2\sigma_{s,k}}\right)}}{\sqrt{-S_k(t)}} dS_k(t) = \\ = \frac{1}{2} e^{\frac{N^{(+)^2}}{D}} \sigma_{n,k}\sqrt{\frac{\sqrt{3}}{\sigma_{s,k}} + \frac{2X_k(t)}{\sigma_{n,k}^2}} \curlyvee_a\left[\frac{1}{4}, \frac{N^{(+)^2}}{D}\right] \tag{3.21}$$

where $\curlyvee_a$ denotes the modified Bessel function of the second kind.

The value of the second integral of eqn. (3.19) when $\frac{\sqrt{3}\sigma n^2}{\sigma_{s,k}} < 2X_k(t)$ becomes:

$$\int_{0}^{\infty} \frac{e^{\left(\frac{X_k(t)S_k(t)-2S_k(t)^2}{{\sigma_{n,k}}^2}-\frac{\sqrt{3}S_k(t)}{2\sigma_{s,k}}\right)}}{\sqrt{S_k(t)}} dS_k(t) = \\ = \frac{\pi\sqrt{-\sigma_{s,k}N^{(-)}}}{2\sqrt{2}\sigma_{s,k}} e^{\frac{N^{(-)^2}}{D}} \\ \times \left(\jmath_a\left[-\frac{1}{4}, \frac{N^{(-)^2}}{D}\right] + \jmath_a\left[\frac{1}{4}, \frac{N^{(-)^2}}{D}\right]\right) \tag{3.22}$$

If $\frac{\sqrt{3}\sigma n^2}{\sigma_{s,k}} \geq 2X_k(t)$, the value of the second integral of eqn(3.19) becomes:

$$\int_{0}^{\infty} \frac{e^{\left(\frac{X_k(t)S_k(t)-2S_k(t)^2}{{\sigma_{n,k}}^2}-\frac{\sqrt{3}S_k(t)}{2\sigma_{s,k}}\right)}}{\sqrt{S_k(t)}} dS_k(t) = \\ = \frac{1}{2} e^{\frac{N^{(-)^2}}{D}} \sigma_{n,k}\sqrt{\frac{\sqrt{3}}{\sigma_{s,k}} - \frac{2X_k(t)}{\sigma_{n,k}^2}} \curlyvee_a\left[\frac{1}{4}, \frac{N^{(-)^2}}{D}\right] \tag{3.23}$$

3.4 Forming the LRTs based on the Distribution of Speech

Embarking on the likelihood ratio of the two hypotheses, for the distribution of speech and noise, a decision statistic is defined by substituting the appropriate speech and noise pdf in the general LRT. This is defined as

$$\Lambda_k \equiv \frac{f_{X_k(t)|H_1}(X_k(t))}{f_{X_k(t)|H_0}(X_k(t))} \tag{3.24}$$

where $f_{X_k(t)|H_1}(X_k(t))$ is the hypothesis of speech presence H_1 and $f_{X_k(t)|H_0}(X_k(t))$ is the hypothesis of speech absence H_0 under the assumption of Gaussian distributed noise both defined in *Section III C*. For the case of Gaussian distributed speech, eqn. (3.24) becomes

$$\begin{aligned}\Lambda_k^G =& \frac{f^{(G)}_{X_k(t)|H_1}(X_k(t))}{f_{X|H_0}(X_k(t))} \\ =& \frac{\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_{s,k}\sigma_{n,k}} e^{-\frac{S_k(t)^2}{2\sigma_{s,k}^2} - \frac{(X_k(t)-S_k(t))^2}{2\sigma_{n,k}^2}} dS_k(t)}{\frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{X_k(t)^2}{2\sigma_{n,k}^2}}}\end{aligned} \tag{3.25}$$

When speech distribution follows LD, the LR becomes

$$\begin{aligned}\Lambda_k^L =& \frac{f^{(L)}_{X_k(t)|H_1}(X_k(t))}{f_{X|H_0}(X_k(t))} \\ =& \frac{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2}\sigma_{s,k}\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{\sqrt{2}|S_k(t)|}{\sigma_{s,k}} - \frac{(X_k(t)-S_k(t))^2}{2\sigma_{n,k}^2}} dS_k(t)}{\frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{X_k(t)^2}{2\sigma_{n,k}^2}}}\end{aligned} \tag{3.26}$$

Finally for the case of TTD distributed speech the LR becomes

$$\begin{aligned}\Lambda_k^\Gamma =& \frac{f^{(\Gamma)}_{X_k(t)|H_1}(X_k(t))}{f_{X|H_0}(X_k(t))} \\ =& \frac{\int_{-\infty}^{\infty} \frac{\sqrt[4]{3}}{4\pi\sqrt[4]{2}\sqrt{\sigma_{s,k}}\sigma_{n,k}|S_k(t)|} e^{-\frac{\sqrt{3}|S_k(t)|}{\sqrt{2}\sigma_{s,k}} - \frac{(X_k(t)-S_k(t))^2}{2\sigma_{n,k}^2}} dS_k(t)}{\frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{X_k(t)^2}{2\sigma_{n,k}^2}}}\end{aligned} \tag{3.27}$$

3.4.1 The Convex Combination Scheme

Since speech is a dynamically changing signal, whose characteristics are highly altered when propagated within dynamic environments, its distribution cannot be solely Laplacian, Gaussian, or Gamma. It is rather a combination of these distributions. Thus, instead of forcing the system to generate decision under single model hypothesis (14), a linear combination of GD, LD, and TΓD is proposed

$$\begin{aligned}\Lambda_k^{Convex} \equiv & w_G\Lambda_k^G + w_L\Lambda_k^L + w_\Gamma\Lambda_k^\Gamma \\ = & w_G\frac{f^{(G)}_{X_k(t)|H_1}(X_k(t))}{f_{X|H_0}(X_k(t))} + w_L\frac{f^{(L)}_{X_k(t)|H_1}(X_k(t))}{f_{X|H_0}(X_k(t))} + w_\Gamma\frac{f^{(\Gamma)}_{X_k(t)|H_1}(X_k(t))}{f_{X|H_0}(X_k(t))} \\ = & \frac{w_G f^{(G)}_{X_k(t)|H_1}(X_k(t)) + w_L f^{(L)}_{X_k(t)|H_1}(X_k(t)) + w_\Gamma f^{(\Gamma)}_{X_k(t)|H_1}(X_k(t))}{f_{X|H_0}(X_k(t))}\end{aligned} \tag{3.28}$$

where w_G, w_L, w_Γ, are the weights of the employed Gaussian, Laplacian and TΓD models respectively, derived through the evaluation of the KS-test. Elaborating eqn. (3.28) yields

$$\Lambda_k^{Convex} \equiv \frac{\int_{-\infty}^{\infty}\left(w_G f_s^G + w_L f_s^L + w_\Gamma f_s^\Gamma\right) f_n^G dS_k(t)}{f_{X|H_0}(X_k(t))} \tag{3.29}$$

The decision rule that is used to determine voice presence is established as the geometric mean of the LRs of every frequency bin. Notice that both the real and imaginary parts of DFT are taken into consideration. This is done by combining the corresponding LRs using a geometric mean. Thus, decision is drawn based on

$$\mathcal{D} = \frac{1}{K}\sum_{k=0}^{K-1}\log\Lambda_k^{Convex} \underset{H_0}{\overset{H_1}{\gtrless}} \theta \tag{3.30}$$

where K denotes the total number of frequency bins and θ the decision threshold.

3.4.2 Estimating the Weights for the Convex Model

To develop a set of weights for each pdf in the mixture model of eqn. (3.29), which will reflect the participation of each distribution in the convex combination, the KS-test distance of the input from the three theoretical distributions is evaluated for every frequency bin presented in Chapter 2 . The weights for each distribution in the convex combination will be

$$w_{G,k} = \left(n_k \cdot \hat{T}_{G,k}\right)^{-1} \quad \text{GD weight} \tag{3.31}$$

$$w_{L,k} = \left(n_k \cdot \hat{T}_{L,k}\right)^{-1} \quad \text{LD weight} \tag{3.32}$$

$$w_{\Gamma,k} = \left(n_k \cdot \hat{T}_{\Gamma,k}\right)^{-1} \quad \text{TΓD weight} \tag{3.33}$$

where n_k is a normalization factor for the weights so that they all sum to 1, given by

$$n_k = \hat{T}_{G,k}^{-1} + \hat{T}_{L,k}^{-1} + \hat{T}_{\Gamma,k}^{-1} \tag{3.34}$$

and $\hat{T}_{G,k}$, $\hat{T}_{L,k}$, $\hat{T}_{\Gamma,k}$ the smoothed estimates of the KS distances for each distribution in the convex combination. They are estimated by

$$\hat{T}_t = \lambda_T \hat{T}_{t-p} + (1 - \lambda_T)\, T_t\,(X_m) \tag{3.35}$$

with λ_T being the memory parameter. Parameters m and p denote the memory and step sizes, respectively. They are set to 20 and 10, respectively, as proposed by (14) to achieve fast convergence without dramatically increasing the computational load. The KS distance T of the captured speech signal from each distribution is evaluated on the recent input data $X_m = \{X(t), X(t-1), ...X(t-m)\}$ every p frames for m previous values. The distance of the input from each distribution is defined as

$$T_{G,k} = \max\left(|F_{X,k} - F_{G,k}|\right) \quad \text{KS distance from GD} \tag{3.36}$$

$$T_{L,k} = \max\left(|F_{X,k} - F_{L,k}|\right) \quad \text{KS distance from LD} \tag{3.37}$$

$$T_{\Gamma,k} = \max\left(|F_{X,k} - F_{\Gamma,k}|\right) \quad \text{KS distance from TΓD} \tag{3.38}$$

where $F_{X,k}$ is the empirical cdf of the current input for the k^{th} frequency bin and $F_{G,k}$,$F_{L,k}$,$F_{\Gamma,k}$, the theoretical cdfs.

3.5 SNR Estimation

An essential intermediate step toward the evaluation of the individual models of the convex LRT is the estimation of the *a priori* SNR. Thus, the values of speech and noise power spectrum have to be continuously tracked. The most popular way to do this is to apply the *Decision Directed (DD)* algorithm (29). The authors in (14) presented an alternative method, namely *Predicted Estimation (PD)*. This can overcome the limitations of *DD* within speech activity intervals. According to *PD* method, the *a priori* SNR is estimated on the power spectrum of noise $\lambda_{n,k}(t) = \sigma_{n,k}(t)^2$ and speech $\lambda_{s,k}(t) = \sigma_{s,k}(t)^2$ which are given by

$$\begin{aligned}\hat{\lambda}_{n,k}(t+1) &= \zeta_n \hat{\lambda}_{n,k}(t) + (1-\zeta_n) E\left[|N_k(t)|^2 |X_k(t)\right] \\ \hat{\lambda}_{s,k}(t+1) &= \zeta_s \hat{\lambda}_{s,k}(t) + (1-\zeta_s) E\left[|S_k(t)|^2 |X_k(t)\right]\end{aligned} \tag{3.39}$$

where $\hat{\lambda}_{s,k}(t), \hat{\lambda}_{n,k}(t)$ are estimates of $\lambda_{s,k}(t), \lambda_{n,k}(t)$ and ζ_n ,ζ_s are smoothing parameters both set to 0.99. Following similar considerations as in (14), the expectations in eqn(3.39) can be further analyzed as:

$$\begin{aligned}E\left[|N_k(t)|^2 |X_k(t)\right] =& E\left[|N_k(t)|^2 |X_k(t), H_0\right] p\left(H_0 | X_k(t)\right) \\ &+ E\left[|N_k(t)|^2 |X_k(t), H_1\right] p\left(H_1 | X_k(t)\right) \\ E\left[|S_k(t)|^2 |X_k(t)\right] =& E\left[|S_k(t)|^2 |X_k(t), H_0\right] p\left(H_0 | X_k(t)\right) \\ &+ E\left[|S_k(t)|^2 |X_k(t), H_1\right] p\left(H_1 | X_k(t)\right)\end{aligned} \tag{3.40}$$

where the expectations in speech absence periods are

$$\begin{aligned}E\left[|N_k(t)|^2 |X_k(t), H_0\right] &= |X_k(t)|^2 \\ E\left[|S_k(t)|^2 |X_k(t), H_0\right] &= 0\end{aligned} \tag{3.41}$$

and in the expectations within speech presence intervals

$$\begin{aligned}E\left[|N_k(t)|^2 |X_k(t), H_1\right] &= \frac{\hat{\lambda}_{s,k}(t)}{1+\hat{\xi}_k^{PD}(t)} + \frac{|X_k(t)|^2}{\left(1+\hat{\xi}_k^{PD}(t)\right)^2} \\ E\left[|S_k(t)|^2 |X_k(t), H_1\right] &= \frac{\hat{\lambda}_{s,k}(t)}{1+\hat{\xi}_k^{PD}(t)} + \frac{|X_k(t)|^2 \left(\hat{\xi}_k^{PD}(t)\right)^2}{\left(1+\hat{\xi}_k^{PD}(t)\right)^2}\end{aligned} \tag{3.42}$$

Combining eqn. (3.40), (3.41), and (3.42), the power spectrum of noise and speech are given by

$$\hat{\lambda}_{n,k}(t+1) = \zeta_n \hat{\lambda}_{n,k}(t) + (1-\zeta_n)\Bigg[p\left(H_0|X_k(t)\right)|X_k(t)|^2 + \left(\frac{\hat{\lambda}_{s,k}(t)}{1+\hat{\xi}_k^{PD}(t)} + \frac{|X_k(t)|^2}{\left(1+\hat{\xi}_k^{PD}(t)\right)^2}\right) p\left(H_1|X_k(t)\right)\Bigg] \tag{3.43}$$

$$\hat{\lambda}_{s,k}(t+1) = \zeta_s \hat{\lambda}_{s,k}(t) + (1-\zeta_s)\times \left(\frac{\hat{\lambda}_{s,k}(t)}{1+\hat{\xi}_k^{PD}(t)} + \frac{|X_k(t)|^2\left(\hat{\xi}_k^{PD}(t)\right)^2}{\left(1+\hat{\xi}_k^{PD}(t)\right)^2}\right) p\left(H_1|X_k(t)\right) \tag{3.44}$$

respectively, where the *a priori SNR* $\hat{\xi}^{PD}$ at time instant t is estimated as

$$\hat{\xi}_k^{PD}(t) \equiv \frac{\hat{\lambda}_{s,k}(t)}{\hat{\lambda}_{n,k}(t)} \tag{3.45}$$

and the speech absence probability is

$$p\left(H_0|X_k(t)\right) = \frac{1}{1+\frac{P(H_1)}{P(H_0)}\Lambda_k^{Convex}} \tag{3.46}$$

The speech presence probability is therefore given by

$$p\left(H_1|X_k(t)\right) = 1 - p\left(H_0|X_k(t)\right) \tag{3.47}$$

Notice that similar to earlier approaches, the proposed algorithm takes also into account both the real and the imaginary parts of the spectrum, by computing the geometrical mean in eqn. (3.30) using both parts of the complex spectrum. Thus, $|X_k(t)|^2$ depends on the complex part that is evaluated at every iteration that is either $\left|X_k^R(t)\right|^2$ or $\left|X_k^I(t)\right|^2$. The estimation of the variance of noise $\lambda_{n,k}(t) = \sigma_{n,k}(t)^2$ and speech $\lambda_{s,k}(t) = \sigma_{s,k}(t)^2$ is performed separately for real and imaginary frequency parts DFT based on eqn. (3.39).

3.6 Adaptive Estimation of Threshold

The LRTs employed here introduce, by definition, a bias towards speech detection H_1 (16). This is attributed to the fact that the model of noise (GD) is present both at the numerator and the denominator of the ratio in eqn. (3.25), (3.26), and (3.27). This bias introduces an offset, which tends to increase as SNR drops (higher noise). As proposed in (16), this could be tackled by

introducing a weight α to cope with the bias

$$\frac{p(X_k(t)|H_n = H_1)}{p(X_k(t)|H_n = H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \frac{P(H_n = H_0)}{P(H_n = H_1)} \tag{3.48}$$

where H_n denotes the correct hypothesis in the current frame and $\alpha \geq 1$ a small number catering to the offset.

Using this method, to cope with the offset requires the calculation of a different α value for several SNR conditions to be able to use a fixed threshold. Nevertheless, the proposed system can perform with a fixed threshold (for $\theta = 0.065$) for high SNR values down to 10dB. To overcome the bias for lower SNR values and to enhance the overall performance of the system an adaptive threshold is introduced. The value of the threshold is initialized to

$$\hat{\theta}(t) \equiv \max \left[1.2 \cdot \overline{N_{buf}}, \frac{\max\left(N_{buf}\right) + \overline{N_{buf}}}{2} \right] \tag{3.49}$$

where N_{buf} is a buffer holding past values of $\mathcal{D}$ and $\overline{N_{buf}}$ its mean (8). After initialization, the computation of the threshold is performed by

$$\hat{\theta}(t) \equiv 1.2 \cdot \left(\overline{N_{buf}} + 3 \cdot \sigma_{N_{buf}} \right) \tag{3.50}$$

where $\sigma_{N_{buf}}$ is the standard deviation of the values in N_{buf}, which is updated with values of $\mathcal{D}$ eqn. (3.30) smaller than that of $\hat{\theta}(t)$. For smoothing the threshold estimate $\hat{\theta_{sm}}$, a forgetting factor λ_θ is introduced

$$\hat{\theta_{sm}}(t+1) = \lambda_\theta \cdot \hat{\theta}(t) + (1 - \lambda_\theta)\, \hat{\theta}(t+1) \tag{3.51}$$

To further enhance the performance of the system and to cater to decisions that lack rational justification (i.e. very short speech segments), a hang-over scheme is employed. This is implemented as a state machine that has two major transition states (speech presence, speech absence) and several intermediate ones (15).

For transition from speech presence to speech absence state, at least $s_0 = 10$ consecutive indications of silence should be detected, while $s_1 = 4$ consecutive speech detections are required for the opposite. This lowers the probability of false rejections, by reducing the risk of the low-energy portion of speech at the end of an utterance being falsely rejected. This hangover scheme comes at a cost: it introduces a bias over speech detection.

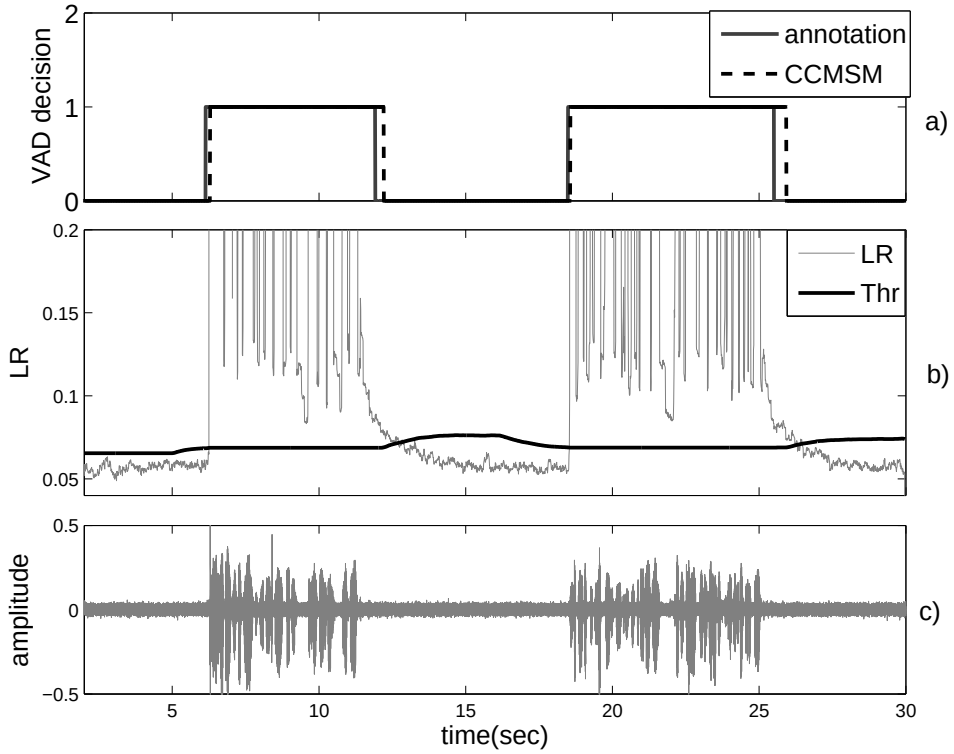

Figure 3.2: Decision based on the adaptive threshold. a) Decision and annotation limits, b) Response of the adaptive threshold given the likelihood ratio of CCMSM on c) Speech samples at 15 dB of babble noise.

3.7 Performance Discussion

Convex combination of multiple statistical models (CCMSM) VAD performance was compared to that of its core constituting single LR (GD, LD, TΓD) models as defined in Section 3.4, to the MSM VAD of (14), to the GΓD VAD of (30), and to the standardized G.729 annex B VAD. Additionally, the performance of a modified version of the proposed algorithm (SWMSM) that uses the switching scheme of (14) is also presented to assess the effectiveness of the convex combination scheme. A version of (14) with the Gaussian model of noise (MSM-G) is used to show that the reason for the difference in performance among the proposed solution and (14) is not produced by the fixed Gaussian model.

For a fair evaluation of the systems, the same thresholding technique, hangover scheme, and frame/step sizes have been used. For the case of GΓD VAD, the smoothing parameters of test statistics and their ratios, smoothing factors, and learning rates, for both noisy speech and noise,

were tuned manually, following the corresponding description in (30), for every different noise condition and intensity level. The system is not able to adjust those automatically and thus online adaptation to varying conditions for optimal performance cannot be achieved in contrary to the rest systems.

The P_c and P_f were evaluated using the speech recordings performed in the anechoic chamber of Aalborg University Denmark (25) described in section 2.2.

Speech data were contaminated artificially with white, vehicular, and babble noises from NOISEX-92 database (33). Their spectrum is presented in Fig. 3.3. Finally, the data were reverberated using the *Image Method* (27). For the evaluation of the system under combined phenomena of reverberation and noise, the source speech signals were first reverberated and then noise was added at different intensity levels. The input data were sampled at 8 kHz and were segmented into overlapping frames of 40 msec duration (10 msec step size).

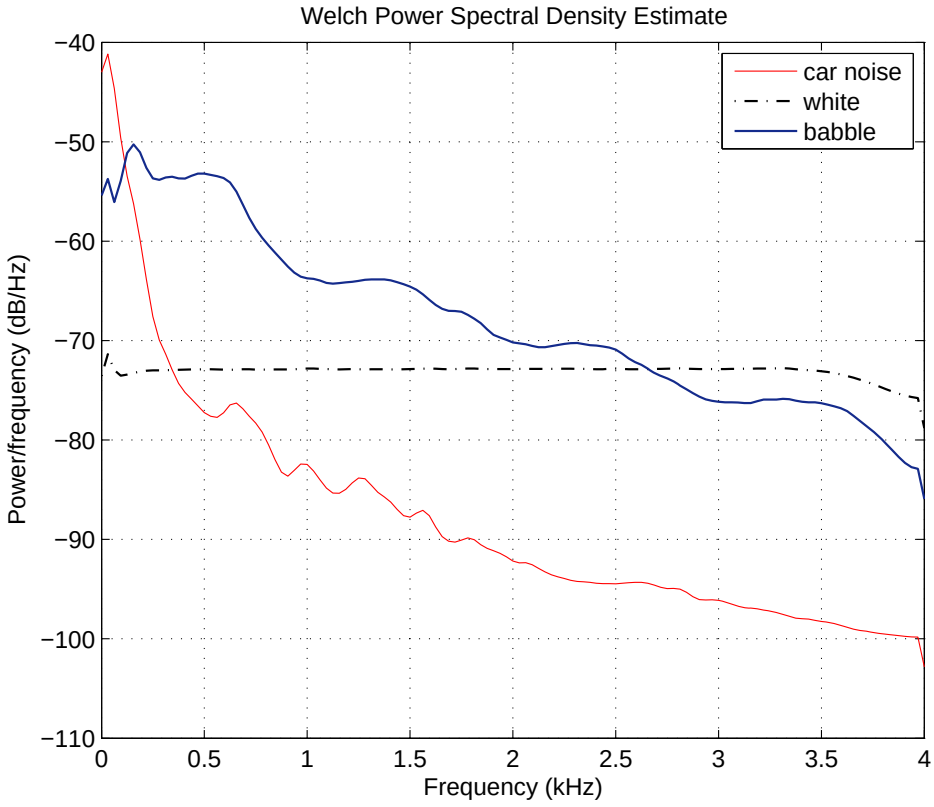

Figure 3.3: Spectrum of noises employed in the experiments

3.7.1 Performance under Additive Noise

Figure 3.4 depicts how the performance of core models and that of the proposed system varies as SNR drops due to AWGN. In this graph, it is shown that the performance of CCMSM surpasses that of the single modalities, with the Laplacian model getting closer for SNRs between 0 and 5 dB. The performance of TΓD is comparable to the CCMSM for only high SNR values. Something worth noting is the fact that the single modalities, proposed as core of our system, perform better than the MSM algorithm (14) and GΓD VAD (30). Additionally, SWMSM performs better than MSM-G, which follows the same system architecture, although it relies on differently defined statistical models as those defined here.

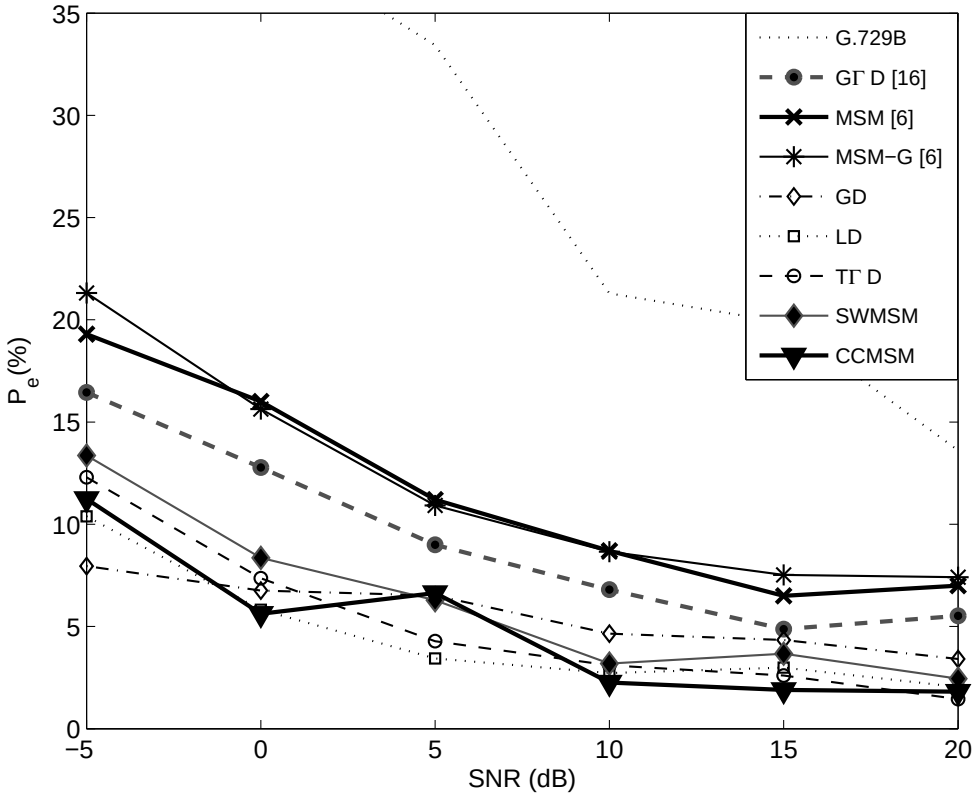

Figure 3.4: P_e performance under different intensities of White noise

Thus, the methodology of (18) for forming the LRT of Laplacian over Gaussian distribution seems to perform better. Additionally, the SWMSM VAD shows good performance, but due to the ripples in its response, caused by switching from one model to another, it fails to perform

better than LD and TΓD under high SNR values although being able to outperform the GD model LTR. SWMSM and MSM-G rely on the same switching architecture although the underlying statistical models of noise and speech are differently expressed. This is indicative of the performance enhancement, shown in Fig. 3.4, when following the approach of (18) that separately models the real and imaginary parts of the FFT and mixes them at the LTR stage as independent terms of the geometric mean in eqn. (25). This can also be observed for GΓD VAD (30) that is underperformed by single modality GD. Additionally, GΓD VAD (30) requires a set of parameters to be tuned manually for every different noise condition. This cannot be expected to be optimal for all cases and thus can lead to poor performance.

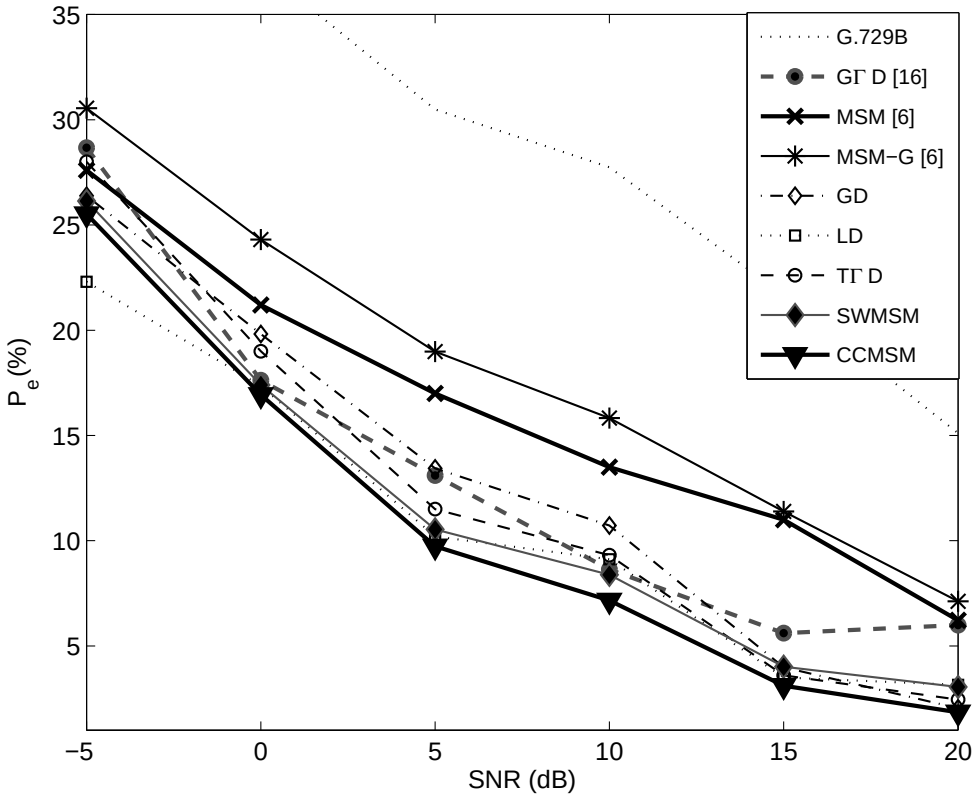

Figure 3.5: P_e performance under different intensities of babble noise

Table 3.1: Performance Results under Various Types of Noise

		CC-MSM VAD			SW-MSM VAD			MSM VAD (14)			MSM-G VAD (14)			GTD VAD (30)		
Noise	**SNR**	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%
AWGN	**20dB**	1.81	1.78	1.84	2.45	1.08	3.82	7.02	6.45	7.59	7.41	6.30	8.52	5.52	4.63	6.41
	15dB	1.90	2.07	1.73	3.67	4.31	2.15	6.47	7.81	5.13	7.52	8.11	6.94	4.87	7.60	2.13
	10dB	2.26	2.55	1.97	3.19	3.93	2.13	8.73	8.13	9.33	8.66	8.00	9.31	6.80	9.16	4.43
	5dB	6.64	2.58	10.71	6.28	4.25	8.31	11.19	5.64	16.74	10.92	6.25	15.59	8.99	12.21	5.77
	0dB	5.62	8.62	2.63	8.36	7.11	9.61	16.03	9.15	22.92	15.63	10.66	20.60	12.78	15.91	9.66
	-5dB	11.24	13.44	9.03	13.36	14.56	12.16	19.31	12.40	26.22	21.31	11.89	30.73	16.45	17.40	15.51
Babble	**20dB**	1.85	2.01	1.68	3.05	3.09	3.01	6.22	4.18	8.26	7.12	5.34	8.90	5.99	8.59	3.38
	15dB	3.11	3.21	3.01	4.01	4.83	3.19	11.09	5.22	16.96	11.39	5.71	17.07	5.61	6.11	5.11
	10dB	7.16	8.11	6.22	8.38	7.58	9.18	13.52	8.75	18.29	15.83	10.43	21.23	8.63	9.02	8.24
	5dB	9.74	9.97	9.50	10.54	9.95	11.14	16.94	9.34	24.54	18.99	12.09	25.89	13.10	12.88	13.31
	0dB	16.93	16.41	17.46	17.42	16.46	18.38	21.23	14.18	28.28	24.31	16.68	31.94	17.62	16.59	18.64
	-5dB	25.52	31.82	19.22	26.14	24.79	27.50	27.60	23.92	31.28	30.55	23.92	37.18	28.67	37.08	20.25
Vehicle	**20dB**	1.88	1.94	1.82	3.34	2.41	4.27	6.74	5.33	8.15	7.93	5.96	9.91	4.48	3.52	5.43
	15dB	3.09	1.71	4.47	4.23	4.34	4.12	7.12	6.29	7.95	8.11	5.44	10.78	3.62	2.74	4.51
	10dB	4.64	3.00	6.28	8.33	7.15	9.51	9.93	7.38	12.48	10.03	8.21	11.85	6.57	4.24	8.90
	5dB	9.61	7.98	11.24	10.52	7.45	13.59	10.87	9.80	11.94	12.57	9.47	15.67	10.54	9.04	12.03
	0dB	15.88	17.96	13.81	14.50	14.36	14.64	14.65	11.87	17.43	16.67	14.19	19.13	11.93	7.83	16.06
	-5dB	20.74	9.54	31.94	19.04	16.87	21.20	17.21	10.43	23.99	21.32	17.88	24.76	16.06	8.86	23.26

Babble noise (Fig. 3.5) is considered to be one of the most adverse conditions for VAD (17). In this case, the conclusions are similar to those for AWGN. The performance for all the assessed systems begins degrading significantly under 5 dB. CCMSM performs better than the rest in almost all cases. SWMSM shows very good performance, and especially below 5 dB it almost matches the performance of CCMSM. GΓD VAD (30) performs better than MSM and as SNR drops, it tends to perform better than MSM and as SNR drops, it tends to perform better than the employed GD single modality.

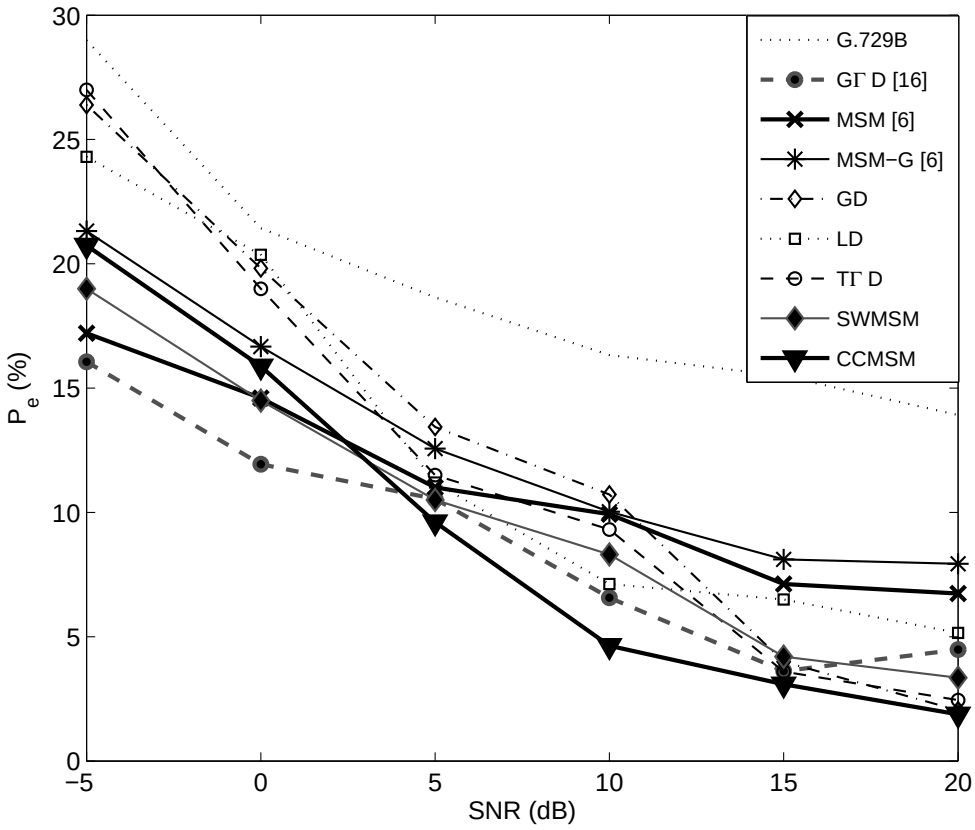

Figure 3.6: P_e performance under different intensities of vehicular noise

In the case of car noise (Fig. 3.6), the performance of MSM (14) and that of GΓD VAD (30) are closer to that of single modalities we use, comparing with the case of white noise, and under 0 dB they perform better than CCMSM and SWMSM. In fact, this is reasonable enough given that the core of CCMSM and SWMSM is based on the assumption of Gaussian noise, something not really efficient when the statistics of noise start to deviate a lot from the GD approaching other distributions.

The distribution of the specific car noise is closer to Laplacian, which complies with the fundamental assumption of (14) (speech and noise have the same distribution). However, this algorithm shows better performance than CCMSM only under severe conditions (SNR<0 dB).

It is interesting that although CCMSM is governed by the bad performance of its single modalities below 5 dB, it does not fail significantly when compared to MSM (14). SWMSM also works on average better than its single modalities.

The performance of the two combination schemes has been further examined. Figure 3.7 shows the likelihood that emerges from CCMSM and SWMSM when fed the same input (speech contaminated by additive car noise at 15 dB). The middle graph illustrates the log-likelihood difference of the two systems. It is evident that CCMSM produces higher likelihoods than SWMSM, indicating its increased modelling abilities. The conclusions are also supported by Table I.

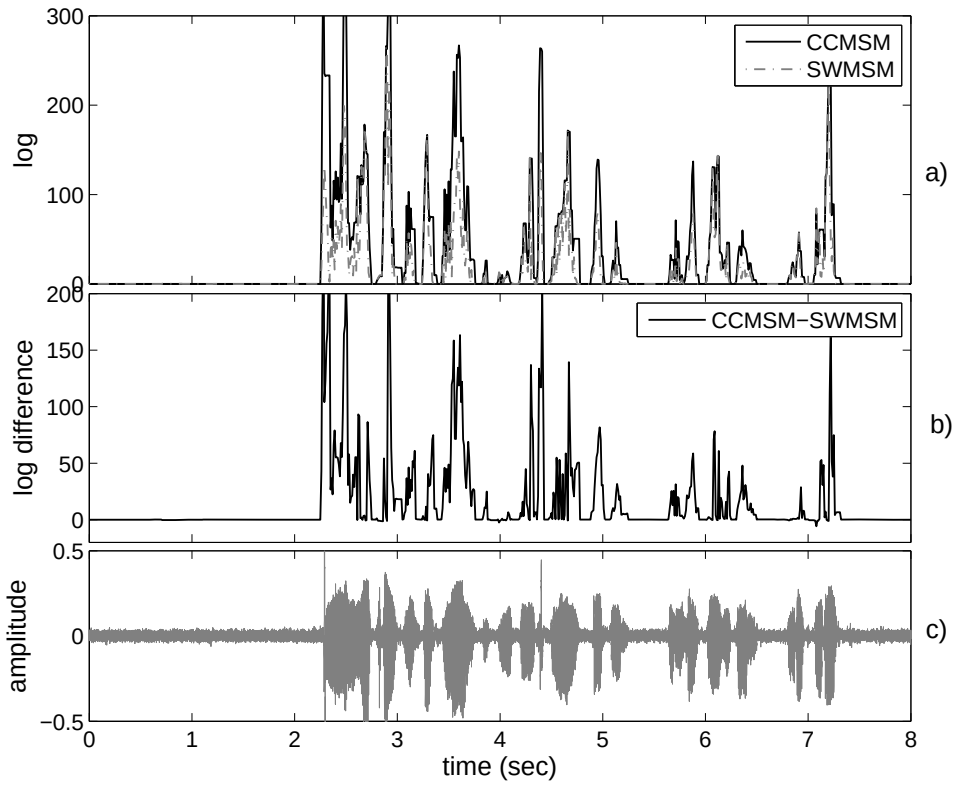

Figure 3.7: a) Likelihoods of SWMSM and CCMSM given a speech input at 15dB of vehicular noise, b) Likelihood difference between CCMSM and SWMSM responses, c) Captured speech sample

3.7.2 Performance within Reverberant Environments

The modification of captured speech statistics triggered by the presence of reverberation has an impact on estimated LRT and thus on the performance of the proposed VAD system. Somewhat surprisingly, this impact is not necessarily negative. The 'spreading' of speech content, through the presence of reverberation, facilitates the operation of VAD in some cases, although in other cases it degrades its performance. Indeed, reverberation reduces the clipping error effect (P_c), which is observed when VAD systems operate on long speech intervals, acting as a sophisticated hang over scheme. The prolonging of phrases caused by reverberation, however, might result in late detection of the offset of speech signals and thus increase P_f, especially for large T_{60} values. Therefore, its overall impact on captured speech cannot be usually determined *a priori* since it depends on a number of parameters ranging from the reverberation time to the percentage of speech within the acoustic signal. Similar performance enhancing effects due to room reflections have been presented also in (34), where the authors examine the performance of ASR within reverberant environments.

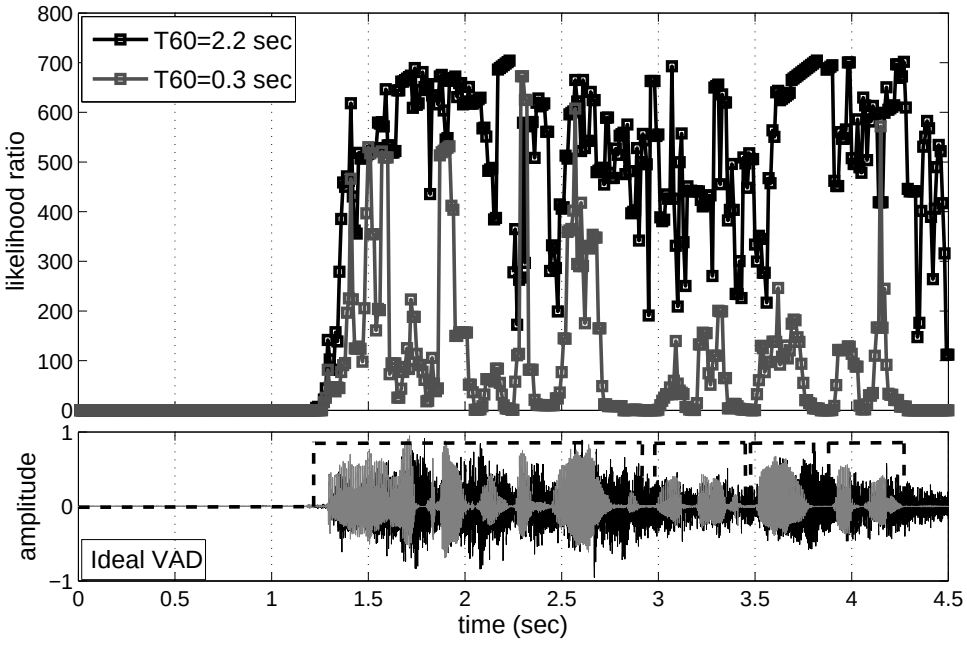

Figure 3.8: Likelihood ratio of CCMSM for reverberation times 0.3 and 2 sec. The LRT is significantly enhanced for the increased reverberation case.

In our experiments, we employed the audio signals recorded at AAU anechoic chamber and reverberated artificially using the *Image Method* (27) and we observed that for reverberation time values up to 2 sec the P_c of the proposed CCMSM algorithm is reduced due to reverberation

for speech intervals. This can also be observed in Fig. 3.8, where the likelihood is higher for highly reverberated speech ($T_{60} = 2.0$ sec) and short silence intervals within words are masked by reverberation allowing the usage of more distinct thresholds of decision.

Table 3.2: Performance Results under Reverberation and Noise

		CCMSM w/ fixed θ			CCMSM w/adapt. θ		
Noise	T_{60}	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%
25dB	**0.1s**	2.04	1.23	2.86	3.08	0.81	5.35
	0.3s	2.24	1.17	3.30	3.44	1.09	5.78
	0.6s	3.09	1.53	4.64	4.56	1.78	7.33
	1.0s	4.14	1.62	6.66	5.80	1.90	9.71
	1.5s	5.23	1.60	8.87	6.87	1.92	11.82
	2.0s	6.13	1.60	10.66	7.54	1.91	13.18
20db	**0.1s**	2.14	2.21	2.08	2.62	3.30	1.95
	0.3s	2.84	1.96	3.72	2.63	2.98	2.29
	0.6s	3.19	1.91	4.48	2.67	1.95	3.38
	1.0s	5.17	1.66	8.67	2.31	2.43	2.20
	1.5s	6.24	1.65	10.83	2.94	2.38	3.49
	2.0s	6.47	1.68	11.26	3.90	2.40	5.39
15db	**0.1s**	2.27	2.27	2.26	2.49	1.97	3.00
	0.3s	2.89	1.90	3.87	2.82	1.01	4.64
	0.6s	4.70	1.72	7.68	3.18	0.64	5.73
	1.0s	4.73	1.38	8.08	2.16	1.62	2.69
	1.5s	5.37	1.19	9.55	3.61	2.61	4.61
	2.0s	8.97	1.72	16.22	5.37	2.74	7.99
10db	**0.1s**	2.48	2.52	2.45	2.47	3.89	1.05
	0.3s	4.77	2.74	6.80	2.63	1.65	3.60
	0.6s	4.88	2.05	7.70	2.81	1.07	4.55
	1.0s	7.56	2.63	12.48	5.16	2.74	7.58
	1.5s	7.90	2.56	13.25	6.47	2.24	10.69
	2.0s	9.35	2.61	16.08	4.91	2.34	7.48
5db	**0.1s**	5.36	5.65	5.07	4.46	0.66	8.25
	0.3s	6.34	4.52	8.16	3.52	1.46	5.58
	0.6s	7.12	4.93	9.32	2.84	1.04	4.64
	1.0s	8.59	2.94	14.24	4.66	2.88	6.43
	1.5s	9.44	3.17	15.71	2.89	1.58	4.21
	2.0s	11.83	5.24	18.42	2.71	2.15	3.28

To evaluate performance of CCMSM under noisy and reverberant conditions, the proposed algorithm has been applied to audio signals that have been initially reverberated artificially and then contaminated with additive noise. The results are illustrated in Fig. 3.9 where it is shown

that the dependence of P_e on the SNR and T_{60} is not monotonic.

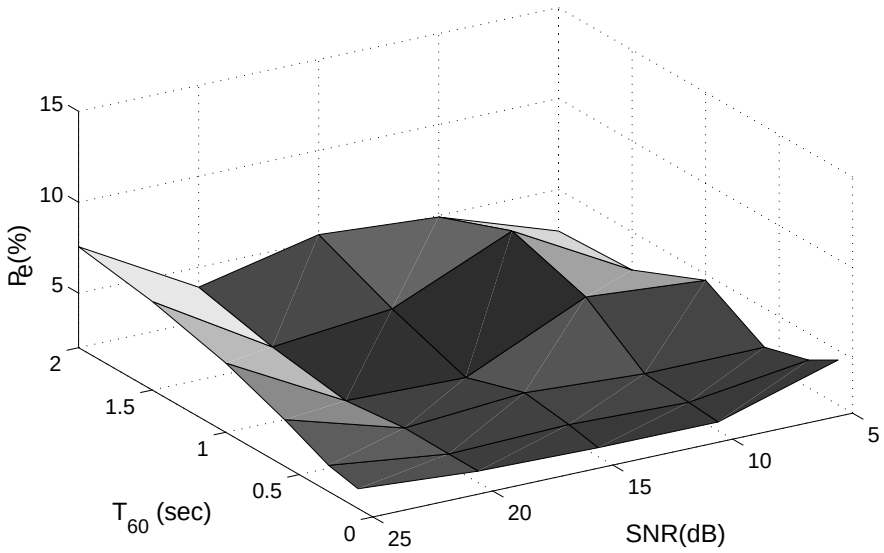

Figure 3.9: Performance of the system under simultaneous phenomena with adaptive thresholding enabled

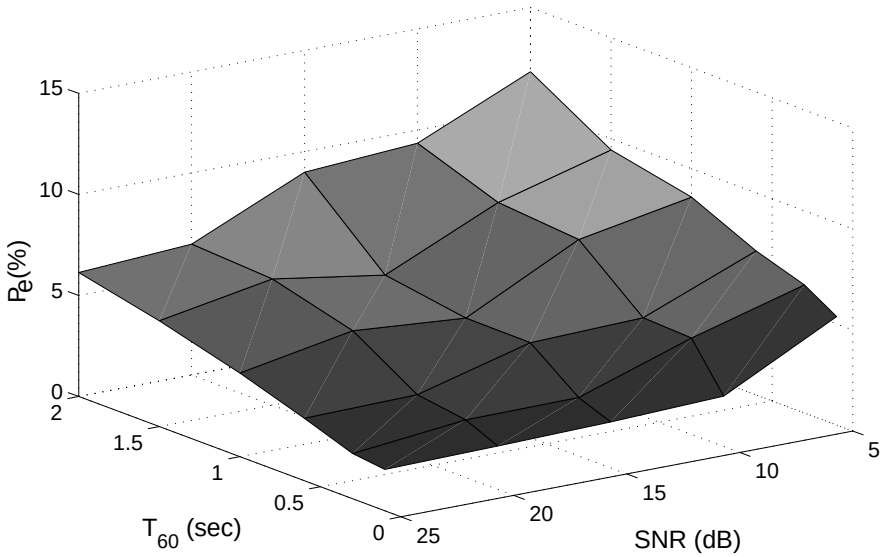

Figure 3.10: Performance of the system under simultaneous phenomena with fixed thresholding

Substituting the adaptive threshold with a fixed one for every noise condition yields Fig. 3.10, which demonstrates that an increase in T_{60} or a decrease in SNR increases P_e and that the non-monotonic behaviour observed in Fig. 3.9 is due to the presence of the adaptive threshold, which often caters to the effects of reverberation and noise on the speech signal. The increase of the P_e, for the case of fixed thresholding, is driven by the increase of P_f as can be observed from Table II. P_c on the contrary remains almost unaffected by reverberation even for low SNR values.

3.8 Conclusions

In this Chapter a statistical voice activity detector, which relies on the modelling of the distribution of speech as a linear combination of a Gaussian, a Laplacian, and a two-sided Gamma distribution, has been presented. The decision criterion of the proposed algorithm is the weighted sum of three likelihood ratio tests, each one corresponding to one of the fundamental core distributions. The computation of the corresponding weights has been based on the statistical distances of the instantaneous input samples from the Gaussian, the Laplacian, and the two-sided Gamma distribution, estimated using the Kolmogorov-Smirnov test. To further enhance the performance of the proposed voice activity detector, an adaptive threshold and a hangover scheme have been introduced. Experiments that were performed using artificially reverberated and contaminated with additive noise anechoic audio data have shown that the voice activity detector outperforms the existing systems in terms of error rate and that it produces reliable results even under adverse noise conditions and reverberation effects. In the next Chapter, the concept of using the likelihood ratio test to detect voice activity will be expanded to support input from multiple sources so that spatial information is consider towards optimum performance.

Chapter 4

Multiple Microphone Voice Activity Detection

4.1 Introduction

As we've shown in Chapter 3, the performance of VAD systems depends strongly on various factors, including the discriminative ability of the classification criterion employed, the dynamics of the additive noise and the signal to noise ratio. Speech signals transmitted within reverberant enclosures and captured using far-field microphones are subject to superposition of reflected versions of the source signal. Additionally, the movement of the talking person is also affecting the characteristics of the captured audio signal.

Towards overcoming such adversities, part of related research focused on microphone arrays (35, 36). These VAD systems have the advantage of utilizing spatial information while involving multiple and independent observations in contrary to single microphone based methods, which can only utilize time and/or frequency information. Nevertheless, most of microphone array based VAD require precise estimates of the direction-of-arrival (DOA) of speech signals in advance or assume that the speaker's movement is limited (35, 36). DOA estimation can seriously affect when audio signals are captured within reverberant enclosures or by directional noise sources.

An alternative approach proposed by Ramirez et al. (37) is Multiple Observation likelihood ratio test (MO-LRT) VAD. In MO-LRT, the decision rule that is based on the likelihood ratio of the Gaussian modelled conditioned speech absence and presence, is formulated over a sliding window consisting of a set of observation vectors around the frame for which the decision is being made. Nevertheless, this fact imposes a significant delay to the algorithm and increased computations. For several applications, including real-time operating telecommunication systems, this can be a major disadvantage.

In this Chapter a modification of the MO-LRT towards the development of a multiple sensor VAD will be presented. The proposed scheme takes advantage of the additional information pro-

vided by microphone arrays. It operates without the need of DOA estimation or additional delay compared to previous multi-microphone VAD technologies.

4.2 System Description

Assuming that speech is generated by a single speaker (source), the reverberated speech signals captured by the distant microphone array, bearing M microphones, at time t are given by

$$x_m(\tau) = h_m(\tau) * s(t) + n_m(\tau) \tag{4.1}$$

where x_m denotes the signal captured by the m^{th} microphone $s(\tau)$ the source speech signal at time τ, $h_m(\tau)$ the corresponding acoustic impulse response, $n_m(\tau)$ the additive noise, and $*$ denotes convolution.

4.2.1 Single Microphone Binary Hypothesis Testing

Voice activity detection can be expressed as the likelihood ratio of two hypotheses stating speech presence and absence (17). Assuming additive noise the two hypotheses H_1 and H_0 that indicate speech presence and speech absence are accordingly:

$$H_0 : \text{speech absence} : \mathbf{X}(t) = \mathbf{N}(t) \tag{4.2}$$

$$H_1 : \text{speech presence} : \mathbf{X}(t) = \mathbf{S}(t) + \mathbf{N}(t) \tag{4.3}$$

where $\mathbf{X}(t) = [X_0(t), X_1(t), ..., X_{K-1}(t)]^T$, $\mathbf{S}(t) = \big[S_0(t),\ S_1(t), ..., S_{K-1}(t)\big]^T$, $\mathbf{N}(t) = [N_0(t), N_1(t)$ $, ..., N_{K-1}(t)]^T$ are the noisy captured speech, reverberated speech, and noise frequency components.

Real and imaginary parts of noise and speech frequency spectrum are assumed to be zero mean Gaussian distributed. The probability densities for the noise and speech components with k denoting the frequency bin are given by

$$f_n^G(\mathbf{N}_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{\mathbf{N}_k(t)^2}{2\sigma_{n,k}^2}} \tag{4.4}$$

$$f_s^G(\mathbf{S}_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{s,k}^2}} e^{-\frac{\mathbf{S}_k(t)^2}{2\sigma_{s,k}^2}} \tag{4.5}$$

where $\sigma_{n,k}^2$, $\sigma_{s,k}^2$ the slowly varying variances of the Gaussian distributed noise and speech respectively estimated by employing eqn.(3.39) for the k^{th} frequency component. The probability density functions conditioned on H_0 and H_1 are given by

$$p(X|H_0) = \prod_{k=0}^{K-1} \frac{1}{\pi\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\} \tag{4.6}$$

$$p(X|H_1) = \prod_{k=0}^{K-1} \frac{1}{\pi\left[\lambda_{n,k}+\lambda_{s,k}\right]} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}+\lambda_{s,k}}\right\} \tag{4.7}$$

where $\lambda_{n,k}$ and $\lambda_{s,k}$ denote the variances of N_k, S_k respectively.

4.2.2 Single Microphone LRT (SM-LRT)

In the case of single microphone VAD scheme the likelihood ratio for the kth frequency bin is defined as

$$\Lambda_k \equiv \frac{p(X|H_1)}{p(X|H_0)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \tag{4.8}$$

where $\xi_k \equiv \lambda_{s,k}/\lambda_{n,k}$ and $\gamma_k \equiv |X_k|^2/\lambda_{n,k}$ the *a priori* and *a posteriori* signal to noise ratios (29) estimated by employing the *Predicted Estimation (PD)* (14) method used in Section 3.5.

The decision criteria is based on evaluating the geometric mean of the likelihood ratios for the individual frequencies and is given by

$$\log\Lambda = \frac{1}{K}\sum_{k=0}^{K-1}\log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.9}$$

and elaborating on eqn.(4.8)

$$\log\Lambda = \frac{1}{K}\sum_{k=0}^{K-1}\left\{\gamma_k - \log\left(\gamma_k\right) - 1\right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.10}$$

where η denotes the threshold of decision. If the geometric mean is above the value of the threshold η then speech presence H_1 is indicated, whereas for values below η speech absence is indicated.

4.2.3 Multiple Observation LRT (MO-LRT)

Using multiple observations to enhance the likelihood of a VAD system has shown good properties in previous studies (37). In the MO-LRT system the decision rule is formulated over a sliding window consisting of $2D+1$ observation vectors around the frame for which the decision is being made. The likelihood ratio for MO-LRT is given by (37)

$$\log\Lambda^{MO} = \frac{1}{K(2D+1)}\sum_{d=1}^{2D+1}\sum_{k=0}^{K-1}\left\{\gamma_{k,d} - \log\left(\gamma_{k,d}\right) - 1\right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.11}$$

Nevertheless, this approach imposes a D-frame delay to the algorithm and $2D+1$ times increased computations that, for several applications, like real-time operating telecommunication systems can be a major disadvantage.

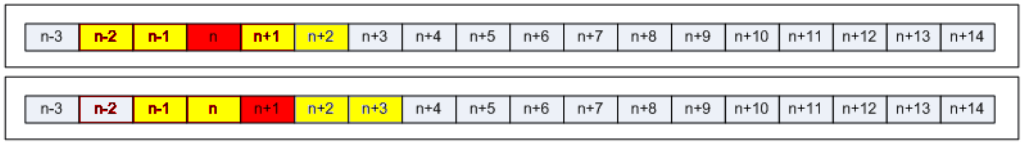

Figure 4.1: Schematic representation of the frame delay imposed by the MO-LRT

4.2.4 Multiple Microphone LRT (MM-LRT)

Nowadays, microphone arrays (Fig. 4.2) have become a commodity in commercial (mobile phones, VOIP terminals) and research environments such as smart rooms (38). The modification of MO-LRT depicted here, takes advantage of the additional information, required to enhance the decision of an LRT based VAD, that can be retrieved by the available microphones rather than using past information through sliding windows that increase the overall delay. This modification on MO-LRT to a Multiple Microphone LRT (MM-LRT) based VAD relies on the following ratio test

$$\log \Lambda^{MM} = \frac{1}{KM} \times \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \{\gamma_{k,m} - \log(\gamma_{k,m}) - 1\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.12}$$

where M denotes the number of available microphones.

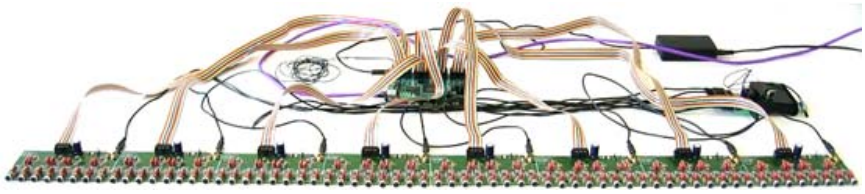

Figure 4.2: Typical hardware used in linear microphone arrays such as NIST Mark-III Microphone array.

4.2.5 Combining MO-LRT and MM-LRT

For the cases that the VAD system doesn't need to operate real-time we propose the combination of MM-LRT eqn.(4.12) and MO-LRT eqn.(4.11) to a Multiple Microphone Multiple Observation LRT (MM-MO-LRT). This combination can potentially enhance even further the performance of such systems, given that the conditions of operation allow for the increased delay. The combined likelihood ratio is given by

$$\log \Lambda^{MM-MO} = \frac{1}{KM(2D+1)} \sum_{m=0}^{M-1} \sum_{d=1}^{2D+1} \sum_{k=0}^{K-1} \{\gamma_{m,k,d} - \log(\gamma_{m,k,d}) - 1\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.13}$$

where M and D indicate the number of the employed microphones and the introduced delay respectively.

4.2.6 Decision Smoothing

In order to enhance the performance of the hypothesis tests the following forgetting scheme is employed

$$\Phi(t) = (1 - \lambda_\Lambda)\Phi(t-1) + \lambda_\Lambda \log \Lambda(t) \tag{4.14}$$

where λ_Λ a smothing factor and $\Phi(t)$ the smoothed likelihood.

4.3 Performance Discussion

The P_c and P_f were evaluated using the speech recordings performed in the anechoic chamber of Aalborg University Denmark using a close talking microphone (*Section II*). Speech data were, as before, contaminated artificially with white and vehicular noises from NOISEX-92 database (33). The microphone array data were artificially generated using the *Image Method* (27) for a reverberation time of $T_{60} = 0.15$sec and room dimensions $[4.4,\ 5.8,\ 2.6]m$. The speaker was $2.5m$ away from the linear array. The input data were sampled at 8 kHz and were segmented into overlapping frames of 40 msec duration (10 msec step size).

The performance was evaluated under several scenarios and has been compared to SM-LRT, MO-LRT, the proposed MM-MO-LRT combination and to the standard ITU-T G.729 Annex B VAD. For a fair evaluation of the systems the same $\lambda_\Lambda = 0.04$ and frame/step sizes have been used. By examining the detection performance under a variety of noisy conditions, a set of thresholds η for each scheme and noise scenario has been heuristically defined. The simulation results are depicted in Table 4.1.

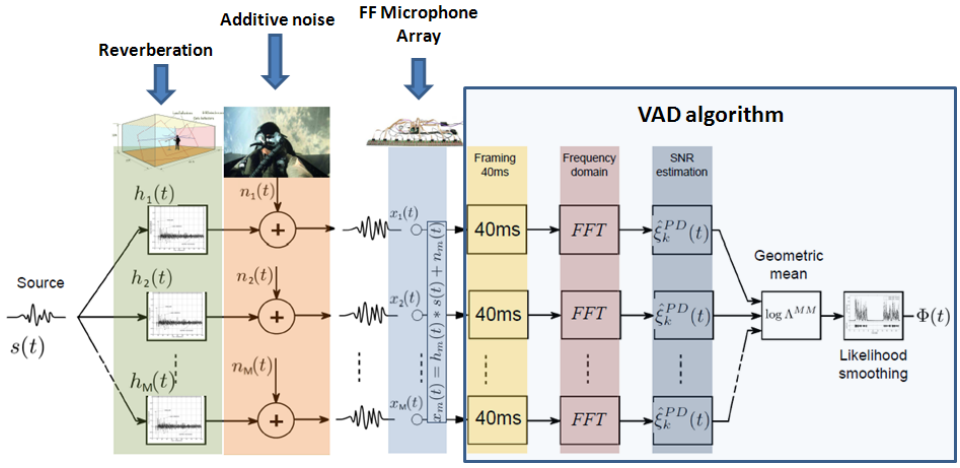

Figure 4.3: System architecture for the Multi Microphone LRT.

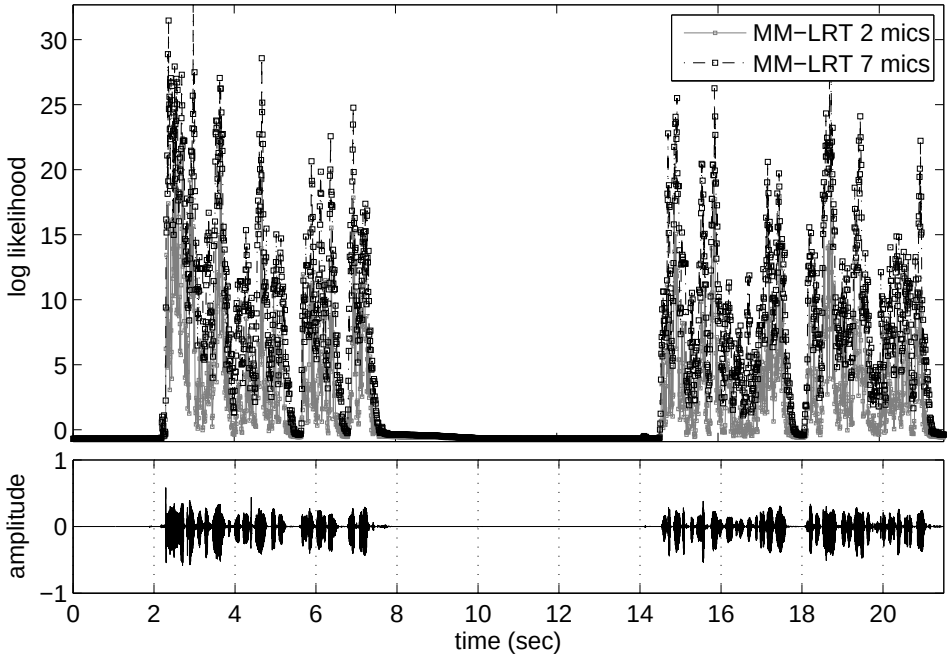

Figure 4.4: Likelihood ratio difference when using 2 or 7 microphones in the estimation of MM-LRT at 10dB of vehicular noise.

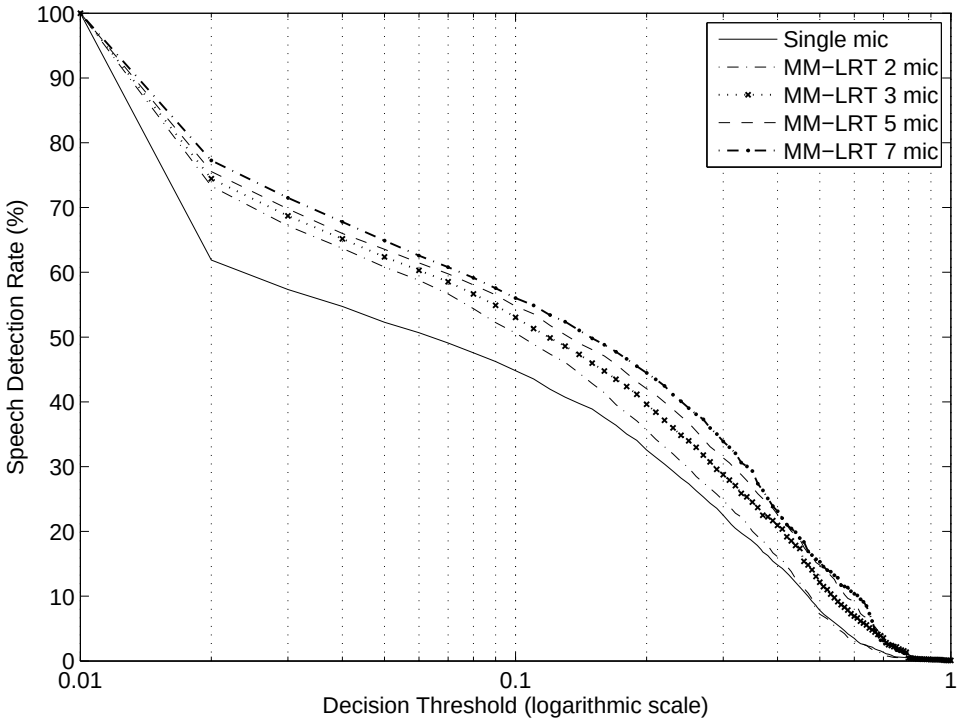

Figure 4.5: Performance enhancement as a function of the number of employed microphones.

Figure 4.4 depicts the difference in likelihood ratio when employing 2 and 7 microphones in eqn.(4.12) at 10dB of vehicular noise. In the latter case the likelihood ratio is significantly enhanced. The LRT value of short silence intervals within words at intervals of speech has been increased. This results in a system the likelihood ratio of which is more uniform within speech segments assisting the overall behaviour of the system towards clipping error reduction.

This type of performance enhancement can be evaluated as a function of speech detection rate versus the normalized value of threshold that is employed every time. To do this the range of values for the likelihood ratio are normalized to 1.

Figure 4.5 depicts the performance gain when increasing the number of microphones in eqn.(4.12) for 10dB of vehicular noise. As shown, the system's performance is significantly enhanced by just introducing a second microphone. Additional microphones also have a positive effect to the response of the system.

Figure 4.6 illustrates the performance of the previously discussed VAD systems under 5dB of vehicular noise. MM-LRT and MO-LRT are compared under the same computational complexity in terms of iterations performed to evaluate eqn.(4.11) and eqn.(4.12) respectively. For a fair

Table 4.1: Performance Results under Various Types of Noise

		SM-LRT			MM-LRT (m=7)			MO-LRT (d=3)			MM-MO-LRT (m=7, d=1)			G.729
Noise	**SNR**	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%	P_e%
AWGN	**20dB**	7.85	6.46	9.24	7.54	7.50	7.57	7.04	7.13	6.94	7.29	6.96	7.62	14.63
	15dB	13.89	19.97	7.81	9.85	10.17	9.53	12.40	16.87	7.93	9.53	8.81	10.25	21.01
	10dB	16.50	19.80	13.20	9.72	9.83	9.60	12.74	16.95	8.53	10.02	9.69	10.35	23.71
	5dB	19.64	18.58	20.71	9.50	8.47	10.54	14.81	17.32	12.13	9.88	10.41	9.35	33.16
Vehicle	**20dB**	6.56	6.26	6.85	5.52	5.37	5.67	5.69	5.93	5.46	5.12	4.63	5.62	15.67
	15dB	8.55	9.45	7.65	6.80	6.49	7.11	8.04	5.45	12.63	6.41	5.14	7.67	17.84
	10dB	14.74	11.72	17.77	11.78	10.72	12.85	13.63	8.23	19.03	11.05	9.77	12.33	19.55
	5dB	18.53	19.57	17.49	16.44	13.71	19.17	17.87	13.15	22.59	15.29	15.96	14.62	21.22

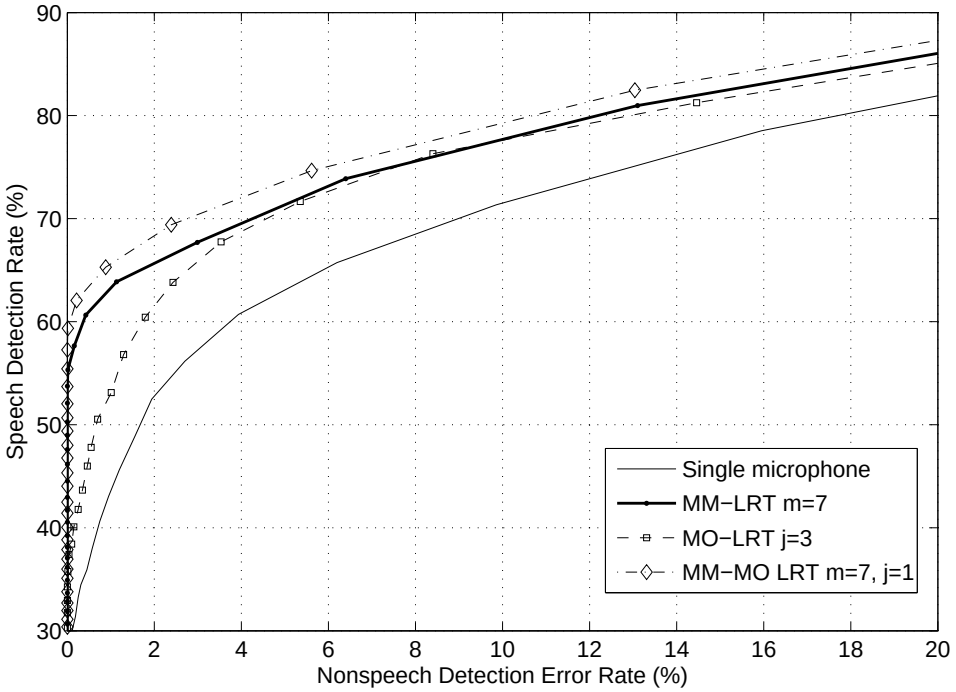

Figure 4.6: Speech Detection Rate vs Non-Speech Error Rate at 5dB SNR of vehicular noise.

evaluation the two systems where set to operate under the same number of iterations thus, $M = 2D + 1$. A number of $M = 7$ microphones has been selected for this case that results in increasing the delay by $D = 3$ frames. The results show that MM-LRT performs better or equal to MO-LRT with the same complexity and significantly lower delay. The scenario of MM-MO-LRT has been also evaluated for the least frame delay increment $D = 1$ showing that it outperforms the rest systems with the cost of additional delay.

The performance of the proposed systems has been also evaluated under white noise as illustrated in Table 4.1, showing similar properties to vehicular noise operation.

4.4 Conclusions

A statistical VAD, which relies on a multiple microphone likelihood ratio test has been demonstrated in this Chapter. The system is based on processing signals captured by far-field microphone arrays. This way the proposed scheme is taking advantage of the spatial information provided by multiple sensors without assuming knowledge of direction-of-arrival estimates. In scenarios that are not real-time critical the system can be further extended to include additional observations employing

a sliding window around the currently processed frame. Through simulations we have demonstrated that the proposed system remains more robust than a set of related counterparts. In Chapter 5, the proposed architecture will be used to merge information derived by a signal decomposition framework, exploiting further the possibilities this VAD scheme.

Chapter 5

Empirical Mode Decomposition based Voice Activity Detection

5.1 Introduction

Data analysis is a necessary part of research and practical applications. Nevertheless, it is very common that data from a process will be likely characterised by several limiting factors. The total data span will be finite, the data will be non-stationary and the data might emerge from a non-linear processes. These characteristics might appear individually in a data set or combined. Furthermore, intervals of data shorter in duration than the total time scale of a stationary process can be non-stationary. Speech generation process is such a non-linear and non-stationary process and especially during fast transitions between phonemes and voicing states it can be considered as highly non-stationary. Thus, its analysis, when employing methods such as Fourier is conducted under specific assumptions of linearity and stationarity.

Nevertheless, Fourier analysis is the dominant one, and has been applied to all kinds of data since it was presented. Although the Fourier transform is valid under general conditions the system must be linear and the data must be periodic or stationary. Otherwise, the resulting spectrum will make little physical sense. The stationarity requirement is not particular to the Fourier spectral analysis, as it is a general case for most of the available data analysis methods.

Apart from stationarity, Fourier spectral analysis also assumes data linearity. Although many processes can be approximated by linear systems, they also have the tendency to be non-linear whenever their variations become finite in amplitude and time. Despite the limitations imposed by Fourier analysis, due to lack of alternatives, Fourier spectral analysis is widely used to process such data and speech. Although the rough adoption of the stationary and linear assumptions may lead in performance reduction of systems based on Fourier analysis.

In this Chapter the system developed as a multi-microphone VAD serves as the platform to merge VAD with a very powerful analysis framework the Empirical Mode Decomposition (EMD)(39).

This highly efficient method relies on local characteristics of time scale of the data to analyse and decompose non-stationary signals into a set of so called intrinsic mode functions (IMF). These functions are injected to the multiple microphone VAD scheme in order to decide upon speech presence or absence.

VAD systems based on EMD analysis have been also presented in the past in (40) where EMD has been used to decompose speech signals self-adaptively and locally. Based on extracting entropy-based features from the resulting IMFs, the experiments have shown that the proposed method was superior to the entropy extracted from original speech especially under intensive background noise. In (41) the signal was first decomposed employing EMD, and then partial decomposition results were processed by Hilbert transform (HT) to obtain the instantaneous frequency. The threshold of noise was estimated by analysing the front of signal's Hilbert amplitude spectrum. The speech segments and non-speech segments were distinguished by the threshold and the whole signal's Hilbert amplitude spectrum resulting in very good performance under low SNR conditions.

5.2 Empirical Mode Decomposition

The key part of the EMD method is that any complicated data set can be decomposed into a finite and often small number of 'intrinsic mode functions' that admit well-behaved Hilbert transforms. This decomposition method is adaptive, and, therefore, highly efficient. Since the decomposition is based on the local characteristic time scale of the data, it is applicable to non-linear and non-stationary processes (39). An IMF must satisfy the following two conditions:

1. In the whole signal set, the number of extrema and the number of zeros crossings must either equal or differ at most by one.

2. At any point, the mean value of the envelope defined by the maxima and the envelope by the minima is zero, that is that the upper envelope and the lower envelope of the signal symmetry in the time axis.

Assuming a single speaker, the speech signal captured by a distant microphone at time τ is given by

$$x(\tau) = h(\tau) * s(\tau) + n(\tau) \tag{5.1}$$

where $s(\tau)$ denotes the source speech signal at time τ, $h(\tau)$ the corresponding acoustic impulse response, $n(\tau)$ the additive noise, and $*$ denotes convolution.

The process of EMD decomposition for $x(\tau)$ is the following:

- Identify all the maxima of the whole signal $x(\tau)$ and then by using cubic spline curve interpolate maxima points to define the upper envelope of the signal.
- Repeat the above methods, find all the minima, fitting the lower envelope.
- The mean value function of the upper and lower envelope is defined as $m_1(\tau)$, and then the first signal component can be calculated as

$$x(\tau) - m_1(\tau) = h_1(\tau). \tag{5.2}$$

- Ideally, $h_1(\tau)$ should be an IMF, however, in reality, it is difficult to obtain the theoretical upper and lower envelope, but only obtain by cubic spline fitting an approximation. The sifting process (39) employed can extract the essential scales from the data. The sifting process serves the purpose to eliminate riding waves and to make the wave-profiles more symmetric. To achieve that the sifting process has to be repeated more times. In the second sifting process, h_1 is treated as the data and then we get

$$h_1(\tau) - m_{11}(\tau) = h_{11}(\tau). \tag{5.3}$$

- The sifting procedure is repeated d times, until h_{1d} is an IMF, that is

$$h_{1(d-1)}(\tau) - m_{1d}(\tau) = h_{1d}(\tau), \tag{5.4}$$

- Then, it is designated as $c_1(\tau) = h_{1d}(\tau)$,the first IMF component from the data. To guarantee that the IMF components retain enough physical sense of both amplitude and frequency modulations, a criterion for the sifting process to stop is determined. This is done by limiting the size of the standard deviation, SD, computed from the two consecutive sifting results as

$$SD = \sum_{\tau=0}^{T} \left[\frac{\left| h_{1(d-1)(\tau)} - h_{1d}(\tau) \right|^2}{h_{1(d-1)}^2(\tau)} \right] \tag{5.5}$$

  A typical value for SD can be set between 0.2 and 0.3. Overall, c_1 should contain the finest scale or the shortest period component of the signal.

- The separation of c_1 from the rest of the data is given simply by

$$x(\tau) - c_1(\tau) = r_1(\tau). \tag{5.6}$$

- Since the residue, r_1 , still contains information of longer period components, it is treated as the new data and subjected to the same sifting process as described above.

$$r_1(\tau) - c_2(\tau) = r_2(\tau), \cdots, r_{i-1}(\tau) - c_i(\tau) = r_I(\tau). \tag{5.7}$$

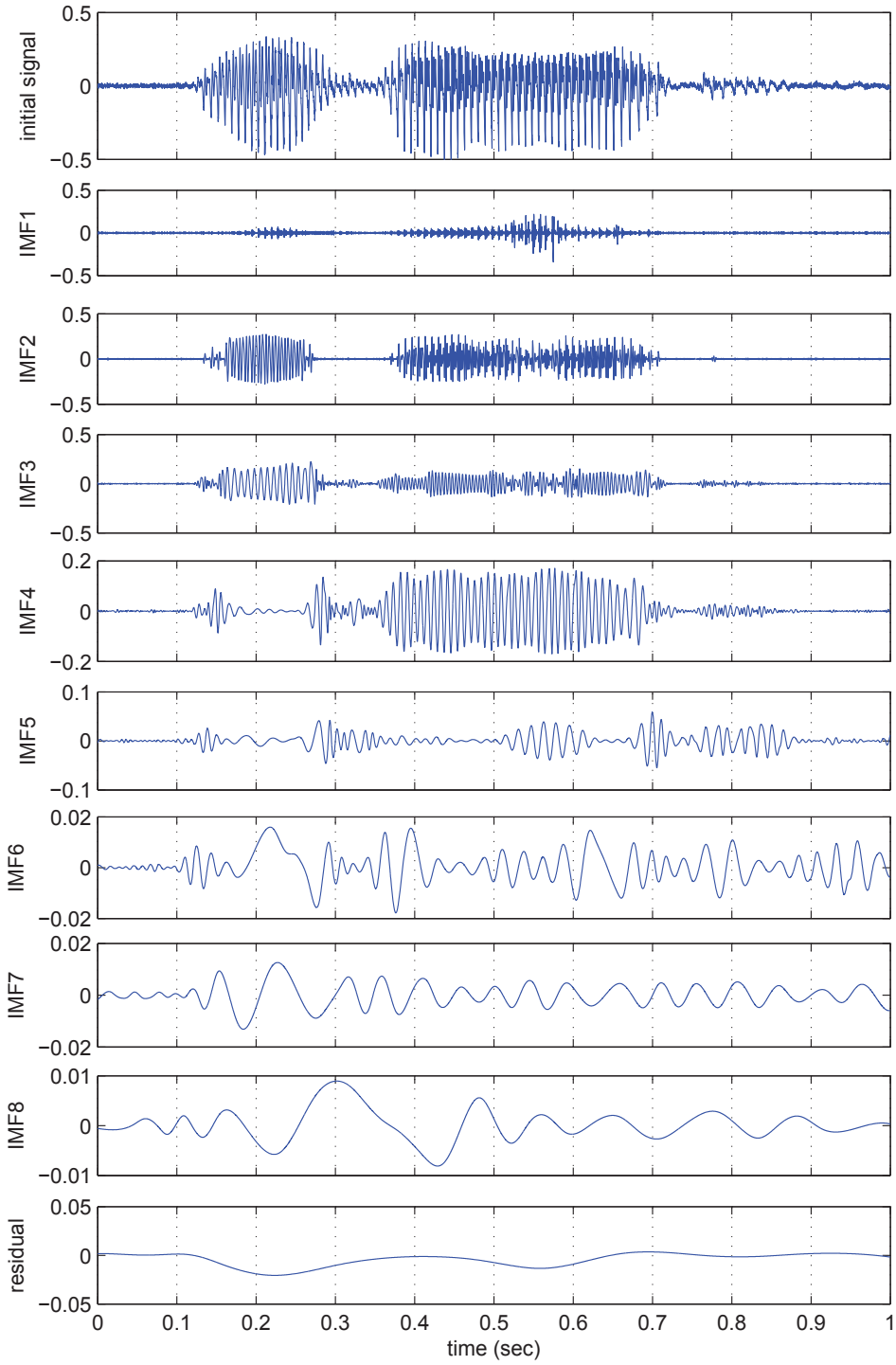

Figure 5.1: Intrinsic Mode Functions that emerged from the Empirical Mode Decomposition of a speech segment.

The sifting process can be stopped by either when the component, c_i , or the residue, r_I , becomes so small that it is less than the predetermined value of substantial consequence, or when the residue, r_I ,becomes a monotonic function from which no more IMFs can be extracted. Finally, the decomposed data into n-empirical modes, and a residue, r_I , which can be either the mean trend or a constant can be recomposed to the initial signal by

$$x(\tau) = \sum_{i=1}^{I} c_i(\tau) + r_I(\tau). \tag{5.8}$$

where I the number of total IMFs that emerged from the decomposition process. In Fig. 5.1 a decomposition example of a speech segment contaminated by Gaussian noise at 20dB is illustrated. During the decomposition of EMD, IMFs with the minimal scale are obtained first (high frequency) and then are IMFs with large scales (low frequency). In the end the IMF with the maximal scale is derived (the trend). As depicted in Fig. 5.1 the larger scales (low frequency IMF) have very low amplitude compared to the small ones. Thus, information contained in the specific IMFs is of lower significance compared to the information contained in the first IMFs.

5.3 Merging EMD with Multiple Microphone VAD

The multiple microphone VAD system developed in 4 serves as the platform to merge VAD and EMD. Although for the scope of this work, the multiple microphone signals are substituted by the corresponding I IMFs (5.8). In essence, the signal $x(\tau)$ is first decomposed with EMD into a set of IMF signals $c_i(\tau)$ that are treated as additional recordings of a microphone array. The trend r_I is not included in the process.

Following 4, VAD is expressed as the likelihood ratio of two hypotheses stating speech presence and absence for each IMF $c_i(t)$. Assuming additive noise the two hypotheses $H_{1,i}$ and $H_{0,i}$ that indicate speech presence and speech absence are accordingly:

$$H_{0,i} : \text{speech absence} : \mathbf{X}_i(t) = \mathbf{N}_i(t) \tag{5.9}$$

$$H_{1,i} : \text{speech presence} : \mathbf{X}_i(t) = \mathbf{S}_i(t) + \mathbf{N}_i(t) \tag{5.10}$$

where $\mathbf{X}_i(t) = [X_{0,i}(t), X_{1,i}(t), ..., X_{K-1,i}(t)]^T$, $\mathbf{S}_i(t) = \left[S_{0,i}(t), S_{1,i}(t), ..., S_{K-1,i}(t)\right]^T$, $\mathbf{N}_i(t) = \left[N_{0,i}(t), N_{1,i}(t), ..., N_{K-1,i}(t)\right]^T$ are the noisy captured speech, reverberated speech, and noise frequency components for the i^{th} IMF $c_i(t)$ with K the total number of frequency bins.

Real and imaginary parts of noise and speech frequency spectrum are assumed to be zero mean Gaussian distributed for every IMF. The probability densities for the noise and speech components

with k denoting the frequency bin are given by

$$f_{n,i}(\mathbf{N}_{k,i}(t)) = \left(2\pi\sigma^2_{n,k,i}\right)^{-\frac{1}{2}} e^{-\frac{\mathbf{N}_{k,i}(t)^2}{2\sigma^2_{n,k,i}}} \tag{5.11}$$

$$f_{s,i}(\mathbf{S}_{k,i}(t)) = \left(2\pi\sigma^2_{s,k,i}\right)^{-\frac{1}{2}} e^{-\frac{\mathbf{S}_{k,i}(t)^2}{2\sigma^2_{s,k,i}}} \tag{5.12}$$

where $\lambda_{n,k,i} = \sigma^2_{n,k,i}$, $\lambda_{s,k,i} = \sigma^2_{s,k,i}$ the slowly varying variances of the Gaussian distributed noise and speech respectively estimated by employing eqn.(3.39) for the k^{th} frequency component of the i^{th} IMF. The probability density functions conditioned on $H_{0,i}$ and $H_{1,i}$ are given by

$$p(X_{k,i}|H_{0,i}) = \prod_{k=0}^{K-1} \frac{1}{\pi\lambda_{n,k,i}} e\left\{-\frac{|X_{k,i}|^2}{\lambda_{n,k,i}}\right\} \tag{5.13}$$

$$p(X_{k,i}|H_{1,i}) = \prod_{k=0}^{K-1} \frac{1}{\pi\left[\lambda_{n,k,i} + \lambda_{s,k,i}\right]} e\left\{-\frac{|X_{k,i}|^2}{\lambda_{n,k,i} + \lambda_{s,k,i}}\right\}. \tag{5.14}$$

In the case of single microphone VAD scheme the likelihood ratio for the kth frequency bin of the i^{th} IMF is defined as

$$\Lambda_{k,i} \equiv \frac{p(X_{k,i}|H_{1,i})}{p(X_{k,i}|H_{0,i})} = \frac{1}{1+\xi_{k,i}} e\left\{\frac{\gamma_{k,i}\xi_{k,i}}{1+\xi_{k,i}}\right\} \tag{5.15}$$

where $\xi_{k,i} \equiv \lambda_{s,k,i}/\lambda_{n,k,i}$ and $\gamma_{k,i} \equiv |X_{k,i}|^2/\lambda_{n,k,i}$ the *a priori* and *a posteriori* signal to noise ratios estimated by employing the *Predicted Estimation (PD)* method (14)

$$\begin{aligned}\hat{\lambda}_{n,k,i}(t+1) &= \zeta_n\hat{\lambda}_{n,k,i}(t) + (1-\zeta_n)E\left[|N_{k,i}(t)|^2 |X_{k,i}(t)\right] \\ \hat{\lambda}_{s,k,i}(t+1) &= \zeta_s\hat{\lambda}_{s,k,i}(t) + (1-\zeta_s)E\left[|S_{k,i}(t)|^2 |X_{k,i}(t)\right]\end{aligned} \tag{5.16}$$

where $\hat{\lambda}_{s,k,i}(t), \hat{\lambda}_{n,k,i}(t)$ are estimates of $\lambda_{s,k,i}(t), \lambda_{n,k,i}(t)$ and ζ_n ,ζ_s are smoothing parameters both set to 0.99.

The decision is drawn through the geometric mean of the likelihood ratios for the individual frequencies of every IMF

$$\log\Lambda_{k,i} = \frac{1}{K}\sum_{k=0}^{K-1}\left\{\gamma_{k,i} - \log\left(\gamma_{k,i}\right) - 1\right\}. \tag{5.17}$$

Thus, the LRT across all IMFs components will be transformed to

$$\Lambda^{\text{EMD}}_{\log} = \frac{1}{IK} \times \sum_{i=1}^{I}\sum_{k=0}^{K-1}\left\{\gamma_{k,i} - \log\left(\gamma_{k,i}\right) - 1\right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{5.18}$$

where η the threshold of decision estimated by 3.6.

5.3.1 Likelihood Smoothing

The following forgetting scheme is employed to enhance the performance of the hypothesis test

$$\Phi(t) = (1 - \lambda_{\Lambda_{\log}^{\text{EMD}}})\Phi(t-1) + \lambda_{\Lambda_{\log}^{\text{EMD}}} \log \Lambda_{\log}^{\text{EMD}}(t) \tag{5.19}$$

where $\lambda_{\Lambda_{\log}^{\text{EMD}}} = 0.9$ the smoothing factor and $\Phi(t)$ the smoothed likelihood.

5.4 Performance Discussion

The P_c and P_f were evaluated using the speech recordings performed in the anechoic chamber of Aalborg University Denmark using a close talking microphone (Chapter 2). Speech data were, as before, contaminated artificially with white and vehicular noises from NOISEX-92 database (33). The microphone array data were artificially generated using the *Image Method* (27) for a reverberation time of $T_{60} = 0.15$sec and room dimensions $[4.4, 5.8, 2.6]m$. The speaker was $2.5m$ away from the linear array. The input data were sampled at 8 kHz and were segmented into overlapping frames of 40 msec duration (10 msec step size). The total number of IMFs used in the EMD decomposition was fixed to ten and the trend was excluded from the estimation of the LRT with $\lambda_{\Lambda}^{\text{EMD}} = 0.1$.

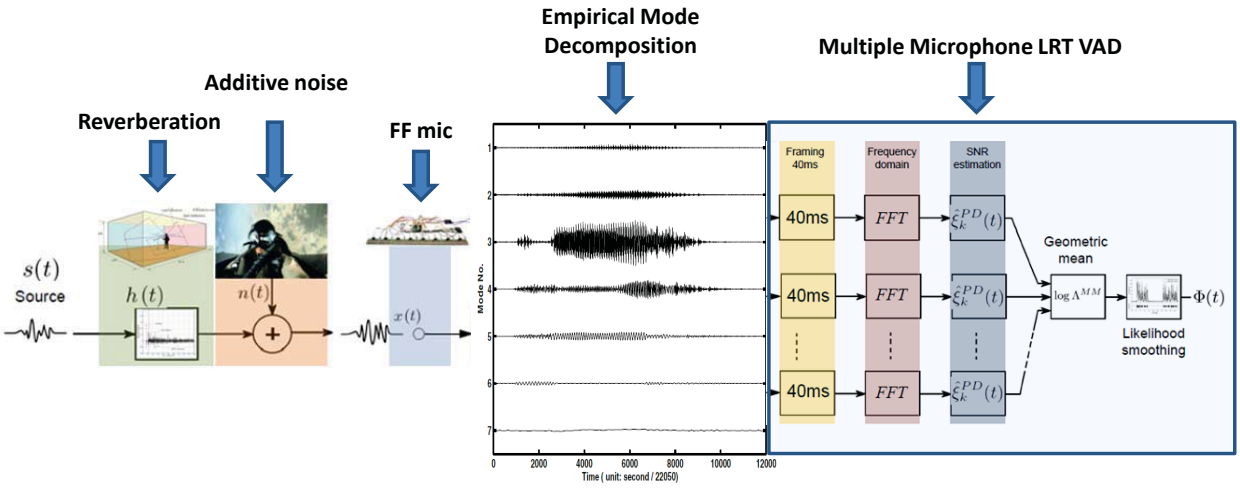

Figure 5.2: System architecture for the EMD based LRT VAD

The performance of the proposed systems was evaluated under several scenarios and has been compared to SM-LRT presented in Section (4.2.2) and to the systems proposed in (40, 41) denoted as 'EMD+HHT' and 'EMD + SpEnt' respectively. Same frame/step sizes have been used for all systems.

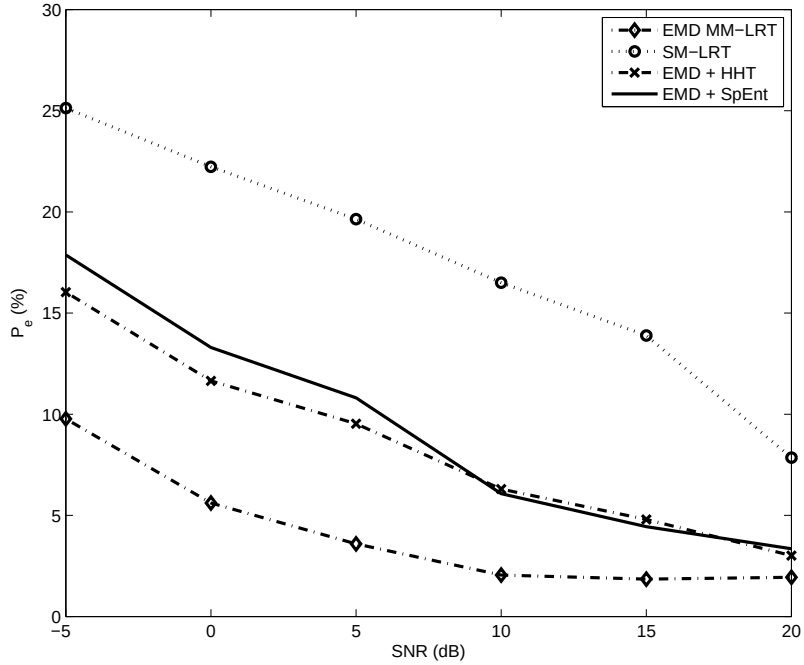

Figure 5.3: P_e performance under different intensities of White noise

Figure 5.3 depicts how the performance of the proposed system varies as SNR drops due to AWGN. In this graph, it is shown that the performance of EMD combined with the Multi-microphone VAD surpasses all other solutions. Additionally, EMD with spectral entropy performs better that EMD with HHT just for up to 10dB. This is due to the fact that the overall performance of the former system is subject to P_f. The single microphone LRT is the worst performer depicting the advantage of performance enhancement with EMD prior likelihood testing.

Babble noise (Fig. 5.4) is considered to be one of the most adverse conditions for VAD. In this case, the conclusions are similar to those for AWGN. The proposed system performs better than the rest in almost all cases. EMD with spectral entropy is slightly better than EMD with HTT system and especially for the case of -5dB it performs better than the proposed solution. The performance for all systems drops significantly faster with SNR dropping, compared to the case of white noise.

In the case of car noise (Fig. 5.5), the performance of the proposed system is again above the rest of the systems. The performance of the EMD + HHT system better than the EMD with spectral entropy for almost all cases apart from the case of 10dB. The simulation results are depicted in detail in Table 5.1.

Table 5.1: Performance Results under Various Types of Noise

		EMD MM-LRT			SM-LRT			EMD + HHT (39)			EMD + SpEnt. (40)		
Noise	**SNR**	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%	P_e%	P_c%	P_f%
AWGN	**20dB**	1.94	2.32	1.65	7.85	6.46	9.24	3.01	3.91	2.12	3.35	1.39	5.31
	15dB	1.85	2.77	0.93	13.89	19.97	7.81	4.80	6.42	3.19	4.44	1.01	7.88
	10dB	2.05	2.09	2.01	16.50	19.80	13.20	6.30	6.23	6.38	6.07	7.95	4.19
	5dB	3.59	4.64	2.55	19.64	18.58	20.71	9.53	10.56	8.51	10.81	9.45	12.18
	0dB	5.61	6.62	4.61	22.23	19.64	24.82	11.65	16.12	7.18	13.30	11.29	15.32
	-5dB	9.77	10.75	8.78	25.13	22.46	27.81	16.03	19.32	12.74	17.87	16.72	19.03
Babble	**20dB**	2.23	1.89	2.57	7.25	6.01	8.49	5.87	5.32	6.42	4.39	2.34	6.45
	15dB	3.00	3.14	2.86	9.75	7.38	12.12	6.43	7.11	5.76	6.77	5.17	8.37
	10dB	5.26	5.52	5.00	13.63	9.94	17.32	8.23	9.28	7.19	10.79	9.78	11.81
	5dB	8.44	7.16	9.73	18.81	14.73	22.89	12.36	12.79	11.93	12.06	10.01	14.12
	0dB	14.16	9.98	18.34	22.74	19.41	26.07	18.09	17.54	18.64	17.78	18.04	17.52
	-5dB	21.18	15.36	26.99	26.85	23.08	30.63	22.09	20.52	23.66	18.44	16.81	20.08
Vehicle	**20dB**	2.18	1.98	2.38	6.56	6.26	6.85	4.24	4.51	3.97	4.66	3.18	6.51
	15dB	1.77	2.25	1.29	8.55	9.45	7.65	4.92	4.03	5.82	5.37	4.31	6.44
	10dB	3.82	3.10	4.54	14.74	11.72	17.77	8.64	6.37	10.91	6.59	5.05	8.13
	5dB	5.18	5.01	5.35	18.53	19.57	17.49	10.10	8.79	11.42	11.45	9.49	13.42
	0dB	9.87	7.77	11.96	22.97	21.33	24.62	11.78	9.44	14.13	13.83	10.99	16.67
	-5dB	11.07	9.81	12.33	24.69	20.05	29.34	14.31	13.18	15.44	15.39	12.34	18.44

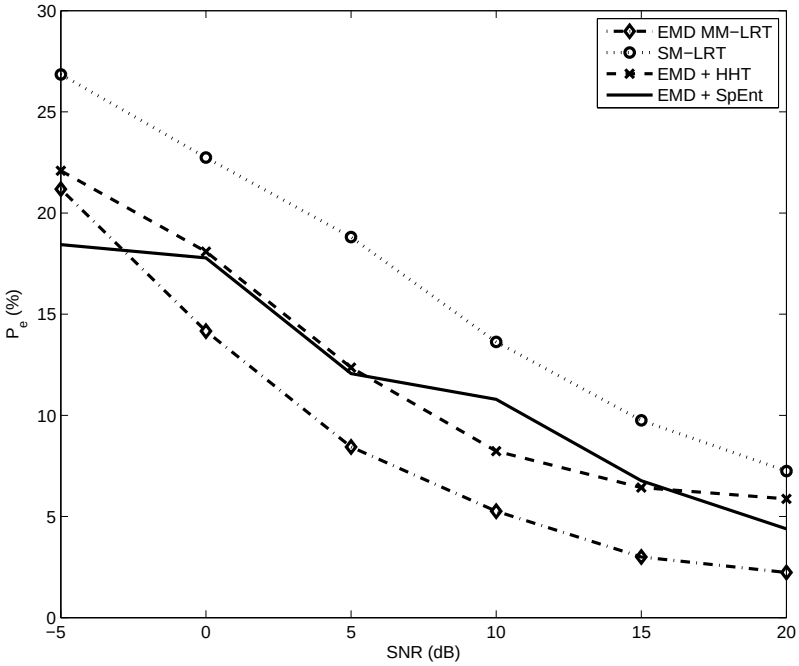

Figure 5.4: P_e performance under different intensities of babble noise

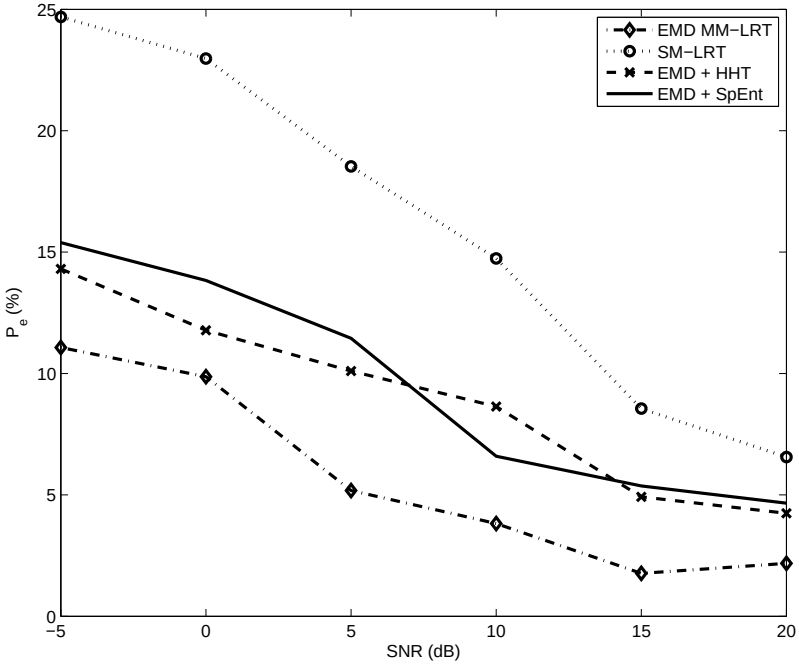

Figure 5.5: P_e performance under different intensities of vehicular noise

5.5 Conclusions

In this Chapter the system developed as a multi-microphone VAD served as the platform to merge VAD with a very powerful analysis framework the Empirical Mode Decomposition (EMD). This highly efficient method relies on local characteristics of time scale of the data to analyse and decompose non-stationary signals into a set of so called intrinsic mode functions (IMF). These functions were injected to the multiple microphone VAD scheme in order to decide upon speech presence or absence. The outcome of this procedure demonstrated significantly enhanced performance compared to single microphone approaches and other competing systems employing EMD. Nevertheless, we think that there is even more space for further improving the performance of the proposed system. Such improvements might emerge by adaptively selecting the number of IMFs at every step and not predefine it heuristically. Additionally, an adaptive way to discard those IMFs that contain very little speech information could possible lead in better performance. Overall, the system proposed here shows the decomposition ability of EMD in terms of significantly improved performance. Nevertheless, the performance comes with a cost. The iterative decomposition procedure followed to produce the IMFs requires significantly increased computational resources. Given that the number of iterations required per IMF are not *a-priori* defined, and is data driven, the whole system might be to slow for real-time implementation although it is designed on a frame-by-frame processing basis as the rest of the systems.

Part III
Supervised Voice Activity Detection

Chapter 6

Hidden Markov Models based VAD

6.1 Introduction

In this Chapter a supervised VAD system that uses far-field microphones is considered. The core of this system consists of two left-right Hidden Markov Models that operate in the feature domain. Taking into consideration the observations emerged from Chapter 2 for the distribution of speech the observation probability distribution function of each state is modelled using Gaussian Mixture Models. This way the distribution of captured speech will be more accurately modelled as a set of Gaussian distributions. An adaptive threshold is derived, that allows for optimum performance even in the case of varying noise statistics. Furthermore, to cater for the inter-frame correlation, especially in the case of speech presence, a hang-over scheme is employed.

Speech generation is a bi-modal process conveying both, audio and visual information an audiovisual VAD that combines the advantages of both modalities is considered. As described in previous Chapters, conventional VAD systems rely solely on audio information (15); their performance is inversely related to the level and the characteristics of the inherent noise. Solely visual VAD systems on the other hand are immune to the interfering noise; however their performance is subject to several issues like the visibility of the lips from the camera angle, poor illumination and low quality of captured images.

A video-VAD is constructed based on the HMM framework of the audio one. The basis of the design is a computationally efficient lip-tracker, operating on the top of a face detection system. Simulations performed for the audio modality are compared to those of the video one. An approach to fuse the two modalities is introduced to illustrate the possibilities of such a combination scheme.

6.2 Audio-VAD System Architecture

The core of the system consists of a pair of Hidden Markov Models (HMM). Each one attempts to model a different audio scene situation; the first is dedicated to the identification of noisy speech

intervals, while the second deals with the speech-free segments. The employed HMMs have the well-documented left-right architecture, in which the index of each state is an increasing, non-monotonical function of time (42). The rationale behind this choice is to accurately model the time evolving properties of speech. Stationary emission processes per model state are modelled using Gaussian Mixtures (GM). Model parameters will be optimised using recorded training data that have been collected from the enclosure within which the system will operate.

The system operates in the feature domain. In particular, 12 Mel-Frequency Cepstral Coefficients-(MFCCs) plus energy are used as input. To effectively reflect the incurring changes of speech signal dynamics and to enhance performance, the first-order derivatives of these features are also taken into account.

During testing mode, the likelihood of the noise model is subtracted from that of the noisy speech model to get a *speech-presence* likelihood. The latter is compared to an adaptive threshold to allow efficient operation even in non-stationary environments. Decisions are smoothed using a hang-over scheme (15) that caters for inter-frame correlation of speech signals. The process is summarized in Fig. 6.1.

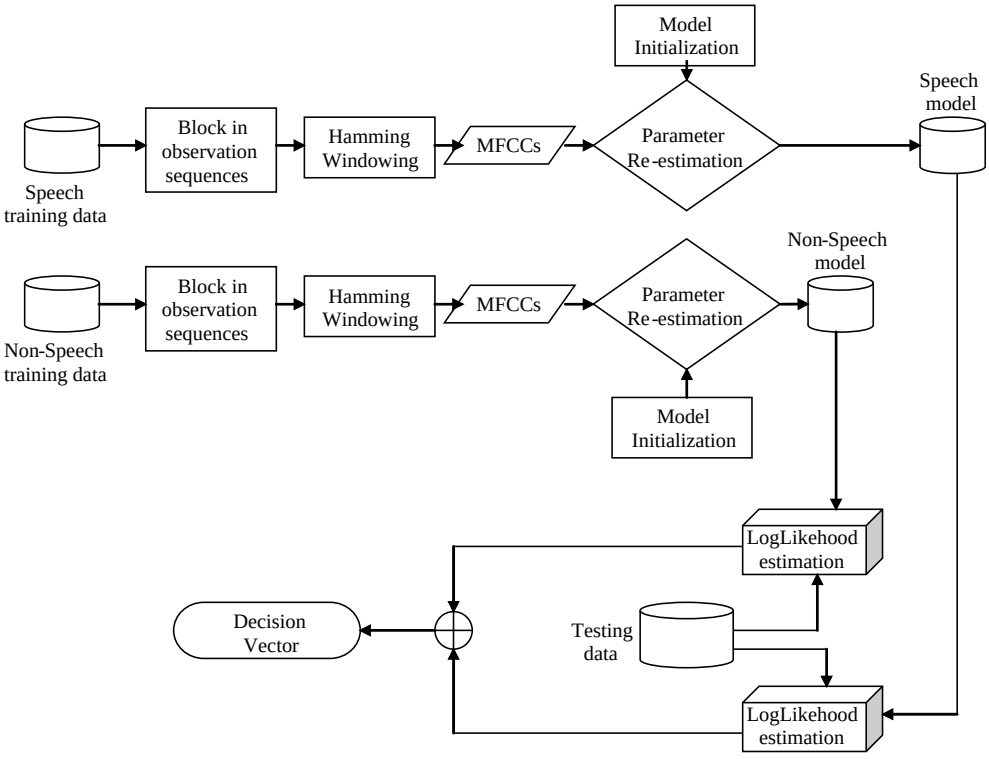

Figure 6.1: Block diagram of the audio HMM-based VAD component.

6.2.1 Model Initialisation

The choice of the initial values of the HMM parameters is crucial for their performance. This is more profound in the continuous distribution case, when employing multiple mixture models (42). Using the *k-means* algorithm for parameter initialisation results in a different HMM parametrisation each time the system is trained, even if the the same training data are used. This, in turn, leads to inconsistent VAD performance. Alternatively a statistical initialization strategy for the Expectation Maximization (EM) algorithm, based on the statistics of the training data can be employed (43). According to this, the pair-wise Euclidean distance between all dimensions of the training dataset is computed and a hierarchical cluster tree is created, based on the Single Linkage (SL) algorithm. The data is partitioned in each dimension into as many subsets as the modes of each HMM state. The *k-means* algorithm (44)is then executed using the central values of each subset as initial values, and the resulting clustered data are fed into EM algorithm for parameter re-estimation (45). Experimental results have shown that this approach results in lower error values than the standard randomly initialised *k-means*.

6.2.2 Training Mode

During training mode the provided *noisy speech* and *noise* observation sequences are split into overlapping frames of duration 45ms and a set of features is extracted for each frame. Subsequently, the parameters of the noisy speech λ_{SN} and the noise model λ_N are optimised, using the EM algorithm. The two models are expected to perform almost identically during speech-absence, provided that a sufficient number of mixtures is used. The use of GMs improves the data-modelling ability of the system and thus, enhances its performance. The optimum order of the GMs per model state is chosen heuristically by assessing performance of the system for various number of states and order of mixtures. To reduce the computational complexity and/or increase the degrees of freedom (number of states, order of mixtures) it is assumed that the elements of the feature vectors are mutually uncorrelated. In practice though this assumption does not hold. Apparently, there is a compromise between optimal performance and computational efficiency.

6.2.3 Classification Mode

To decide on the content of $i-th$ frame, k prior and k posterior frames are considered, forming the observation sequence

$$O^{(i)} = O_{i-k}O_{i-k+1}\cdots O_i \cdots O_{i+k} \tag{6.1}$$

to which the *forward algorithm* is applied (42). The derived conditional probabilities $P(O|\lambda_{SN})$ and $P(O|\lambda_N)$ are then subtracted to obtain an estimate of the probability of speech activity within the $i-th$ frame

$$D^{(i)} = P(O^{(i)}|\lambda_{SN}) - P(O^{(i)}|\lambda_N) \tag{6.2}$$

This estimate is then compared to a decision threshold in order to decide upon the content of the current frame

$$D^{(i)} \gtrless_{silence}^{speech} T \tag{6.3}$$

Notice that the system, when in classification mode, operates in real time; only a slight delay might be introduced depending on the choice of k.

6.2.4 Adaptive Threshold

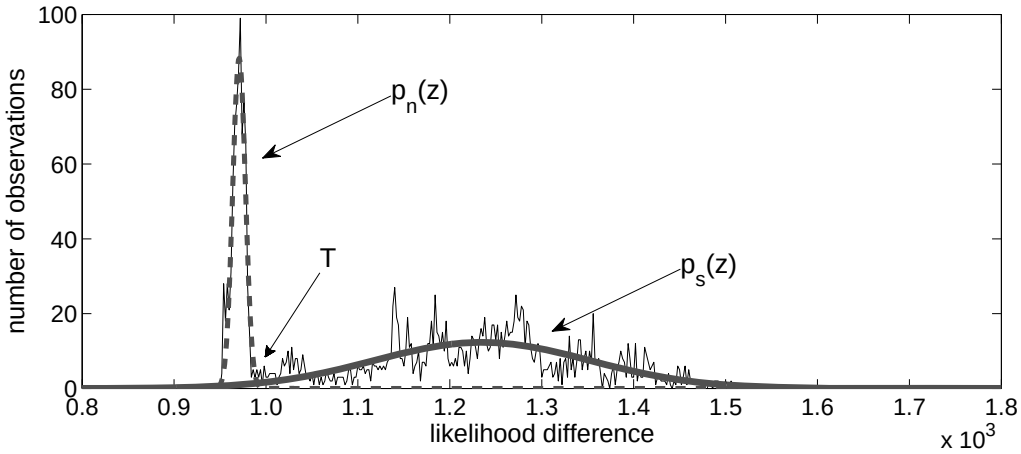

Figure 6.2: Optimum decision threshold, the distributions of speech and noise for the 0 dB case, and the fitted curves.

To deal with the dynamic nature of the signals involved in recordings with FF microphones within enclosures, an adaptive threshold is introduced that aims at the minimisation of the misclassifications (P_f, P_c). The underlying concept behind this threshold is that the derived speech-presence and speech-absence likelihoods follow Gaussian distributions, denoted by $p_n(z;\mu_n,\sigma_n,\alpha_n,i)$ and $p_s(z;\mu_s,\sigma_s,\alpha_s,i)$ respectively, of different mean value (μ_n,μ_s), variance (σ_n,σ_s) and amplitude

(α_n, α_s) for the $i-th$ frame of input data. (Fig. 6.2). Given a decision threshold T, the probability of misclassification will be

$$P_{mis}^{(i)} = \int_{-\infty}^{T} p_s(z)dz + \int_{T}^{\infty} p_n(z)dz \tag{6.4}$$

where the first term of the right hand side stands for the speech classified as noise, while the second is the probability of noise classified as speech (Notice that the dependence of the pdfs on the frame index i, and their parameters has been neglected for simplicity reasons). The optimal threshold for the $i-th$ frame is the argument that minimises the misclassification error.

$$T^{(i)} : \arg\min_{T} P_{mis}^{(i)} \tag{6.5}$$

Taking the gradient of P_{mis} with respect to $T^{(i)}$ yields

$$(\sigma_s^2 - \sigma_n^2){T^{(i)}}^2 - 2(\mu_n\sigma_s^2 - \mu_s\sigma_n^2)T^{(i)} + \left(\mu_n^2\sigma_s^2 - \mu_s^2\sigma_n^2 - 2\sigma_n^2\sigma_s^2 \ln\left(\frac{a_n\sigma_s}{a_s\sigma_n}\right)\right) = 0 \tag{6.6}$$

The solution of the quadratic equation (6.6) returns the value of T for which the misclassification error is minimised.

The threshold $T^{(i)}$ is estimated for every input frame. The noise and speech pdfs are approximated by applying curve fitting techniques to the retrieved histograms. Values of $D^{(i)}$ that are classified as noise are used to update the properties of $p_n(z)$, while those classified as noisy speech are used to update properties of $p_s(z)$. Initially, in the absence of sufficient speech data, $T^{(i)}$ is set to $3\sigma_n$.

6.2.5 Decision Smoothing

To further enhance the performance of the system and to correct decisions that lack a rationale justification, a hang-over scheme is employed. This is a state machine that has 2 major (speech presence, speech absence) and several intermediate transitions states. Thus for transition from speech presence to speech absence state at least 150ms of silence should be detected, while 100ms of speech detections are required for the opposite. This lowers the probability of false rejections, by reducing the risk of a low-energy portion of speech at the end of an utterance being falsely rejected. (15).

6.3 Visual-VAD Architecture

Similarly to Audio-VAD, the structure of a classical dichotomizer, consisting of two HMM models, has been used as the core of the Visual-VAD. The first is used to model lip-movement that occurs

during speech generation while the second identifies lip-movement during speech-absence. The employed features are the vertical lip distance and its first derivative. The log-likelihoods that emerge from the recognition procedure are compared to produce the decisions. These are temporally filtered to reduce false alarms and clipping errors.

Notice that the horizontal opening of the mouth could be also helpful (46); however, this feature can not be robustly extracted in an ambient intelligence set-up where visual features are extracted through automatic processing of facial images captured with FF cameras.

In every frame prior to lip-movement tracking the location of the face is detected using a boosted cascade of simple classifiers structure (47). This frontal detector is trained using 9,000 positive samples (images including faces) and 18,000 negative samples (images without human faces), all of them scaled to 12 pixels width and 16 height (aspect ratio of 3/4). The minimum feature size is 0, hit rate is 99.9% and false alarm per cascade stage 50%. Also horizontal and 45-degrees tilted haar-like features are employed along with non-symmetric faces, four splits and gentle AdaBoost learning (47).

Within each detected face, a Region Of Interest (ROI) is initially defined heuristically based on face geometry; the lips of the speaker are assumed to lie in this area. This ROI is then filtered using a horizontal Sobel gradient operator to emphasize horizontal edges. Laplacian operators are avoided due to their sensitivity to noise (high frequencies) and ringing (double edges) phenomena. Provided that the face is not rotated more than 45 degrees, these edges correspond to the lips. The gradient images are binarised using the adaptive thresholding method proposed by Otsu (48). Morphological opening is finally applied on the binary image to remove noisy pixels. The outcome of each stage of this process is depicted in Fig. 6.3.

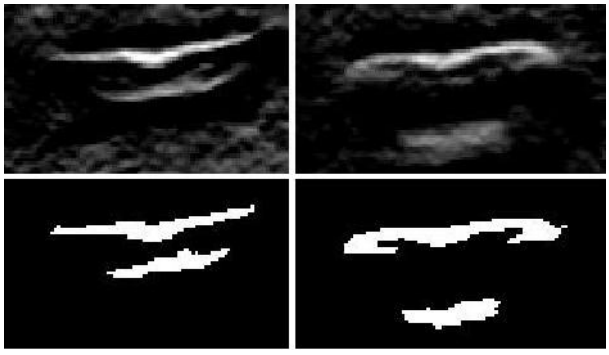

Figure 6.3: Sobel filtering and binarization of the lip ROI.

It is assumed that the upper and lower lips are the two largest components in the lip ROI; the rest are discarded using size thresholding to the second largest object. The vertical opening of lips is calculated as the difference of the mean vertical distance of each lip from the bottom edge of the ROI. By doing so, the effect of the angle of the face in the computation of the lip distance is minimised.

Based on the estimated lip position and the inter-lip distance, a new ROI is defined that is used to locate the lips on the next frame. The exact position and the dimensions of the lip ROIs are temporally smoothed using a recursion. This allows the lip-tracker to cater for the movements of the head of the speaker. A typical run of the lip tracker on the cropped faces is shown in Fig. 6.4.

The video extracted features, namely the vertical mouth opening and its first-order temporal difference, are used to train the speech presence and speech-absence HMMs in the training mode. When the system operates in classification mode the same process is followed; however, features in this case are fed to the Visual-VAD dichotomiser to produce decisions regarding the existence of speech activity.

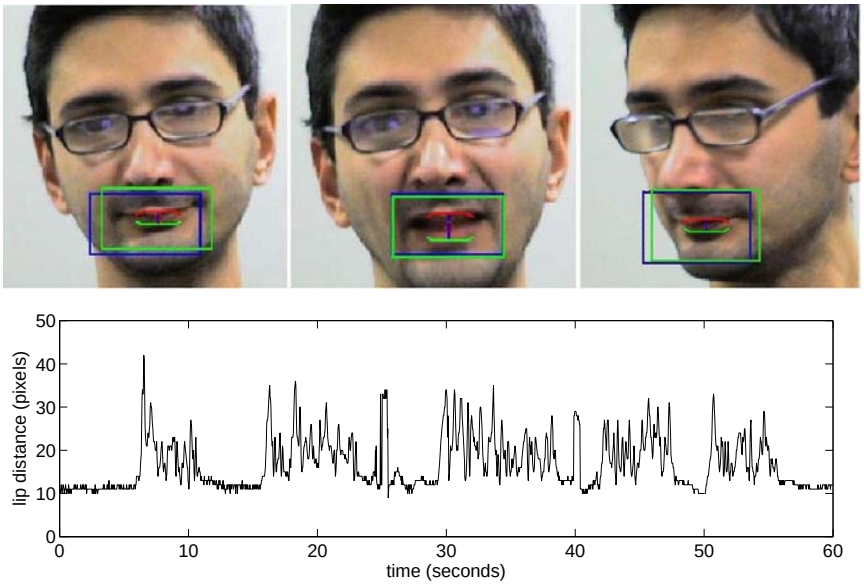

Figure 6.4: Typical example of lip-tracker's outcome. The output of the lip-tracker is rather similar to the energy of a speech-like signal. This is related to the fact that usually, the more we open our mouth, the higher the level of speech gets.

6.3.1 Fusion

In an AudioVisual-VAD system, the objective is to combine the results of the two modalities in order to achieve improved and robust performance even under severe noise conditions. Early

fusion of the audio and visual features in an audio-visual super-vector has been reported to be outperformed by late fusion of the decisions of the individual modalities (49). In addition, in early fusion implementations the following issues occur

- Sensors capture raw data in different rates (22050Hz for microphones and 15-30FPS for camera). Thus the visual features would have to be interpolated in order to be combined with microphone features.
- The resulting super-vector will not be balanced across the two modalities, since the video feature vector has two elements, while the audio vector has 26 (12 MFCCs + Energy + 1st derivative).

To tackle these issues, high-level fusion is employed. The usual approach for decision fusion is to use weights for the two modalities (50). In some cases these weights are allowed to vary, depending on the characteristics of the audio-visual signals (51). However, these approaches require training of the weights and measurement of the chosen audio-visual characteristic, which is not feasible in the current application. Moreover, they do not take into account the occasional absence of lips due to head movement, nor the movement of the lips that might occur in speech absence (coughing, yarning, etc).

In the AudioVisual-VAD a rule-based approach is employed, performing in a hierarchical fashion. More specifically, if a face is detected within the current frame then Visual-VAD becomes the primary component and Audio-VAD is activated only if lip movement is detected. In this case, a speech-absence decision by the Visual-VAD results in a general speech absence decision, while both components should detect speech for a general speech presence decision. On the other hand, if the face detector does not detect a face in the current frame only the Audio-VAD component is taken into account (Table 6.1).

Table 6.1: The Employed Fusion Matrix

Face Detection	Visual-VAD	Audio-VAD	AudioVisual-VAD
1	1	1	1
1	1	0	0
1	0	1	0
1	0	0	0
0	-	1	1
0	-	0	0
0	-	1	1
0	-	0	0

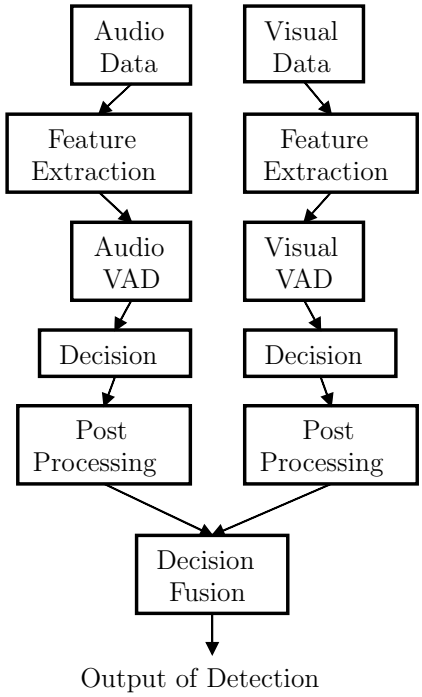

Figure 6.5: Block diagram of fusing the two modalities.

6.4 Experimental Setup

The performance of the system was evaluated using recordings from a typical office room with dimensions 7x4x2.5 meters and reverberation time 150ms. A single omnidirectional microphone placed on a table was used to record the voices of four speakers who were allowed to move freely in space while recording, in order to emulate a non-intrusive, ambient intelligence environment. Subjects were asked to read a specific document, including equally distributed segments of speech and silence. 30% of the recorded data was used for training and 70% for testing purposes. The recorded sequences were manually annotated with precision 1/100 seconds. The SNR of the data was approximately 15dB; however, additive noise components were also artificially introduced, bringing SNR down to -5dB in order to evaluate performance under adverse noise conditions. The sampling rate was 22kHz and the recorded sequences were split into frames of duration 45ms during processing. Adjacent frames were overlapping by 75%. The camera operates at 15 frames per second, with a resolution of 640x480pixels. Heads were less than 250 pixels high in the video sequence.

The left-right HMMs that were used for the modelling of the *noisy speech* and *noise* audio signals had three states each and six Gaussian distributions per state.

The performance of the algorithm is compared to that of the VAD system described in annex B of G.729 audio data compression standard algorithm (4), a very common solution for VoIP applications and cellular telephony.

6.4.1 Performance Results

The performance results of the proposed system, presented here, were evaluated under three different additive white Gaussian noise level scenarios: (a) 15dB, (b) 0dB and (c) -5dB and the results that were obtained are depicted in Table 3.1. For each noise scenario two situations were examined: (i) speaker dependent experiments, where the users used for training are a superset of the users used for testing and (ii) speaker independent experiments, where the speakers used for testing are not included in the training set.

Table 6.2: Performance Results

Noise	**Case**	P_c%	P_f%	P_e%
Audio - VAD				
15dB	SpDEP	0.2580	2.7118	1.4849
	SpIND	0.0864	3.5247	1.8055
0dB	SpDEP	2.6203	2.6437	2.6320
	SpIND	2.5231	2.8635	2.6933
-5dB	SpDEP	9.0798	6.9208	8.0003
	SpIND	10.3073	13.8744	12.0908
Noise	**Case**	P_c%	P_f%	P_e%
Visual - VAD				
	SpIND	6.3864	9.0466	7.7165
AudioVisual-VAD				
15dB	SpIND	0.2927	2.7075	1.5001
0dB	SpIND	0.8824	2.5851	1.7667
-5dB	SpIND	3.8933	5.4793	4.6863
ITU-I Annex B G.729 VAD (4)				
Noise	**Case**	P_c%	P_f%	P_e%
15dB	SpIND	23.4736	14.7329	19.1032
0dB	SpIND	75.3260	1.5916	38.4588
-5dB	SpIND	98.9634	1.2982	50.1308

The performance of the proposed VAD at 15dB SNR is almost optimal, with P_e being as low as 1.48% for the speaker dependent (SpDEP) and 1.80% for the speaker independent (SpIND) scenario (Table 3.1). This can also be observed from Fig. 6.6 where the input signal, the optimal (annotated) decisions and the produced decisions are depicted for this case. As the noise level increases the accuracy of the proposed VAD deteriorates rising to 2.69% for the speaker independent case at 0dB.

However, it can perform satisfactorily even under -5dB of additive white Gaussian noise where it achieves voice activity detection rates of 88% for the speaker independent case.

The performance of the proposed system is superior than that of the G.729 when these perform using FF microphones. This can be observed from Table 6.2, where it is presented that G.729 fails to deliver high accuracy results when performing on reverberant and noisy signals captured with FF microphones. Moreover, its performance is biased toward noise as can be observed from the values (voice clipping effect). The effectiveness of the introduced adaptive threshold is depicted in Fig. 6.6(b) where it is shown that as time evolves the threshold value varies in order to better reflect the dynamic characteristics of the captured audio signals.

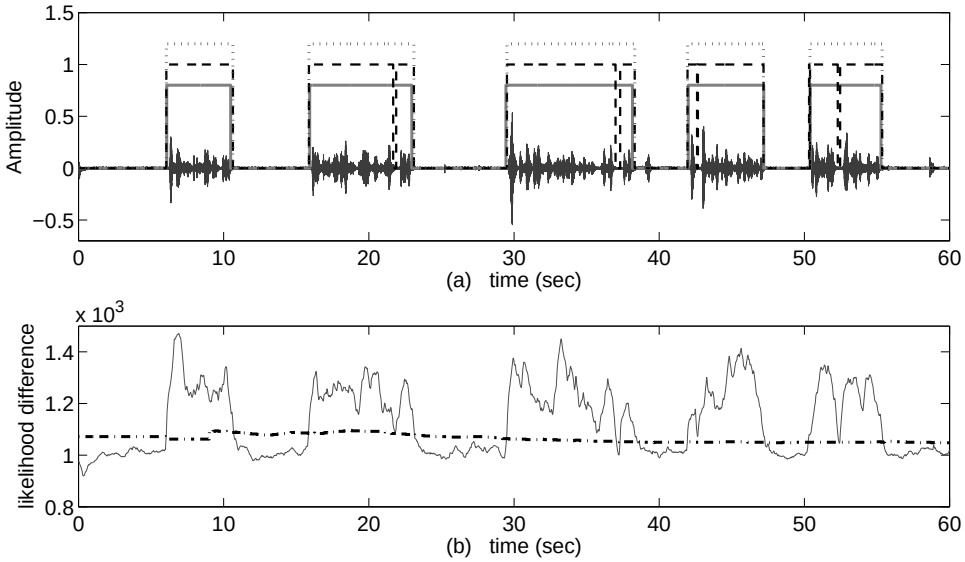

Figure 6.6: (a)Decision of the proposed VAD system before (dashed line) and after (dotted line) the application of the hang-over scheme, along with the optimal decisions (solid line) (b) The difference of the noisy speech from the noise likelihoods plotted along with the adaptive threshold.

The performance of the visual modality is not a function of the noise level. However, its performance is not optimal since lip-movement (a) commences before speed production and stops after speech pause, (b) does not necessarily imply speech production and (c) might not be visible at all times by the camera. So P_e for Visual-VAD is limited to 7.71%. Finally, the performance of the proposed multimodal system is superior than that of both unimodal systems in terms of accuracy and robustness (Table 6.2). The fused system combines the advantages of the Visual-VAD with those of the Audio-VAD. As a consequence it can perform with remarkable accuracy even under adverse noise conditions (P_e=4.68%), or under poor illumination conditions. Moreover, due to the hierarchical fusion system, the performance of the AudioVisual-VAD excels that of its consisting

components. For instance, in the 0dB case, the average detection rate of AudioVisual-VAD is 1.76%, which is smaller than that of the Audio-VAD (2.63%) and of the Visual-VAD (7.71%).

6.5 Conclusions

In this chapter a supervised voice activity detection system suitable for non-intrusive ambient intelligence applications has been presented. The modelling capabilities of hidden Markov models have been combined with Gaussian Mixtures per model state to cater for the variable distribution of speech. Given the bi-modality of speech generation process, conveying both audio and visual information, an AudioVisual VAD that combines the advantages of both modalities has been also considered. The employed hierarchical fusion scheme operated at the decision level. Although the developed system wasn't based on two perfect modalities, fusing of different VAD showed that there is a noticeable increment in performance even under extremely adverse conditions.

Part IV

Applications of Voice Activity Detection

Chapter 7

Time Delay Estimation

7.1 Introduction

Voice activity detection is a fundamental component of several speech processing systems. Its operation is widely combined with other systems to enhance their performance through the discrimination of speech active and inactive periods. Speech processing systems encapsulating VAD benefit from the fact that they can operate only within speech active periods, which results in allowing resource reallocation and error rate reduction. In this Chapter, performance benefits will be demonstrated by combining the multi-microphone VAD developed in Chapter 4 with a direction-of-arrival (DOA) estimation scheme. The core of the DOA system is based on a Mutual Information (MI) Time delay estimation (TDE) scheme. Optimization of the TDE scheme is considered, prior performance enhancement with VAD, by encapsulating speech-shaped distributions in the evaluation of TDE. This is done in respect to the observations made in Chapter 2 and Chapter 3. Towards this direction, the underlying assumption of Gaussian distributed source is replaced by that of Generalized Gaussian distribution that allows evaluating the problem under a larger set of speech-shaped distributions, ranging from Gaussian to Laplacian and Gamma. Univariate and multivariate entropy expressions of the Generalized Gaussian distribution are estimated in order to evaluate the specific information theoretical TDE scheme. The analysis performed, revealed a significant outcome. It is shown that the employed marginal Mutual Information criterion TDE proposed in (52) is not depended on the underlying assumption for the distribution of speech as long as it belongs to the the Generalised Gaussian Distribution (GGD) family.

7.2 Time Delay Estimation

TDE algorithms are embedded in many applications related to localization and tracking of sources, as part of DOA estimating systems. Systems of interest are voice, sonar and radars (53, 54, 55). The principle of operation is a literature standard. For the acoustic source tracking scenario, the

problem is approached by employing distant microphone arrays for the collection of data in frames, so that the current TDE estimate can be provided. DOA estimation relies on identifying the relative delay between pairs of microphones using some statistical measure, that returns a peak at the correct DOA of the source.

The generalized cross-correlation (GCC) algorithm, proposed by Knapp and Carter (56), is generally considered to be the most common method for TDE (57). The delay estimate function is provided, by calculating the cross-correlation between the microphone signals and then searching for the time-delay that maximizes it. The typical limitation of GCC is that if the system is used in reverberant and noisy environments, the maximum cross-correlation could occur in an erroneous time-delay created due to the room's reverberation. Several methods have been proposed to overcome such issues, including the multichannel cross-correlation coefficient (MCCC) (58), able to perform adequately in higher noise and reverberation levels by taking advantage of the redundant information from multiple sensor pairs. However, for non-Gaussian source signals or with limited number of sensors, it is not any better than GCC.

More recently, the information-theoretic metric of Mutual Information, which can be considered indirectly as higher order statistics (HOS), was employed in order to evolve TDE (52). Based on characterizing the speech source as Gaussian, the marginal MI measure was used for TDE. In addition, in order to overcome reverberation problems more effectively, the MI scheme was modified to encapsulate information about reflections, improving significantly the estimator's robustness against reverberation.

However, all of the aforementioned systems operate under the assumption of Gaussian distributed (GD) source, something not true for speech. It has been shown (23, 24) that in several feature domains including time, Fourier Transform (DFT), Discrete Cosine Transform (DCT), and Karhunen Loeve Transform (KLT), distributions of clean speech, with SNR down to 20dB, can be very well approximated by Laplacian (LD) and Gamma distributions (ΓD). In addition, speech distribution varies with time and can be affected by several unpredictable factors including speaker's temper, mood, and environmental characteristics.

Additionally, when far-field microphones are used instead of the conventional close-talking, reverberation effects, competitive sound sources, and speaker movement can alter the distribution of captured speech. It becomes apparent that speech is not solely GD, LD, or ΓD distributed, given its nonstationarity in time and its dependence on external interferences. Thus, speech processing systems relying solely on the Gaussian assumption fail to perform adequately under varying conditions.

Towards the direction of embedding speech shaped distributions for TDE, the authors in (59) worked on modeling speech with a Laplacian distribution. The relative delay was estimated via

minimizing the joint entropy of the multiple microphone output signals. A comparison study (60) presenting performance differences when employing either Laplacian or Gaussian modeling on the information theoretical TDE of (52) was performed showing similar performance for both systems with in fact the Gaussian one performing marginally better. Nevertheless, the Laplacian framework presented in (59) was based on empirical approximations in order to evaluate the expectations involved in TDE estimation and the multivariate LD, not allowing for a fair comparison.

7.3 Information Theory in Time Delay Estimation

In general, GCC (56) and its variants like the GCC-PHAT (57) are the most common methods for TDE. They have the property of exhibiting a global maximum at the lag value that corresponds to the correct sample delay δ that is syncing the mic recordings. The corresponding delay δ can be converted to the source's DOA angle θ by using

$$\theta = \arcsin\left[\frac{\delta c}{f_s d}\right] \tag{7.1}$$

where f_s is the sampling frequency of the recording system, and c is the speed of sound (typically defined as 343 m/s). Thus, the DOA can be obtained by estimating the TDE δ. Note that we restrict the estimation system to integer-valued delays δ, for which several of the values of θ will correspond to the same integer delay. This defines the *resolution* of the array, and it is a function of the chosen values of d and f_s (61).

The problem of the GCC family algorithms is that they cannot perform adequately within reverberant environments described by the model of (4.1), failing to return accurate estimates of the relative delay δ. This becomes even more evident for relatively high T_{60} values.

7.3.1 Mutual Information based TDE

In order to overcome this drawback, some researchers steered their focus on methods employing information theory, aiming to remain robust under adverse conditions. One of them is Mutual Information, a measure of how much information one random variable contains about another random variable. Without loss of generality, we may consider the signals $\mathbf{x_1}$ and $\mathbf{x_2}$ captured by two distant microphones m_1 and m_2 to be stationary stochastic processes, for which the MI between them is defined as (62)

$$I = H[\mathbf{x_1}] + H[\mathbf{x_2}(\delta)] - H[\mathbf{x_1}, \mathbf{x_2}(\delta)] \tag{7.2}$$

where $H[\mathbf{x_m}]$ is the differential entropy of $\mathbf{x_m}$, and $H[\mathbf{x_1}, \mathbf{x_2}(\delta)]$ is the joint entropy of the captured signal $\mathbf{x_1}$ and the delayed by δ samples $\mathbf{x_2}(\delta)$. In the scope of TDE, the problem of finding the correct relative delay between the two signals is equivalent to finding the delay δ that maximizes (7.2). This practically means that when we determine this delay and synchronize the two microphone signals, the information that one microphone signal has about the other will be maximum.

If we let $\mathbf{x_n}$ be an observation vector of a random variable, with density function $p(\mathbf{x_n})$, the differential entropy is defined as

$$\begin{aligned} H(\mathbf{x_n}) &= -\int p(\mathbf{x_n}) \ln p(\mathbf{x_n}) dx \\ &= -E\{\ln p(\mathbf{x_n})\} \end{aligned} \tag{7.3}$$

where $E\{\cdot\}$ denotes mathematical expectation.

If we now consider N observation vectors of random variables

$$\mathbf{x} = [\mathbf{x_1} \quad \mathbf{x_2} \quad \dots \quad \mathbf{x_N}]^T \tag{7.4}$$

with joint density $p(\mathbf{x})$ their joint entropy will be

$$\begin{aligned} H(\mathbf{x}) &= -\int p(\mathbf{x}) \ln p(\mathbf{x}) dx \\ &= -E\{\ln p(\mathbf{x})\} \end{aligned} \tag{7.5}$$

Assuming that the random variables $\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}$ are Gaussian distributed signals their multivariate normal distribution with zero mean and covariance matrix

$$\begin{aligned} \mathbf{R} &= E\{\mathbf{x}\mathbf{x}^T\} \\ &= \begin{bmatrix} \sigma_{x_1}^2 & r_{x_1x_2} & \dots & r_{x_1x_N} \\ r_{x_1x_2} & \sigma_{x_2}^2 & \dots & r_{x_2x_N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_1x_N} & r_{x_2x_N} & \dots & \sigma_{x_N}^2 \end{bmatrix} \end{aligned} \tag{7.6}$$

will be given by

$$p_G(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^N [\det(\mathbf{R})]^{\frac{1}{2}}} \exp^{-\frac{1}{2}\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}}. \tag{7.7}$$

By substituting (7.7) into (7.5) we compute the joint multivariate Gaussian entropy

$$
\begin{aligned}
H_G^m(\mathbf{x}) &= -\int p_G(\mathbf{x}) \ln p_G(\mathbf{x}) d\mathbf{x} \\
&= -\int p_G(\mathbf{x}) \ln \left[\frac{1}{\left(\sqrt{2\pi}\right)^N [\det(\mathbf{R})]^{\frac{1}{2}}} \right. \\
&\qquad \left. \times \exp^{-\frac{1}{2}\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}} \right] d\mathbf{x} \\
&= \int p_G(\mathbf{x}) \left[\ln \left\{ \left(\sqrt{2\pi}\right)^N [\det(\mathbf{R})]^{\frac{1}{2}} \right\} \right. \\
&\qquad \left. + \frac{1}{2}\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x} \right] d\mathbf{x} \\
&= \ln \left\{ \left(\sqrt{2\pi}\right)^N [\det(\mathbf{R})]^{\frac{1}{2}} \right\} \int p_G(\mathbf{x}) d\mathbf{x} \\
&\qquad + \frac{1}{2} \int p_G(\mathbf{x}) \mathbf{x}^T\mathbf{R}^{-1}\mathbf{x} d\mathbf{x} \\
&= \frac{1}{2} \ln \left\{ (2\pi)^N \det(\mathbf{R}) \right\} + \frac{1}{2} E\left\{ \mathbf{x}^T\mathbf{R}^{-1}\mathbf{x} \right\} \qquad (7.8)
\end{aligned}
$$

In order to evaluate the expectation $E\left\{\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right\}$ the trace property is employed so that $E\left[U^TVU\right] = E\left[tr(U^TVU)\right] = tr(VE\left[UU^T\right]) = tr(I) = N$ where N the size of the identity matrix (62).

Thus the joint multivariate Gaussian entropy will be

$$
\begin{aligned}
H_G^m(\mathbf{x}) &= \frac{1}{2} \ln \left\{ (2\pi)^N \det(\mathbf{R}) \right\} + \frac{1}{2} E\left\{ \mathbf{x}^T\mathbf{R}^{-1}\mathbf{x} \right\} \qquad (7.9) \\
&= \frac{1}{2} tr\left\{ E\left[\mathbf{R}^{-1}\mathbf{x}\mathbf{x}^T \right] \right\} + \frac{1}{2} \ln \left\{ (2\pi)^N \det(\mathbf{R}) \right\} \\
&= \frac{1}{2} N + \frac{1}{2} \ln \left\{ (2\pi)^N \det(\mathbf{R}) \right\} \\
&= \frac{1}{2} \ln \left\{ (2\pi e)^N \det(\mathbf{R}) \right\} \qquad (7.10)
\end{aligned}
$$

Intuitively, the corresponding univariate entropy for any of the random variables $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}$ is given by

$$
H_G^u(x_n) = \frac{1}{2} \ln \left\{ 2\pi e \sigma_{x_n}^2 \right\} \qquad (7.11)
$$

If we assume that the source signal is zero-mean Gaussian distributed, the MI of (7.2) will be equal to (62)

$$
I = -\frac{1}{2} \ln \frac{\det[\mathbf{C}(\delta)]}{C_{11}C_{12}} \qquad (7.12)
$$

with $det[\cdot]$ the determinant operator and $\mathbf{C}(\delta)$ the joint covariance matrix of the microphone signals. For large frame size L (ideally $L \to \infty$) $\mathbf{C}(\delta)$ can be approximated as

$$\mathbf{C}(\delta) \approx \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2}(\delta) \end{bmatrix} \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2}(\delta) \end{bmatrix}^T = \begin{bmatrix} C_{11} & C_{12}(\delta) \\ C_{21}(\delta) & C_{22} \end{bmatrix}. \tag{7.13}$$

Note that C_{11} and C_{22} are time-shift independent variables. The relative delay is obtained as that delay that maximizes (7.12) through the evaluation of $\hat{\delta} = \arg\max_\delta \{I\}$.

As described in (52), given the theoretical equivalence between maximizing the MI in (7.12) and the GCC algorithm, which is in fact the time-domain interpretation of the basic form of the GCC method, the MI-based estimator suffers from the same limitations of GCC and its PHAT variant, i.e., it would not be robust enough in multi-path environments. Thus, the MI calculation of (7.12) is not representative enough in the presence of reverberation. In order to estimate the information between the microphone signals, we use the marginal MI that considers jointly neighbouring samples and can be formulated as follows (62):

$$\begin{aligned} I_{N(G)} =& H_G^u[\mathbf{x_1}] + H_G^u[\mathbf{x_1}(1)] + \ldots + H_G^u[\mathbf{x_1}(D)] \\ &+ H_G^u[\mathbf{x_2}(\delta)] + H_G^u[\mathbf{x_2}(\delta+1)] + \ldots \\ &+ H_G^u[\mathbf{x_2}(\delta+D)] - H_G^m[\mathbf{x_1}, \mathbf{x_1}(1), \ldots, \\ &\mathbf{x_1}(D), \mathbf{x_2}(\delta), \mathbf{x_2}(\delta+1), \ldots, \mathbf{x_2}(\delta+D)] \end{aligned} \tag{7.14}$$

which reduces to the following expression for the Gaussian distributed signals

$$I_{N(G)} = -\frac{1}{2} \ln \frac{\det[\mathbf{C}(\delta)]}{\det[\mathbf{C_{11}}]\det[\mathbf{C_{22}}]} \tag{7.15}$$

with the joint covariance matrix

$$\begin{aligned} \mathbf{C}(\delta) &\approx \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_1}(1) \\ \vdots \\ \mathbf{x_1}(D) \\ \mathbf{x_2}(\delta) \\ \mathbf{x_2}(\delta+1) \\ \vdots \\ \mathbf{x_2}(\delta+D) \end{bmatrix} \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_1}(1) \\ \vdots \\ \mathbf{x_1}(D) \\ \mathbf{x_2}(\delta) \\ \mathbf{x_2}(\delta+1) \\ \vdots \\ \mathbf{x_2}(\delta+D) \end{bmatrix}^T \\ &= \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12}(\delta) \\ \mathbf{C}_{21}(\delta) & \mathbf{C}_{22} \end{bmatrix}. \end{aligned} \tag{7.16}$$

If D is chosen to be greater than zero, the elements of $\mathbf{C}(\delta)$ are now themselves matrices. In fact, for any value of δ, the size of $\mathbf{C}(\delta)$ is always $2(D+1) \times 2(D+1)$ where $N = 2(D+1)$. We call D the order of the tracking system.

7.4 Employing Laplacian Distribution for TDE

A Gaussian random variable has the highest entropy of all random variables for a given variance. Hence, a Gaussian random variable is, in some sense, the least predictable of all, which is why the GD is usually associated with noise. Information bearing signals contain structures that make them more predictable than GD random variables (63). Those characteristic structures directly affect the distributions of such signals which deviate significantly from GD. Hence, given that speech is fundamentally an information bearing signal, one should look for more accurate representation of its distribution rather than employing GD.

Furthermore, when it comes to acoustic environments, where the signal of interest is typically speech, GD modelling can be accurate only under specific conditions of reverberation and noise as was shown in Section II. Thus, under the rough and inaccurate assumption of GD speech for TDE, performance reduction should be expected. Towards the direction of substituting the Gaussian entropy assumption with entropies the distributions of which fit better speech characteristics, the authors in (59) worked on the derivation of the Laplacian Entropy.

Using an approximation of the multivariate Laplacian pdf of $x_1, x_2, \ldots, x_N$ given by

$$p(\mathbf{x}) = 2(2\pi)^{-\frac{N}{2}} \left[\det(\mathbf{R})\right]^{-\frac{1}{2}} \left(\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right)^{\frac{P}{2}} \times K_P\left(\sqrt{2\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}}\right) \tag{7.17}$$

where $P = (2-N)/2$ and $K_P(\cdot)$ is the modified Bessel function of the second kind that is given by

$$K_P(\alpha) = \frac{1}{2}\left(\frac{\alpha}{2}\right)^P \int_0^\infty z^{-P-1} \exp\left(-z - \frac{\alpha^2}{4z}\right) dz,\ a > 0. \tag{7.18}$$

The multivariate differential entropy Laplacian for random variables $x_1, x_2, \ldots, x_N$ was given by

$$H_L^m(\mathbf{x}) = \frac{1}{2}\ln\left[\frac{(2\pi)^N}{4}\det(\mathbf{R})\right] - \frac{P}{2}E\left\{\ln\left(\frac{\theta}{2}\right)\right\} - E\left\{\ln K_P\left(\sqrt{2\theta}\right)\right\} \tag{7.19}$$

where $\theta = \mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}$.

The two quantities $E\left\{\ln\left(\frac{\theta}{2}\right)\right\}$ and $E\left\{\ln K_P\left(\sqrt{2\theta}\right)\right\}$ cannot be represented in a closed form. Thus, a numerical method to estimate the expectations has been proposed, assuming that all

processes are ergodic so as to replace ensemble averages by time averages. For K samples of each element of the observation vector the following estimators were proposed:

$$E\left\{\ln\left(\frac{\theta}{2}\right)\right\} \approx \frac{1}{K}\sum_{k'=0}^{K-1}\ln\left[\theta(k-k',m)/2\right] \tag{7.20}$$

$$E\left\{\ln K_P\left(\sqrt{2\theta}\right)\right\} \approx \frac{1}{K}\sum_{k'=0}^{K-1}\ln K_P\left[\sqrt{2\theta(k-k',m)}\right] \tag{7.21}$$

where

$$\theta(k-k',m) = \mathbf{x}(k-k',m)^T\mathbf{R}^{-1}(m)\mathbf{x}(k-k',m) \tag{7.22}$$

In practice, $\mathbf{R}(m)$ is first estimated with K observations of $\mathbf{x}(k,m)$. When the covariance matrix is estimated, the same data is used to estimate (7.20) and (7.21).

The univariate zero-mean Laplace distribution pdf is given by

$$p(x) = \frac{\sqrt{2}}{2\sigma_x}\exp^{-\frac{\sqrt{2}|x|}{\sigma_x}} \tag{7.23}$$

and its entropy is

$$H_L^u(\mathbf{x}) = 1 + \ln\left(\sqrt{2\sigma_x}\right) \tag{7.24}$$

The comparison study performed in (60) on the effectiveness of different information-theoretical TDE techniques revealed that, maximizing the MI for TDE gives more consistent results compared to minimizing the joint entropy given that MI is insensitive to the variance changes of sensor outputs. Furthermore, the authors employed the Laplacian entropy proposed by (59) in the estimation of the marginal MI in (7.14) to investigate possible performance alterations. The marginal MI for Laplacian distribution assumption is estimated as

$$\begin{aligned} I_{N(L)} =& H_L^u[\mathbf{x_1}] + H_L^u[\mathbf{x_1}(1)] + \ldots + H_L^u[\mathbf{x_1}(D)] \\ &+ H_L^u[\mathbf{x_2}(\delta)] + H_L^u[\mathbf{x_2}(\delta+1)] + \ldots \\ &+ H_L^u[\mathbf{x_2}(\delta+D)] - H_L^m[\mathbf{x_1}, \mathbf{x_1}(1), \ldots, \\ &\mathbf{x_1}(D), \mathbf{x_2}(\delta), \mathbf{x_2}(\delta+1), \ldots, \mathbf{x_2}(\delta+D)]. \end{aligned} \tag{7.25}$$

Due to the approximations employed in (7.20) and (7.21) a closed form for (7.25) cannot be derived. Simulations performed under various reverberant conditions demonstrated that employing GD models results in performing similarly or slightly better than employing LD for TDE. Nevertheless, the Laplacian framework proposed in (59) includes several approximations that do not allow

for a fair comparison (i.e. the multivariate LD, empirical approximations for the expectations) something that can possibly explain the reduced performance of LD. The next section provides solutions for the problems caused by the inaccuracies of the investigated Laplacian framework through a more generalized approach regarding the underlying distribution.

7.5 Employing Generalized Gaussian Distribution

In order to further investigate how the adaptation of distributions of higher super-gaussianity affects the performance of TDE, we have to evaluate the output changes of (7.14) as we employ different underlying distributions. For the scope of the this work, in order to deal with such comparisons along a wider set of distributions we took advantage of the properties of the Multivariate Generalized Gaussian distribution (MGGD).

The generalized Gaussian distribution represents an extension of the standard Gaussian distribution which comprises of three parameters, mean, variance and the shape parameter. The latter is a measure of the peakedness of the pdf, and allows the GG to approximate a large class of statistical distributions, including the Gaussian, the Laplacian, and the Gamma distributions which are very close to the distribution of speech.

The N-dimensional zero-mean Generalized Gaussian (GG) distribution for $x_1, x_2, \ldots, x_N$ is defined as (64)

$$p_{GG}(\mathbf{x}) = \frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^N} \exp\left\{-\frac{1}{2}[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}]^{\frac{\beta}{2}}\right\} \tag{7.26}$$

where β is the shape parameter. $Z(\beta) = \frac{2}{\beta}\Gamma\left(\frac{1}{\beta}\right)$ and $A(\beta) = \sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}}$ with Γ the Gamma function. The Gamma, Laplacian, Gaussian and Uniform distributions are special cases of the GGD, with $\beta = \frac{1}{2}$ $\beta = 1$, $\beta = 2$ and $\beta = \infty$ respectively.

Through the GGD all multivariate expressions of distributions can be represented in a closed form avoiding the usage of approximations like in (7.17) that can potentially result in performance degradation.

The joint entropy for the generalized Gaussian random variables $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}$ is given by

$$
\begin{aligned}
H_{GG}^{m}(\mathbf{x}) &= -\int p_{GG}(\mathbf{x}) \ln p_{GG}(\mathbf{x}) d\mathbf{x} \\
&= -\int p_{GG}(\mathbf{x}) \ln \left[\frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^{N}} \right. \\
&\qquad \left. \times \exp\left\{ -\tfrac{1}{2}\left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \right\} \right] d\mathbf{x} \\
&= -\int p_{GG}(\mathbf{x}) \left[\ln \left\{ \frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^{N}} \right\} \right. \\
&\qquad \left. - \frac{1}{2}\left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \right] d\mathbf{x} \\
&= \int p_{GG}(\mathbf{x}) \left[-\ln \left\{ \frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^{N}} \right\} \right. \\
&\qquad \left. + \frac{1}{2}\left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \right] d\mathbf{x} \\
&= \frac{1}{2}\int p_{GG}(\mathbf{x}) \left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} d\mathbf{x} \\
&\qquad - \ln \left\{ \frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^{N}} \right\} \int p_{GG}(\mathbf{x}) d\mathbf{x} \\
&= \frac{1}{2} E\left\{ \left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \right\} - \ln \left\{ \frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^{N}} \right\}
\end{aligned} \tag{7.27}
$$

The expectation $E\left\{ \left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \right\}$ in the left part of (7.27) cannot be evaluated through the *trace* property as for the case of multivariate Gaussian entropy in (7.8). The methodology proposed in (59) for evaluating the expectations for the Laplacian joint entropy could be followed instead. Thus, we have to evaluate the quantity

$$
E\left\{ \left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \right\} \approx \frac{1}{K}\sum_{k'=0}^{K-1} \left[\theta(k-k',m)\right]^{\frac{\beta}{2}} \tag{7.28}
$$

Nevertheless, using such approximations will result in an instant input dependent system, something definitely not beneficial for comparing TDE systems based on different distribution assumptions.

The specific expectation $E\left\{ \left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \right\}$ is actually a Dirichlet integral of type 1 (64, 65). We note that the expectations over the whole parameter space $\Re^N$ of a function $\phi(\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}) = \phi(\mathbf{z}^T\mathbf{z}) \equiv \phi(u)$ with $u > 0$ for $\mathbf{x} \neq 0$ can be reduced to integrals over $\Re^+$ (for non-negative functions

$\phi(u)$). Thus, for $\phi(u) = u^{(\beta/2)}$ the expectation becomes

$$\begin{aligned}
E\left\{\left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}}\right\} = \\
&= \int_{\Re^N}\left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} p_{GG}(\mathbf{x})d\mathbf{x} \\
&= \frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^N}\int_{\Re^N}\left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \\
&\qquad\qquad \times \exp\left\{-\tfrac{1}{2}\left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}}\right\} d\mathbf{x} \\
&= \frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^N}\int_{\Re^N}\phi\left(\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right) \\
&\qquad\qquad \times \exp\left\{-\tfrac{1}{2}\left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}}\right\} d\mathbf{x} \\
&= \frac{\beta}{\Gamma\left(\frac{N}{\beta}\right)2^{\left(\frac{N}{\beta}+1\right)}}\int_{\Re^+}\phi(u)\,u^{\frac{N}{2}-1}\exp^{\left(-\frac{1}{2}u^{\frac{\beta}{2}}\right)} du \\
&= \frac{\beta}{\Gamma\left(\frac{N}{\beta}\right)2^{\left(\frac{N}{\beta}+1\right)}}\int_{\Re^+}u^{\frac{\beta}{2}+\frac{N}{2}-1}\exp^{\left(-\frac{1}{2}u^{\frac{\beta}{2}}\right)} du \\
&= \frac{\beta}{\Gamma\left(\frac{N}{\beta}\right)2^{\left(\frac{N}{\beta}+1\right)}}\frac{2^{2+\frac{N}{\beta}}\Gamma\left(\frac{\beta+N}{\beta}\right)}{\beta} = \frac{2\Gamma\left(\frac{\beta+N}{\beta}\right)}{\Gamma\left(\frac{N}{\beta}\right)}
\end{aligned} \tag{7.29}$$

Substituting the expectation of (7.27) with (7.29) we get

$$\begin{aligned}
H^m_{GG}(\mathbf{x}) &= -\int p_{GG}(\mathbf{x})\ln p_{GG}(\mathbf{x})d\mathbf{x} \\
&= \frac{2\Gamma\left(\frac{\beta+N}{\beta}\right)}{\Gamma\left(\frac{N}{\beta}\right)} - \ln\left\{\frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^N}\right\}
\end{aligned} \tag{7.30}$$

For the scope of this work the range value of β is $0.5 \leq \beta \leq 2$ and thus, (7.30) reduces to

$$H^m_{GG}(\mathbf{x}) = \frac{N}{\beta} - \ln\left\{\frac{[\det(R)]^{-1/2}}{[Z(\beta)A(\beta)]^N}\right\} \tag{7.31}$$

For the univariate case of the generalized Gaussian distributed variable x the entropy is

$$H^u_{GG}(\mathbf{x_n}) = \frac{1}{\beta} + ln\left[2\Gamma\left(1+\frac{1}{\beta}\right)\sigma_{x_n}\sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}}\right] \tag{7.32}$$

The theoretical solutions of (7.32),(7.31), derived through the evaluation of Dirichlet integrals match exactly all the theoretically evaluated multi- and uni- variate entropies presented in (7.23),(7.11),(7.10) and at the same time provide the multi- and uni- variate entropies over a wider family set of distributions that can be represented through the MGGD or the GGD.

Based on the evaluated expressions for generalized Gaussian Entropies the marginal MI (7.14) has been modified and used with the GG assumption to estimate the sample delay between signals received by a two speaker microphone array. The evaluation has been conducted for values of β in the range $0.5 \leq \beta \leq 2$ that correspond to distributions ranging from Gaussian, to Gamma shaped. The resulting MI is depicted in Fig.7.1

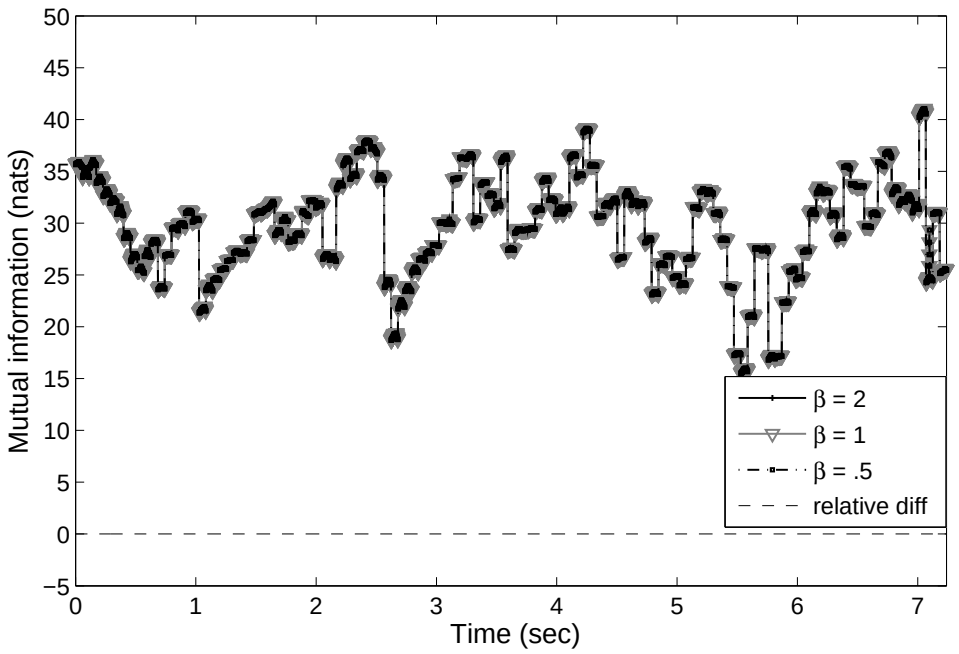

Figure 7.1: Marginal MI estimated for different values of the shape parameter β employing GGD and their sum of absolute relative differences.

The system's response is identical for the different values of β resulting in exactly the same sample delay estimation regardless of the assumed underlying distribution. Those results actually indicate that the estimation of marginal MI does not depend on the underlying distribution.

Indeed, by substituting (7.31) and (7.32) in the estimation of the marginal MI (7.14) we have

$$
\begin{aligned}
I_{N(GG)} =& H^u_{GG}[\mathbf{x_1}] + H^u_{GG}[\mathbf{x_1}(1)] + \ldots + H^u_{GG}[\mathbf{x_1}(D)] \\
&+ H^u_{GG}[\mathbf{x_2}(\delta)] + H^u_{GG}[\mathbf{x_2}(\delta+1)] + \ldots \\
&+ H^u_{GG}[\mathbf{x_2}(\delta+D)] - H^m_{GG}[\mathbf{x_1}, \mathbf{x_1}(1), \ldots, \\
&\mathbf{x_1}(D), \mathbf{x_2}(\delta), \mathbf{x_2}(\delta+1), \ldots, \mathbf{x_2}(\delta+D)] \\
=& \frac{2(D+1)}{\beta} + \ln \left\{ \left[2\Gamma\left(1+\frac{1}{\beta}\right) \sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}} \right]^{2(D+1)} \right. \\
&\left. \times \sqrt{\det[\mathbf{C_{11}}]\det[\mathbf{C_{22}}]} \right\} \\
&- \frac{2(D+1)}{\beta} - \ln \left\{ \left[\frac{2}{\beta}\Gamma\left(\frac{1}{\beta}\right) \sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}} \right]^{2(D+1)} \right. \\
&\left. \times \sqrt{\det[\mathbf{C}]} \right\} \\
=& \ln \frac{\beta\Gamma\left(1+\frac{1}{\beta}\right) \sqrt{\det[\mathbf{C_{11}}]\det[\mathbf{C_{22}}]}}{\Gamma\left(\frac{1}{\beta}\right) \sqrt{\det[\mathbf{C}]}} \\
=& -\frac{1}{2} \ln \frac{\det[\mathbf{C}(\delta)]}{\det[\mathbf{C_{11}}]\det[\mathbf{C_{22}}]}
\end{aligned}
\tag{7.33}
$$

The result is identical to the closed form marginal MI of (7.15) estimated for Gaussian assumed signals. This indicates that the evaluation of marginal MI is distribution independent, verifying the results depicted in Fig.7.1. This outcome directly exploits MI invariance property that implies that if $\acute{X} = F(X)$ and $\acute{Y} = G(Y)$ are homeomorphisms, then $I(X,Y) = I(\acute{X},\acute{Y})$ (66). The underlying distribution assumption for the marginal MI estimation is a function of the β shaping factor i.e. a linear invertible transformation. Thus, MI will be the same regardless of the β value for the given distribution family of GGD. Although the invariance property is numerically extremely useful, it would not hold in general for other interdependence measures. Entropy for example, changes in general under a homeomorphism.

Furthermore, the solution shows that the illustrated difference in performance results in (60) indicating that the Laplacian based TDE in (59) is surpassed by the Gaussian case, is caused due to the approximations used for the Laplacian multivariate distribution employed by the authors in (59) and the proposed empirical approximation for the evaluation of $E\left\{ \left[\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{\frac{\beta}{2}} \right\}$.

7.6 Employing VAD towards assisting DOA Estimation

VAD schemes are being widely used to assist DOA estimators (67, 68, 69, 70). When VAD is adopted before DOA estimation, the system estimates the sound sources position only when the target sound occurs. Speech emission is a discontinuous sound source. Due to short pauses between words and phrases, the performance can be interfered greatly in the presence of a competing sound source during those silent pauses. Therefore, in a similar fashion to speech recognition applications, VAD is indispensable to achieve better performance by restricting operation of DOA only in between speech presence intervals.

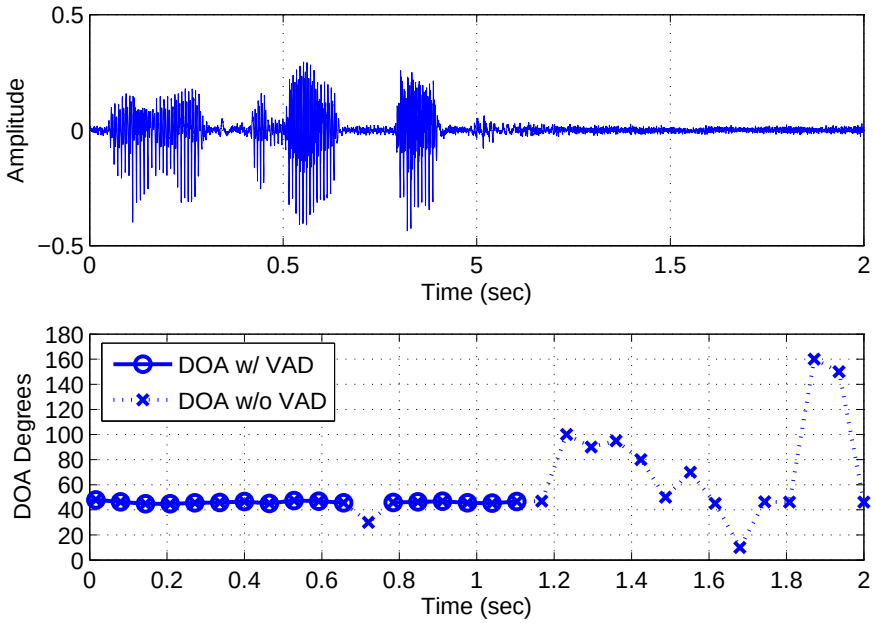

Figure 7.2: Estimating direction of arrival with a source placed at approximately 45 degrees in respect to the array.

In order to demonstrate this effect, the following simulation was carried out by employing the *Image Model for Small room Acoustics* (27) to generate artificial sound waves of a speaker placed at 4 different angles 1 meter away of the center of a linear microphone array. The microphone array consisted of two microphones placed at 0.20m away. In the experiment, the sampling rate was set at 22.1kHz and the reverberation time T_{60}=0.15ms. The simulated room dimensions are [5, 4, 3]m. Framing of 0.128s and 50% overlapping has been used. The estimation of the source's direction is based on the the MI based TDE scheme described in the previous sections for each data frame.

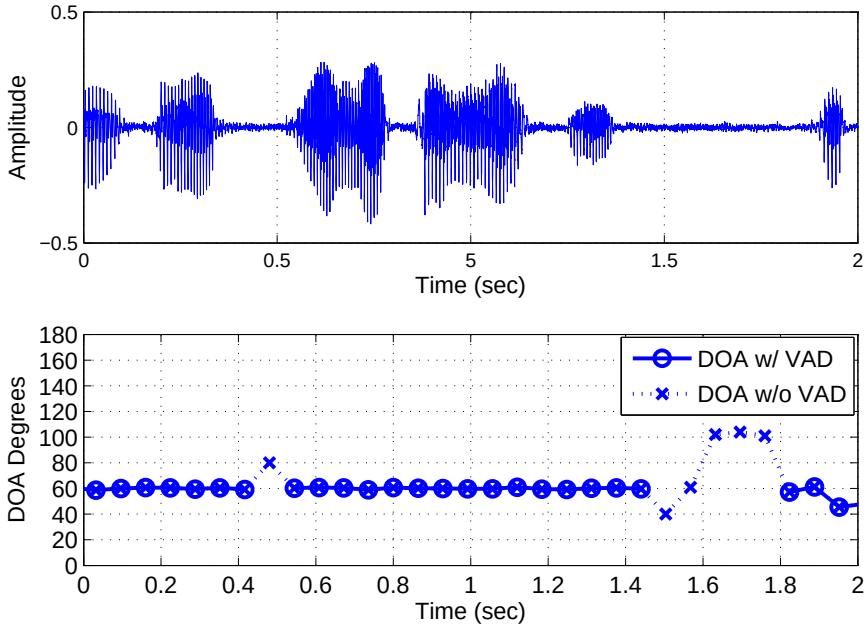

Figure 7.3: Estimating direction of arrival with a source placed at approximately 60 degrees in respect to the array.

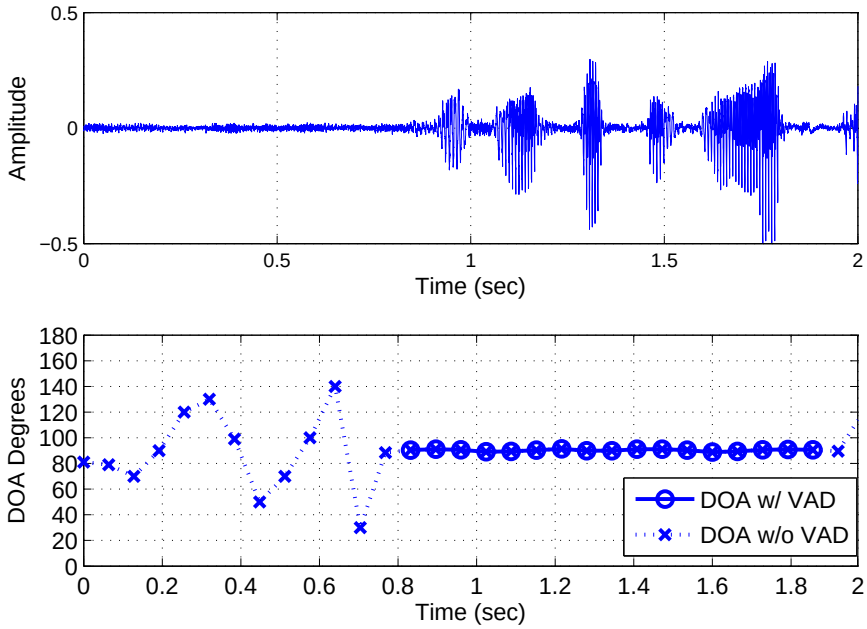

Figure 7.4: Estimating direction of arrival with a source placed at approximately 90 degrees in respect to the array.

Moreover, 15 dB of additive babble noise, one of the most adverse scenaria for DOA, was also introduced to the signals to act as the competitive sound source from NOISEX-92 database. The simulation was performed with and without the presence of a VAD. The VAD employed is the multi-microphone on described in Chapter 4, selected for its ability to encapsulate spatial information of multiple microphones without DOA dependency. Fig. 7.2 to 7.5 depict the outcome of the simulation.

Figures 7.2 to 7.5 show that with high SNR, within speech presence periods, DOA it is almost immune to the reflective environment in and the competitive noise. However, DOA estimation is influenced by noise when the sound source pauses. This is very reasonable because actually there is nothing to listen to at that interval but ambient noise. Nevertheless, the performance of a speaker localization system degrades greatly at those intervals. Unreliability is introduced since without VAD the system is not capable to recognise that voice emission has stoped and thus, tracking the target has to be halted. Nevertheless, when VAD is adopted before DOA estimation, the system estimates the sound sources position only when the target sound occurs.

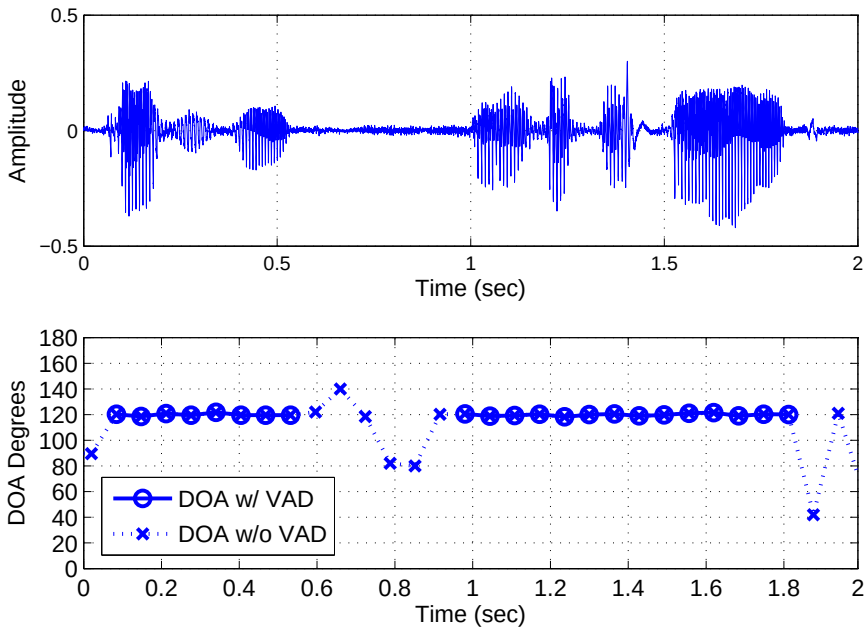

Figure 7.5: Estimating direction of arrival with a source placed at approximately 120 degrees in respect to the array.

7.7 Conclusions

In this Chapter, one of the speech processing technologies the operation of which is usually combined with a VAD has been presented. More specifically, the performance benefits of combining the multi-microphone VAD developed in Chapter 4 with a direction-of-arrival (DOA) estimation scheme were demonstrated. Like most speech processing systems, TDE encapsulating VAD, benefits from the fact that it operates only within speech active periods, which results in allowing resource reallocation and error rate reduction. Given that speech emission is a discontinuous sound source, the performance can be interfered greatly in the presence of background noise during silent pauses. The employed DOA system was based on information theoretical TDE system. The optimization of this TDE scheme was also considered in the context of encapsulating speech shaped distributions in the underlying assumption of the speech model employed. Based on the observations of Chapter 2 and the outcomes of Chapter 3 we investigated how the performance of a robust information-theoretical TDE algorithm, changes as we switch between different underlying assumptions for the distribution of speech. For the scope of the study, the generalized Gaussian distribution has been employed that allowed to investigate the problem under a wide range of super-Gaussian distributions ranging from Gaussian to Gamma. The analysis performed, indicates that the employed marginal MI criterion TDE is not depended on the underlying assumption for the distribution of speech when it belongs to the family of generalized Gaussian distribution, exploiting the invariance property of MI. To support the analysis, closed forms of the multivariate and univariate differential entropies for the Generalized Gaussian distribution were derived, that encapsulate the entropies of other well known distributions like Gaussian, Laplacian and Gamma.

Chapter 8

Noise Reduction

8.1 Introduction

Speech processing systems often operate in noisy and reverberant environments. Their operation is subject to the accuracy of the underlying noise reduction algorithm, that aims to reduce noise present in the signals that are captured by the employed microphones. Under adverse conditions a noise reduction scheme, failing to perform adequately, will produce results characterised by speech distortions (metallic or clipping voice) and/or fluctuating residual background noises, the result of inaccuracy in estimating the noise spectrum, known as musical noise. In this chapter, performance enhancement when using VAD in combination with noise reduction will be discussed in terms of residual noise suppression, present within silence intervals and short pauses during speech production. For this scope an efficient noise reduction architecture has been developed based on cascading the scheme of spectral subtraction based on minimum statistics noise estimation (71, 72, 73). The idea behind the cascaded structures is to initially subtract primary noise and then consequently use the same technique, with specific parameter adjustments, to remove spectral subtraction artefacts such as musical noise or noise leftovers that were generated by the first stage. Moreover, the VAD presented in Chapter 3 is employed to further suppress noise residuals acting as a musical noise reduction scheme.

8.2 Spectral Subtraction Based on Minimum Statistics

Noise reduction is a speech processing technology, aiming at the reduction of the noise level in audio signals and thus the enhancement of speech. Its objective is to increase the SNR and the intelligibility of speech signals captured by microphones. Its operation relies on the performance of two subsystems; a noise estimator that produces estimates of the additive noise signal based on measurements of the noisy speech and a rule that subtracts the estimated noise signal from the noisy one in order to derive the clean speech.

An ideal noise reduction system should remove the noise completely without affecting the voice quality. However, in practice this is not the case; noise reduction systems fail to remove noise completely and to leave voice unaffected. Moreover, their output signal is usually contaminated with residuals of noise subtraction called musical noise. One of the reasons is that, very often, the estimation of noise power spectrum is updated based on the speech presence/absence indication of an underlying VAD being subject to its accuracy. Tracking of varying noise levels is slow given that it is updated only within periods of speech inactivity.

Towards eliminating the problem of incorporating a VAD for updating the estimation for the psd of noise, a minimum tracking of smoothed power estimate values algorithm was introduced in (71). The algorithm has been updated (72, 73) to cater for the bias that is introduced given that the short term minimum power is always smaller than - sparsely equal to - the mean power , the minimum noise power estimate is a biased estimate of the mean power. The approach tracks spectral minima in every frequency band without distinction between speech activity and speech pause. The noisy speech signal is smoothed using an optimal in the mean square sense smoothing parameter and an unbiased noise estimator is developed. Spectral subtraction is performed by computing gain factors for every frequency bin based on the a priori SNR (74). The a priori SNR is computed using the estimated noise spectrum as proposed in (29).

8.2.1 The Algorithm

According to (71) the captured noisy signal $x(\tau)$ is assumed to be the sum of the zero mean speech signal $s(\tau)$ and the zero mean noise signal $n(\tau)$ thus,

$$x(\tau) = s(\tau) + n(\tau) \tag{8.1}$$

where t the time index. $s(\tau)$ and $n(\tau)$ are assumed to be statistically independent. The process of denoising can be summarized in the following steps

- The input signal is processed based on DFT in buffers denoted $w(\tau)$. The DFT of the windowed signal will therefore be
$$X(\lambda, k) = \sum_{n=0}^{W_{DFT}-1} x(\lambda R + n) \cdot w(n) \cdot \exp\left(-j\frac{2\pi nk}{W_{DFT}}\right) \tag{8.2}$$
where W_{DFT} the length of DFT, $k = 0, 1, 2, \ldots, W_{DFT} - 1$ the frequency bin, R denotes the decimation/interpolation factor and λ is the decimated time index.

- To compute the short time signal power $\overline{|X(\lambda,k)|^2}$ subsequent magnitude squared input spectra are smoothed based on an forgetting scheme with $\gamma \leq 0.9$

$$\overline{|X(\lambda,k)|^2} = \gamma\overline{|X(\lambda-1,k)|^2} + (1-\gamma)|X(\lambda,k)|^2 \tag{8.3}$$

- The short time subband signal power $P_x(\lambda,k)$ is computed recursively based on smoothed periodograms and

$$P_x(\lambda,k) = aP_x(\lambda-1,k) + (1-a)|X(\lambda,k)|^2 \tag{8.4}$$

where the smoothing factor is set heuristically within the range $(0.9 \leq a \leq 0.95)$ based on performance evaluation.

- In order to compute the noise power estimate $P_n(\lambda,k)$ we take the short time power estimate $P_x(\lambda,k)$ within a window of D subbund power samples

$$P_n(\lambda,k) = ominP_{min}(\lambda,k) \tag{8.5}$$

where $P_{min}(\lambda,k)$ is the estimated minimum power and $omin$ is the factor used to compensate for the bias of the minimum estimate given by

$$omin = \frac{1}{E\{P_{min}\}|_{\sigma^2(k)=1}} \tag{8.6}$$

with $\sigma^2(k)$ the noise power for the k^{th} frequency bin.

- P_{min} is computed based on the min of the sub-frame when $\mod(\lambda,M)=0$ that is

$$P_{Mmin}(qM-1,k) = \min(P_x(\lambda_0+(q-1)M), P_x(\lambda_0+(q-1)M+1), \ldots, P_x(\lambda_0+qM-1)) \tag{8.7}$$

- The overall mean is given by

$$P_{min}(\lambda_0,k) = \min P_{Mmin}(qM-1,k) \tag{8.8}$$

where $q = 0,1,2,\ldots,W-1$ and W the number of the decomposed windows of size M.

- Finally the power spectrum of the denoised signal is given by

$$|Y(\lambda,k))| = \max(\sqrt{subfP_n(\lambda,k)}, |X(\lambda,k)|Q(\lambda,k)) \tag{8.9}$$

where

$$Q(\lambda,k) = \left(1 - \sqrt{osub(\lambda,k)\frac{P_n(\lambda,k)}{|X(\lambda,k)|^2}}\right) \tag{8.10}$$

that according to (74) the spectral magnitudes subtraction is performed by employing the oversubtraction $osub(\lambda, k)$ factor and by limiting the maximum subtraction by a spectral floor constant $subf(0.01) \leq subf \leq 0.05$.

8.3 Cascaded Noise Reduction

The estimation of the oversubstraction factor $osub$, involved in the processes of spectral subtraction employing minimum statistics, is critical for the quality of output. A large oversubtraction factor per frequency bin $osub(\lambda, k)$ could significantly suppress the residuals of noise reduction process (musical noise). This comes with the cost of reduced speech quality as large oversubtraction is responsible for metallic voice phenomenon and corrupted speech. Towards controlling those effects the oversubtraction factor is computed as a function of the subband SNR $SNR_x(\lambda, k)$ and frequency bin k that is $osub(\lambda, k) = f(\lambda, k, SNR_x(\lambda, k))$ (74). Notice that the subband SNR is estimated by

$$SNR_x(\lambda, k) = 10log\left(\frac{P_x(\lambda, k) - \min(P_n(\lambda, k), P_x(\lambda, k))}{P_n(\lambda, k)}\right) \tag{8.11}$$

Thus, the value of oversubtruction is lower for high SNR conditions (speech presence) and high frequency than low SNR and low frequency. The efficiency of the proposed solution is not optimal under particular conditions of low SNR, where the intensity of source speech is low or the noise is extremely variable. The result when the scheme fails to track accurately instant SNR is that the presence of noise reduction residuals increases and speech is corrupted.

The idea behind cascaded noise reduction is to initially subtract primary noise using Martins spectral subtraction algorithm (71, 72) and then consequently use the same technique to remove spectral subtraction remaining artefacts (musical noise) that are generated in the first subtraction stage. Two approaches have been devised. The first one, directly performs re-estimation of noise parameters on the power spectrum of the denoised signal $|Y(\lambda, k))|^2$ of eq.(8.9). Thus, the input of the second cascaded system is actually the the power spectrum of the denoised signal of the first noise reduction stage

$$|X_{2cas}(\lambda, k)|^2 = |Y_{1cas}(\lambda, k))|^2 \tag{8.12}$$

In order to estimate the output spectral magnitude of the second stage $|Y_{2cas}(\lambda, k))|$ the procedure of spectral subtraction is repeated in order to update the parameter set of eq.(8.3) to (8.9). A block diagram of the process is shown in Fig. 8.1.

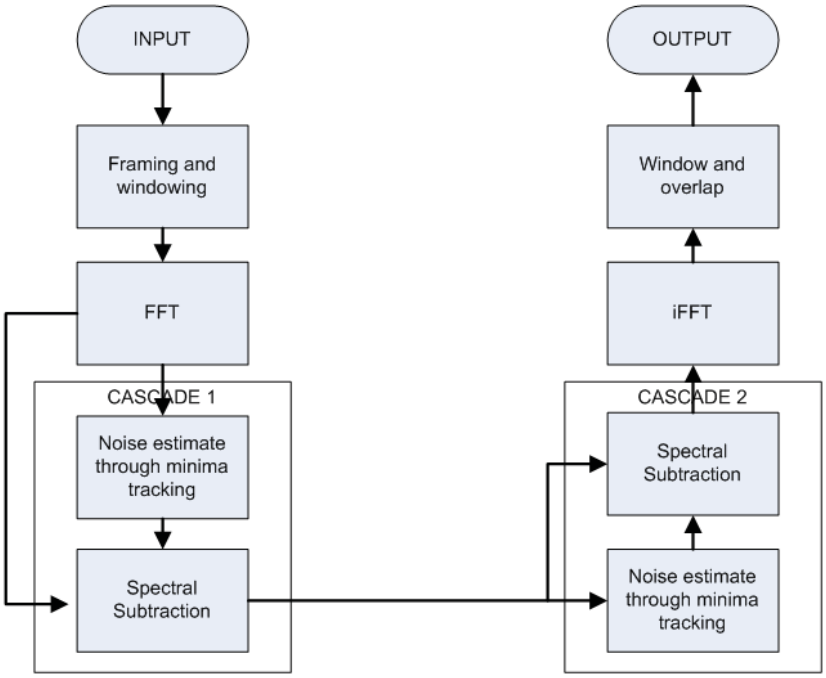

Figure 8.1: Cascaded noise reduction block diagram.

The second approach followed employs cascading before performing overlap and adding to join frames. This way introduces an extra delay (equal to a step) but alters the performance of the cascaded system by reducing framing effects of the signal.

In the two denoising stages, the employed maximum oversubtraction ($ammax$) factor that determines the aggressiveness of spectral subtraction differs. Every stage has different maximum oversubtraction factor value. The first stage is quite aggressive ($ammax = 12$), while the second stage has a lower maximum oversubtraction factor ($ammax = 6$), intending to gently correct the leftovers of the first stage.

The noise reduction enhancement when employing a two-stage cascaded systems is illustrated in the following figures where a noisy speech segment Fig.8.3 (10dB of "destroyer engine" noise from NOISEX-92 database (33)) is denoised by employing the single stage noise reduction scheme Fig.8.4 proposed by Martin (71) and the proposed two-stage cascaded system Fig.8.5.

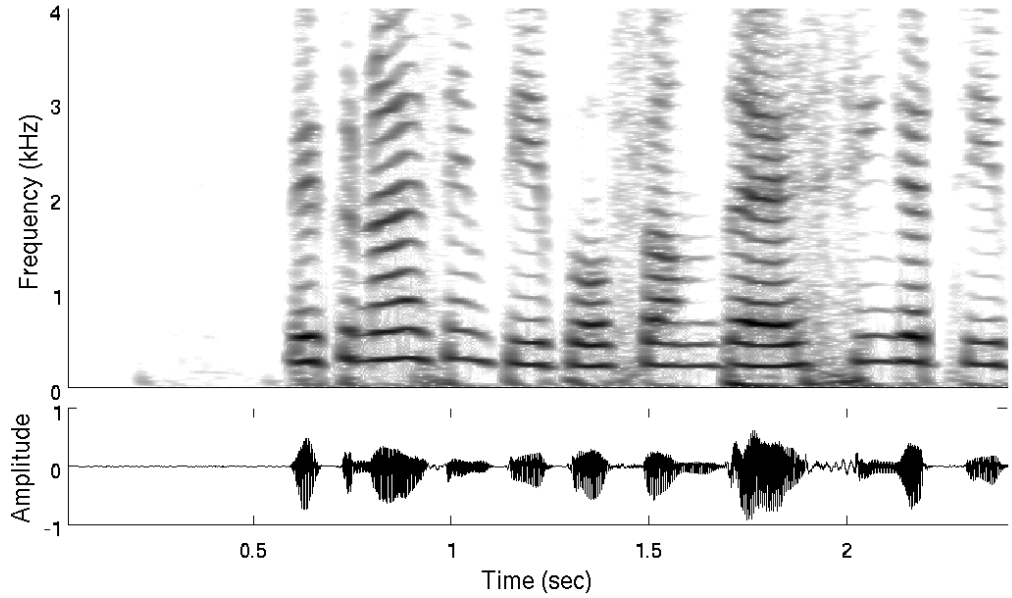

Figure 8.2: Initial speech signal.

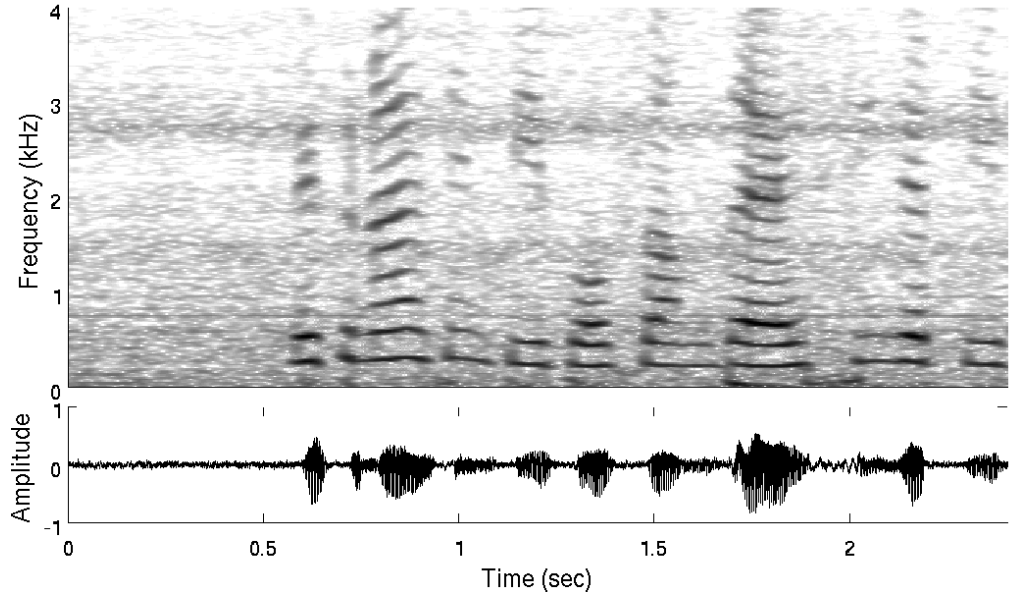

Figure 8.3: Noisy signal 10dB of "destroyer engine" additive noise.

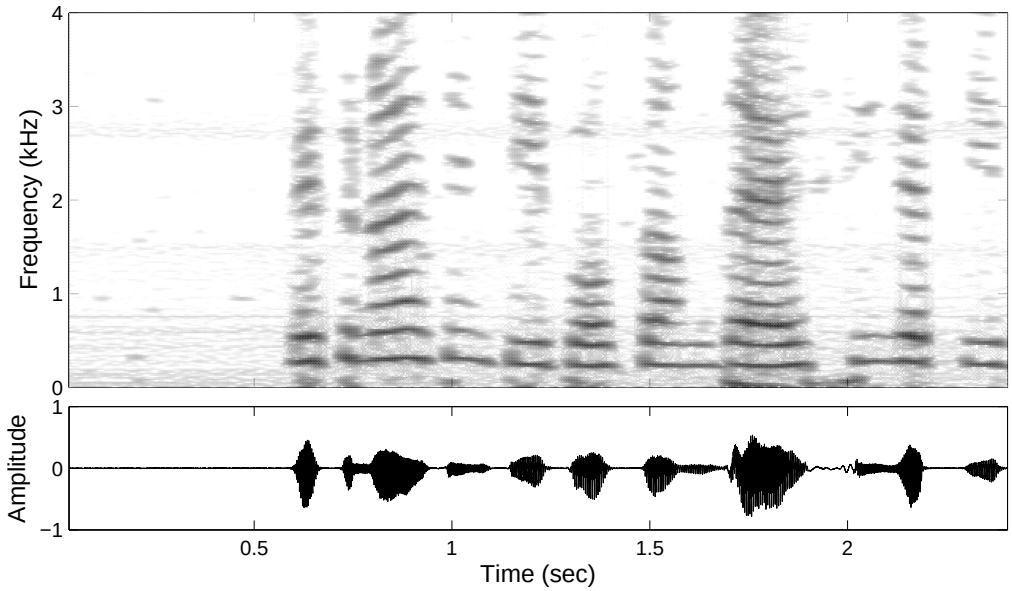

Figure 8.4: Single stage noise reduction output spectrogram.

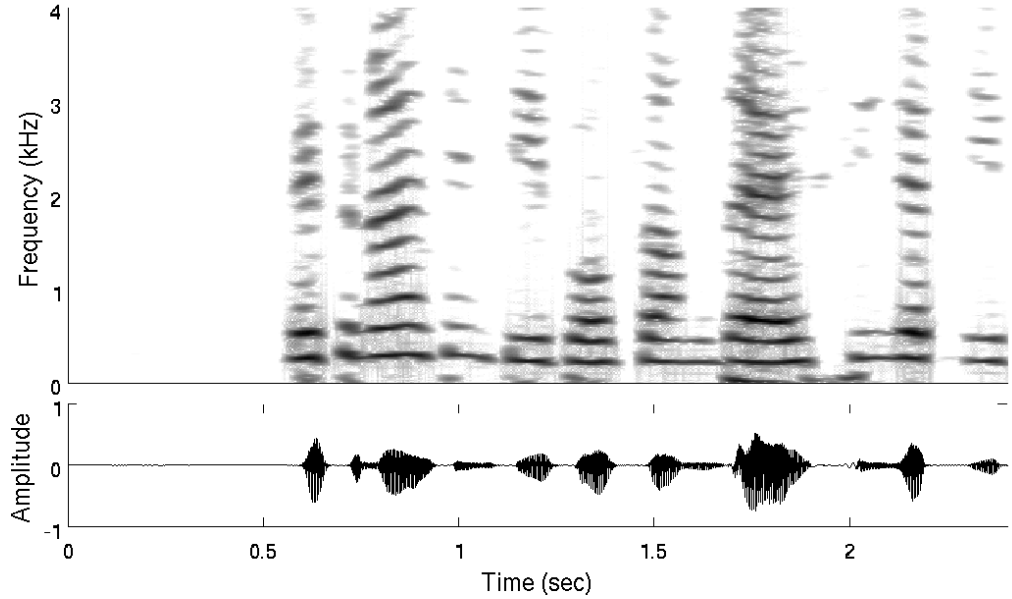

Figure 8.5: Cascaded noise reduction output spectrogram.

As depicted in Fig. 8.4, where the single stage noise reduction scheme of Martin (71) is employed, although most of the frequencies noise is cancelled there are still evident residuals from the spectral subtraction process around the characteristic frequencies of the specific noise type, the psd of which is shown in Fig. 8.6. When employing the proposed cascading scheme, the noise reduction is even higher. The outcome of the cascaded denoising process shown in Fig. 8.5 is very close to the initial clean signal of Fig. 8.2.

8.4 Evaluation of Noise Reduction

The performance of the systems described has been evaluated under several artificial noise-condition scenarios. This extensive evaluation was necessary in order to agree on a generic parameter setup for each state of the cascaded noise reduction scheme. The generic parameter setup ensures that the performance of the system will be satisfactory for most of the noise schemes.

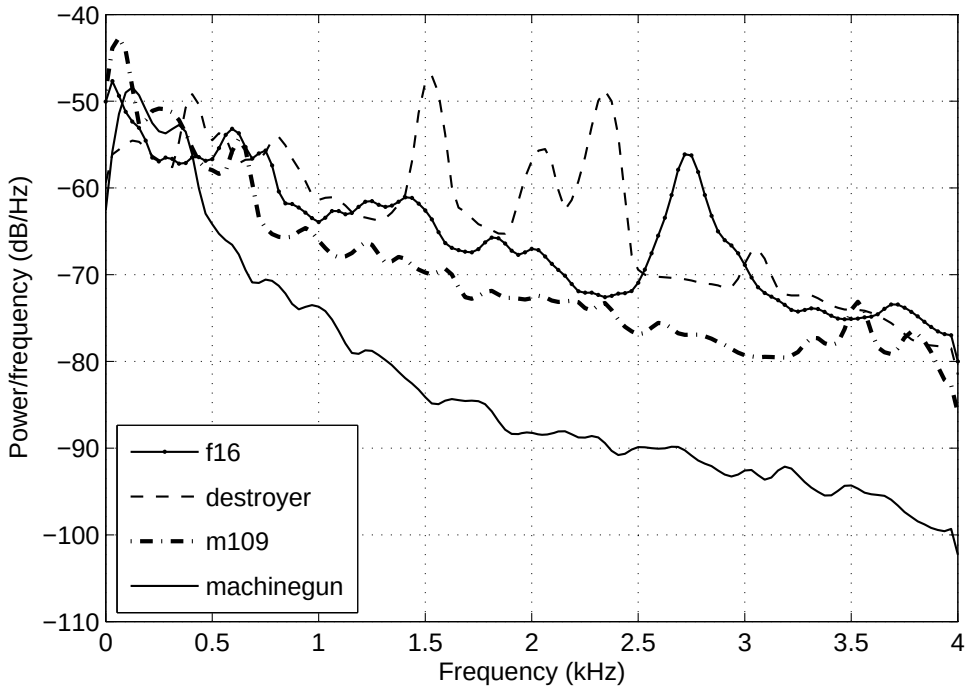

Figure 8.6: Welch Power Spectral Density Estimate of the employed noises.

Speech data captured in the anechoic chamber of AAU were contaminated artificially with "vehicular", "buccaneer engine", "m109", "destroyer engine", "machine gun" and "F-16" noise schemes from NOISEX-92 database (33) at different noise levels. Power spectral density is pre-

sented in Fig. 3.3 and 8.6. Finally, the data were reverberated using the *Image Method* (27) with T_{60}=0.15ms. The input data were sampled at 8 kHz and were segmented into overlapping frames of 256 samples with 64 step size. Thus, the delay introduced by the noise reduction scheme of Martin -tagged as "Single Martin"- and the Cascaded version of this -tagged as "Cascaded"- is 32ms. The cascaded system, the second stage of which operates on the output of the first after overlap and adding the data frames, introduces an extra delay equal to an additional step, with the overall delay being 40ms -tagged as "Cascade + delay".

Moreover, the convex combination VAD presented in Chapter 3 is used as a post-processing block after the "Cascade + delay" system tagged as "Cascade + d + VAD". The combination of the two technologies aims at revealing the noise suppression benefits within silence intervals and short pauses between words, when employing VAD. When VAD is used to trigger a Voice Operated switch (VOX) spectral subtraction residuals ("musical noise") can be significantly suppressed down to zero. Eliminating noise reduction residuals to zero is of crucial importance, especially in the cases where denoising is performed in "open" communication channels, to avoid amplification of the effects due to their circulation in the acoustic channel.

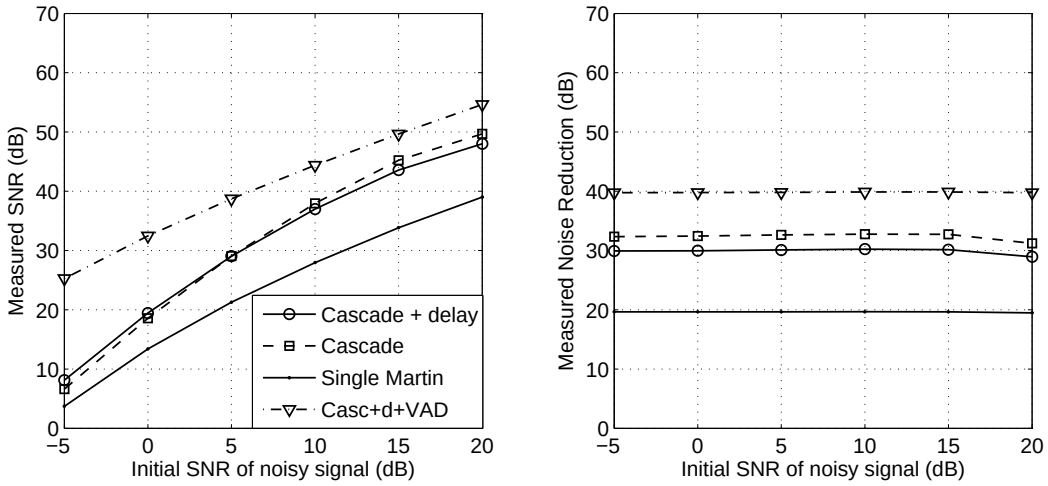

Figure 8.7: Performance of noise reduction schemes in terms of SNR and NR for various levels of vehicular noise.

The systems are compared in terms of SNR (dB) achieved after denoising and total noise reduction - NR (dB) within silence intervals. Voice quality is evaluated through Perceptual evaluation of speech quality (PESQ) (75). Fig. 8.7 to 8.12 summarize the results. The total noise reduction is computed as the average energy of noise over time, in the initial testing pattern, over the average energy of the filtered noise. SNR is estimated as the average energy of speech intervals minus the

average energy of speech absence intervals, over the average energy of speech absence intervals. For the case VAD is used, the SNR calculation the power of background noise present in speech intervals is also considered, measured within silence intervals prior the operation of VAD.

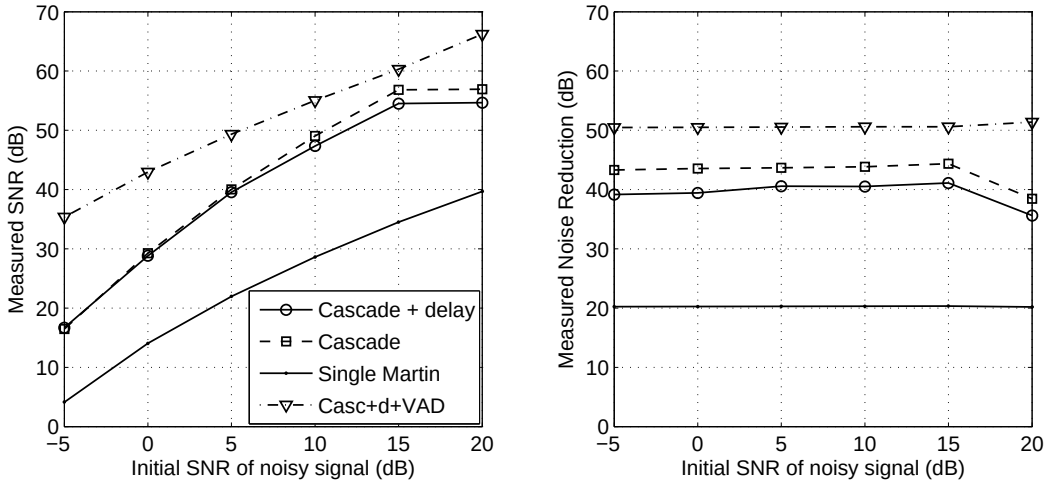

Figure 8.8: Performance of noise reduction schemes in terms of SNR and NR for various levels of buccaneer engine noise.

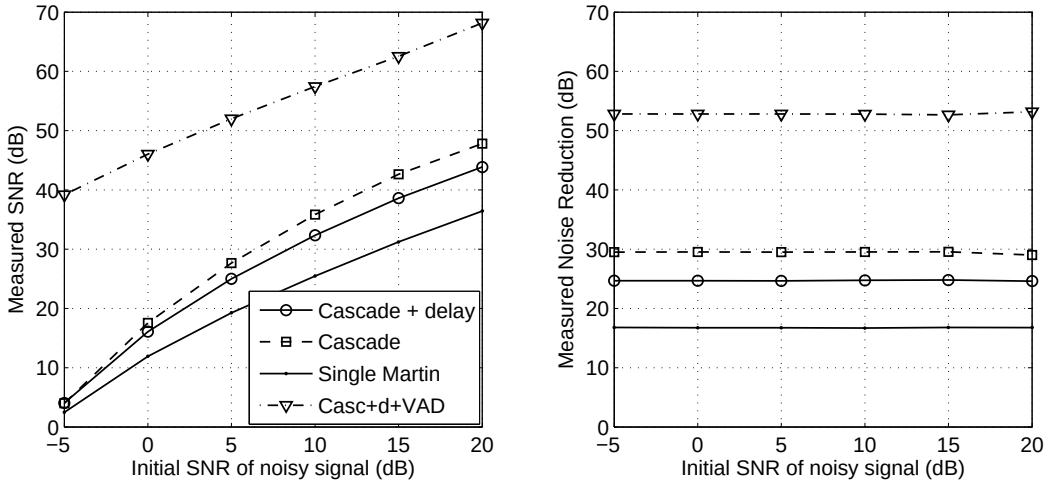

Figure 8.9: Performance of noise reduction schemes in terms of SNR and NR for various levels of m109 engine noise.

As depicted in the figures, the proposed systems achieve high rates of noise reduction, approximately 20-30dB lower than that of Martin's single stage noise reduction scheme. Furthermore, musical noise and residuals from the spectral subtraction process are drastically compressed when

employing the cascaded noise reduction schemes. The version that introduces and additional one step delay is more gentle treating voice (Fig. 8.13) although, the SNR achieved and overall noise reduction within silence intervals is slightly lower that the "Cascade" system.

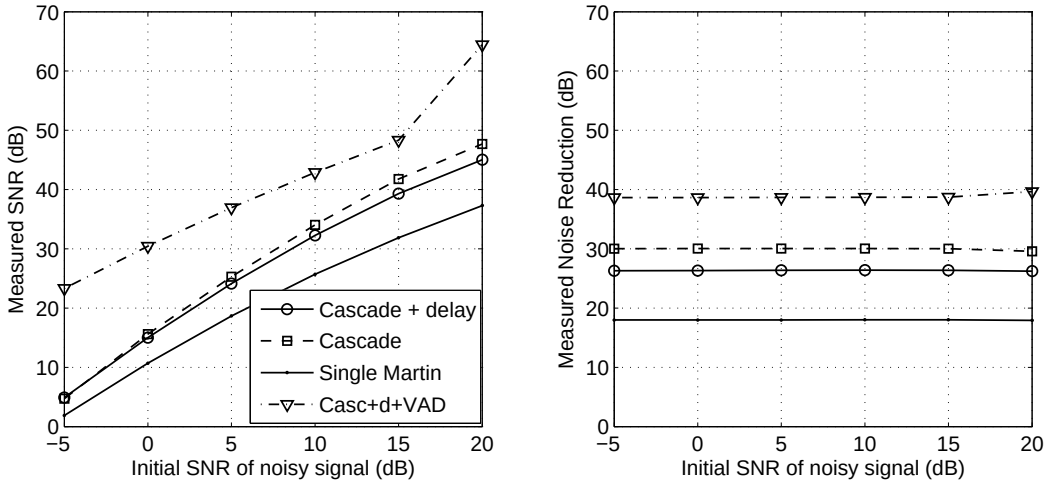

Figure 8.10: Performance of noise reduction schemes in terms of SNR and NR for various levels of destroyer engine noise.

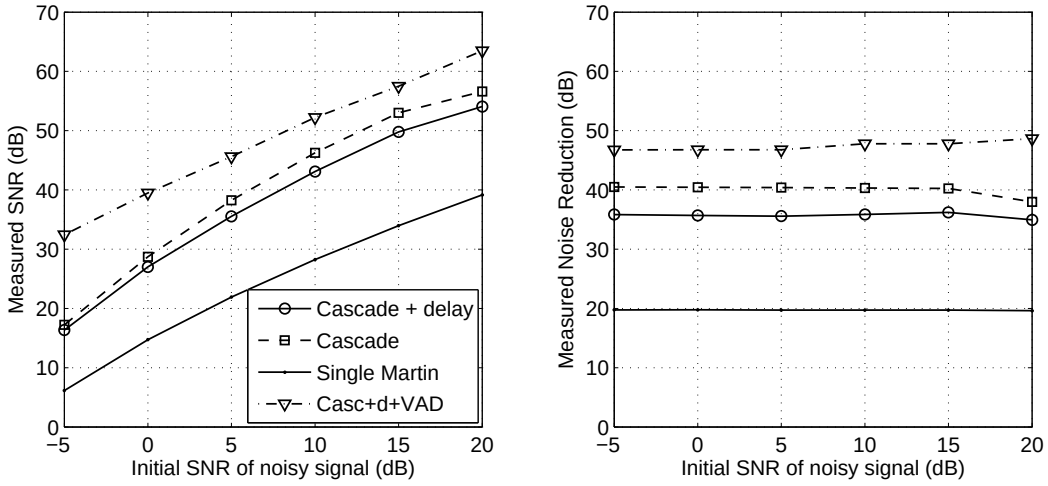

Figure 8.11: Performance of noise reduction schemes in terms of SNR and NR for various levels of m109 and machine gun synthetic noise.

The cascaded system employing the VAD performs remarkably better that the rest. The cascaded system with additional one step delay was combined with VAD due to its better performance in terms of speech quality. Nonetheless, one should expect that by encapsulating VAD in the output

of the system the overall noise reduction and achieved SNR would approach infinity, although this is not the case here. Due to non-speech detection error rate P_f (false alarms) the average power of the signal within silence intervals is not 0 and thus the SNR and NR rates are restricted. This is usually due to the hangover employed that introduces delay in state switching from "speech" to "non-speech" state allowing very small segments of noise get through the acoustic channel.

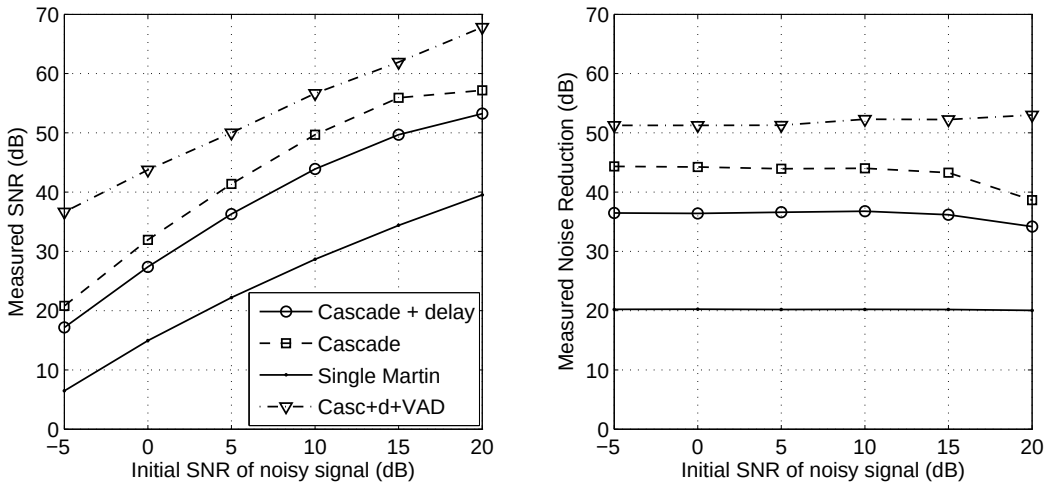

Figure 8.12: Performance of noise reduction schemes in terms of SNR and NR for various levels of F-16 cockpit noise.

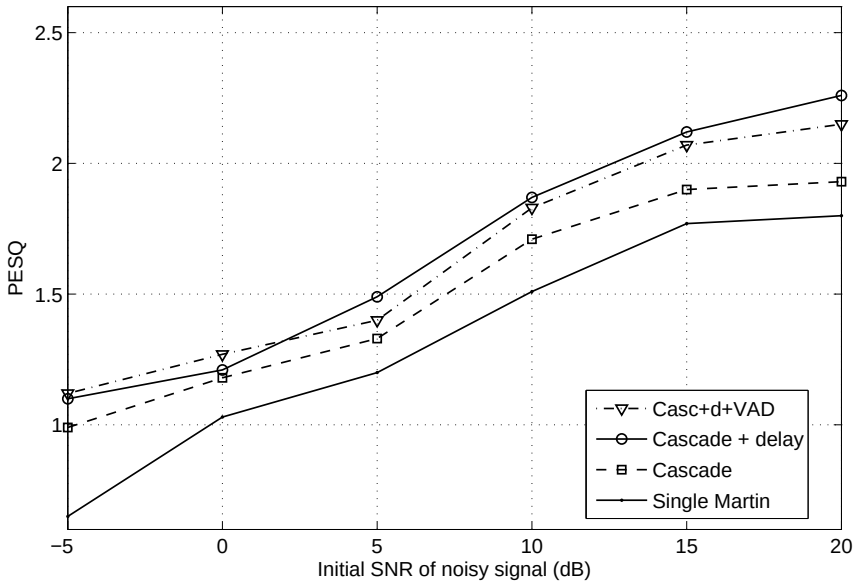

Figure 8.13: Average perceptual evaluation of speech quality (PESQ) for the full noise dataset at different noise intensities.

Figure 8.13 depicts the performance enhancement in terms of PESQ metric. In general the cascaded schemes proposed show significantly better performance especially for adverse conditions below 5dB SNR. The cascaded system with the extra delay performs better than the rest. When combining its operation with the VAD the performance slightly drops for SNR down to 5dB due to voice clipping errors. For the adverse conditions of 0dB and -5 dB there is a slight gain in performance due to suppression of residuals that appear within short pauses of speech between words.

In order to reduce overall delay, we attempted to reduce the frame and step sizes so that the noise subtraction is performed on shorter time intervals. Nevertheless, the transition from 256 frame length to 128 aiming to reduce overall delay resulted to a slight performance degradation, in terms of voice quality and musical noise artefacts.

8.5 Conclusions

In this chapter, one of the speech processing technologies the operation of which is usually combined with a VAD has been considered. More specifically, the performance benefits of combining VAD with noise reduction in noisy environments has been examined. Within this context a cascaded noise reduction framework has been proposed. The idea behind cascading noise reduction in to two stages is to initially subtract primary noise and then consequently use the same technique to remove spectral subtraction remaining artefacts and musical noise that are generated in the first subtraction stage. Two approaches have been followed. The first one, directly performs re-estimation of noise parameters on the power spectrum of the first stage denoised signal. The second one operates after the first stage denoised frames have been reconstructed to form the output in order to cater for the windowing effects. As depicted by the results under a large set of conditions of different types of noise and intensities the proposed architectures achieve high rates of noise reduction, approximately two times than that of Martin's single stage noise reduction scheme. The VAD presented in Chapter 3 has been employed to further suppress noise residuals acting as a musical noise reduction scheme showing significant performance enhancement within silence and short pause intervals.

Chapter 9

Indicating Apnea using VAD

9.1 Introduction

Chapter 9 deals with the detection of one of the major breathing-related sleep disorders, apnea. Apnea is considered to cause significant impact on patients health. Symptoms include disruption of oxygenation, snoring, choking sensations, apneic episodes, poor concentration, memory loss, and daytime somnolence. Diagnosis of apnea and breath disorders involves monitoring patients biosignals and breath during sleep in specialized clinics requiring expensive equipment and technical personnel. In this Chapter the design of a system capable for preliminary detection of sleep breath disorders at patients home utilizing patient sound signals will be described. The core of the system is based on a likelihood ratio test voice activity detector modified to detect snoring and heavy breathing events. The design of the employed VAD is actually a modification of convex combination scheme presented in Chapter 3. The basic feature of the proposed systems is the capability of unobtrusive monitoring of patients at home improving this way the reliable detection of sleep disorders in home environments offering comfort and time saving to patients.

9.2 Apnea Characteristics

Sleep is a basic human need in which there is a transient state of altered consciousness with perceptual disengagement from ones environment. Sleep Disordered Breathing (SDB) describes a group of disorders characterized by abnormalities of respiratory pattern or the quantity of ventilation during sleep. SDB causes disruptions in sleep, yielding waking somnolence, diminished neurocognitive performance, adverse cardiovascular outcomes, insulin resistance and other metabolic dysfunctions. One major sleep disorder is Obstructive Sleep Apnea (OSA), which is a sleep disorder characterized by pauses in breathing during sleep. It can occur due to complete or partial obstruction of the airway during sleep. Sleep Apnea is also known to cause loud snoring, oxyhemoglobin desaturations and frequent arousals (76). Each apnea episode lasts long enough so that one or more breaths

are missed, while such episodes occur repeatedly throughout sleep. The standard definition of an apneic event includes a minimum of 10 seconds interval between breaths, with either a neurological arousal, a blood oxygen desaturation of 3-4% or greater, or both arousal and desaturation. Clinically significant levels of sleep apnea are defined as five or more episodes per hour of any type of apnea. There are three distinct forms of sleep apnea: central, obstructive, and complex (i.e., a combination of central and obstructive) constituting 0.4%, 84% and 15% of cases respectively (76). Breathing is interrupted by the lack of respiratory effort in central sleep apnea. Regardless of type, the individual with sleep apnea is rarely aware of having difficulty breathing, even upon awakening.

Symptoms may be present for years (or even decades) without identification, during which time the sufferer may become conditioned to the daytime sleepiness and fatigue associated with significant levels of sleep disturbance. As a result, affected persons have unrestful sleep and excessive daytime sleepiness (76, 77). The disorder is also associated with hypertension impotence and emotional problems (77). Because obstructive sleep apnea often occurs in obese persons with comorbid conditions, its individual contribution to health problems is difficult to discern. The disorder has, however, been linked to angina, nocturnal cardiac arrhythmias myocardial infarction stroke and even motor vehicle crashes (78, 79, 80, 81, 82).

It is estimated that 20 million Americans are affected by sleep apnea. That would represent more than 6.5%, or nearly 1 in 15 Americans, making sleep apnea as prevalent as asthma or diabetes. It is also estimated that 85-90 percent of individuals affected are undiagnosed and untreated. The Wisconsin Sleep Cohort Study found that, among the middle-aged, nine percent of women and 24 percent of men had sleep apnea (83, 84). 2.500 patients in average per year are examined at sleep disorder centers in Greece and almost 80% of them are diagnosed with obstructive sleep apnea (85). The costs of untreated sleep apnea reach further than just health issues. It is estimated that the average untreated sleep apnea patient's health care costs $1,336 more than an individual without sleep apnea. If approximations are correct, 17 million untreated individuals account for $22,712 million, or almost 23 billion in health care costs (86). All the above facts prove the significance of sleep apnea as a medical problem and justify the research done in this field.

9.3 Common Methods for Apnea Detection

Polysomnography (PSG, see Fig. 9.1) is the most common method for diagnosing obstructive sleep apnea. In this technique, multiple physiologic parameters are measured while the patient sleeps in a laboratory. Typical parameters in a sleep study include eye movement observations (to detect rapid-eye-movement sleep), an electroencephalogram (to determine arousals from sleep), chest wall monitors (to document respiratory movements), nasal and oral airflow measurements,

an electrocardiogram, an electromyogram (to look for limb movements that cause arousals) and oximetry (to measure oxygen saturation). Apneic events can then be documented based on chest wall movement with no airflow and oxyhemoglobin desaturation. PSG requires special equipment of high cost to be installed and specialized personnel to be present. It offers limited resources for patient assessment (e.g., sleeping beds). In addition, elderly or sick patients often find the PSG equipment too cumbersome, and may be reluctant to spend the night in the sleep laboratory (87).

Figure 9.1: Patients being assessed for Obstructive Sleep Apnea (OSA) using Polysomnography equipment.

Additional methods to Polysomnography have been proposed in literature for sleep disorders detection or Apnea assessment. Mendez et al. present in (88) a method for screening OSA based on single ECG signals. Signal processing is used for the detection of RR intervals and QRS complexes and then the latter are classified using neural networks. The accuracy of the method in identifying patients with OSA is up to 88% according to authors. This method however requires from the patient to wear specific equipment and therefore cannot be characterized totally non-invasive. Furthermore the method relies on the existence of a training set of healthy patients and patients diagnosed with OSA. EEG arousal is utilized by Sugi et. al in (89) for sleep apnea syndrome detection. A sensitivity of 86% was achieved in successfully detecting apneic cases using this method. Still, the patient needs to be assessed in the Sleep Clinique wearing some uncomfortable

equipment. A body-fixed accelerometer sensor is used in (90) for acquiring vibration sounds during patients sleep. The latter technique is less invasive than PSG but still can cause discomfort to the patient and results can be easily biased by the sensor placement. In (91) Brunt et al. present a pneumatic bio-measurement method installed on patients bed for monitoring heartbeat, respiration, snoring and body movements. The latter achieves maximum patient comfort but still requires specialized hardware, a lot of data preprocessing and training and can only be used in Sleep Clinics.

Less invasive methods that have been used more extensively utilize sound processing of breath and snore sounds generated by patients during sleep. The feasibility of sleep apnea characterization through specific snore signal features has been proved in previously published studies (92, 93, 94, 95). Sound data acquisition is performed through microphones that are installed near patients beds at Sleep Clinic. For example the bed of the collaborating in this work Evagelismos Sleep Disorder Clinic is depicted in Fig. 9.1. Proper processing for noise removal, and feature extraction for further characterization of the snore as apneic or benign follow sound capturing in such sound analysis systems. Noise removal can be performed by applying adaptive cancellation filters (94), Linear Predictive Coding for speech removal (95), Kalman filtering (96) and Wavelet transformation (97).

All the aforementioned works that utilize snore signal processing for OSA characterization are based on microphone installations as already mentioned at Sleep Clinics. The proposed system is based on a mobile device that can be installed at patients home and can transmit snore sound data to the Sleeping Clinics remotely. Maximum patient comfort during sleep is achieved and a greater number of patients can be examined, resulting in better and faster prognosis of the sleep disorders. The following sections present technical details regarding the proposed system architecture, hardware specifications and the proposed snore sound analysis methodology.

9.4 Applying VAD for Apnea Indication

According to the clinical protocol, an apnea incident occurs when patient breath is interrupted for more than 6 seconds (77). In addition, the majority of the patients suffering from OSA, snore during sleep and present apneic events during the pause of snore events (76, 77, 78). Thus, in order to detect apnea during patients sleep from the acquired snore signal, snoring and breathing events have to be identified and quantified. Based on conducted experiments, when analyzing the captured snore sound signal, and applying short-term (i.e. frame lengths below 100ms) Discrete Fourier Transform (DFT), the distribution of real and imaginary parts of snore coefficients can be modeled by a two-sided Gamma distribution (TΓD) (23, 24), as illustrated in Fig. 9.2 and 9.3 as shown for for anechoic voice distribution in Section 2.3.

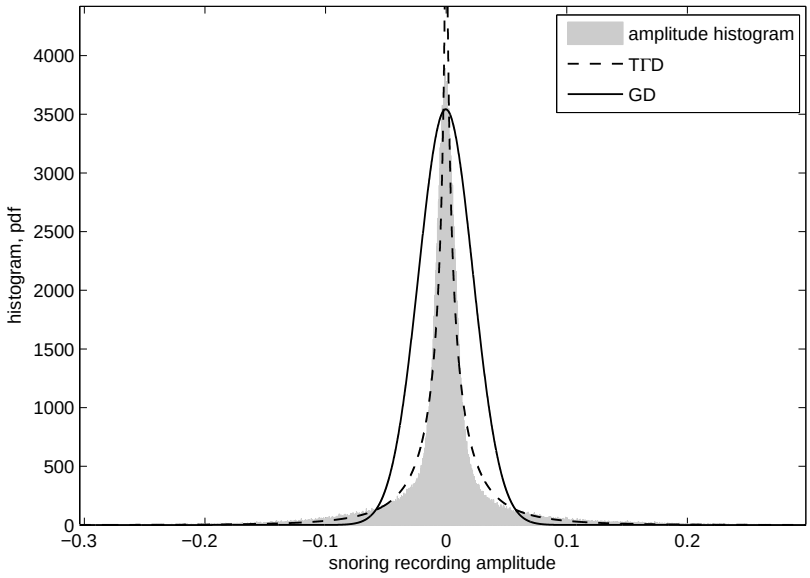

Figure 9.2: Snore amplitude distribution in time.

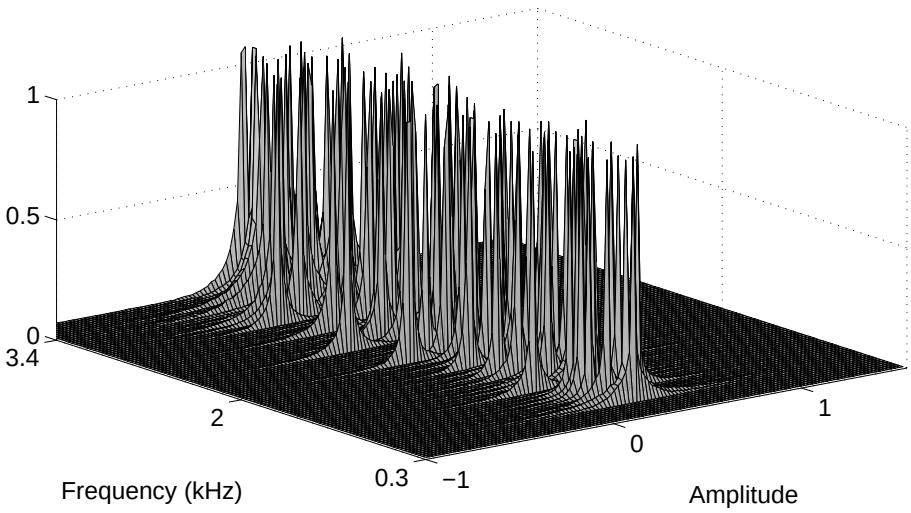

Figure 9.3: Snore amplitude distribution of frequencies. Histograms have been normalized to their maximum value per frequency bin.

The whole process results in the automated annotation of snore events. This way, silent periods between two sequential snores (i.e., the time patient does not breath or exhales) are quantified and depending on their duration, an apneic event can be detected. This way a preliminary assessment

can be provided to the experts. Subsection presents this method in detail.

9.4.1 Snore Hypothesis Testing

As in the case of voice, snore detection can be performed by evaluating the ratio of two distinct hypotheses, snore presence, and snore absence, denoted by H1 and H0 respectively. This approach is analogous to the evaluation of voice activity detection presented in Section 2.2.

$$H_0 : \text{snore absence} : \qquad \mathbf{X}(t) = \mathbf{N}(t) \tag{9.1}$$

$$H_1 : \text{snore presence} : \qquad \mathbf{X}(t) = \mathbf{S}(t) + \mathbf{N}(t) \tag{9.2}$$

where $\mathbf{X}(t) = [X_0(t), X_1(t), ..., X_{M-1}(t)]^T$, $\mathbf{S}(t) = \left[S_0(t), S_1(t), ..., S_{M-1}(t)\right]^T$, $\mathbf{N}(t) = [N_0(t), N_1(t), ..., N_{M-1}(t)]^T$ are the captured snore signal, source snore signal, and noise frequency components.

9.4.1.1 Probability Distribution of Noise

Both the real and the imaginary parts of noise frequency components are assumed to be zero mean following GD. The pdf of $\mathbf{N}_k(t)$ for the case of noise with k denoting the frequency bin is given by

$$f_n^G(N_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{N_k(t)^2}{2\sigma_{n,k}^2}} \tag{9.3}$$

where $\sigma_{n,k}^2$ is slowly varying with time variance factor of the Gaussian assumed distributed noise for the k^{th} frequency component. The imaginary part follows a similar distribution.

9.4.1.2 Probability Distribution of Snore Signal

As shown before in Figs(9.2, 9.3), it can be assumed that both the real and the imaginary parts of the frequency distribution of captured snore signal are better modeled using a TΓD

$$\text{TΓD} : f_s^{\Gamma}(S_k(t)) \frac{\sqrt[4]{3}}{2\sqrt{\pi\sigma_{s,k}}\sqrt[4]{2}} \left|S_k(t)\right|^{-\frac{1}{2}} e^{-\frac{\sqrt{3}|S_k(t)|}{\sqrt{2}\sigma_{s,k}}} \tag{9.4}$$

for the k^{th} frequency component, where $\sigma_{s,k}^2$ is the slowly varying variance factor.

9.4.1.3 Conditional Probability Density Functions

Using the predefined statistical model for snore and assuming Gaussian noise, the conditional pdfs of snore absence can be expressed as

$$H_0 : f_{X|H_0}(X_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{\mathbf{x}_k(t)^2}{2\sigma_{n,k}^2}} \tag{9.5}$$

The snore presence hypothesis is derived by

H_1 for TΓD snore signal model

$$H_1 : f_{X|H_1}(X_k(t)) = \int_{-\infty}^{\infty} \frac{\sqrt[4]{3}\,|S_k(t)|^{-\frac{1}{2}}}{4\pi\sqrt[4]{2}\sqrt{\sigma_{s,k}}\sigma_{n,k}} \times e^{-\frac{\sqrt{3}|S_k(t)|}{\sqrt{2}\sigma_{s,k}} - \frac{(X_k(t)-S_k(t))^2}{2\sigma_{n,k}^2}} dS_k \tag{9.6}$$

9.4.2 Snore Detection Likelihood Ratio Test

The likelihood ratio of those two conditional pdfs of snore presence and absence as proposed in Chapter 3, is defined as

$$\Lambda_k \equiv \frac{f_{X_k|H_1}(X_k)}{f_{X_k|H_0}(X_k)} = \frac{\int_{-\infty}^{\infty} \frac{\sqrt[4]{3}}{4\pi\sqrt[4]{2}\sqrt{\sigma_{s,k}}\sigma_{n,k}|S_k(t)|} e^{-\frac{\sqrt{3}|S_k(t)|}{\sqrt{2}\sigma_{s,k}} - \frac{(X_k(t)-S_k(t))^2}{2\sigma_{n,k}^2}} dS_k}{\frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{X_k(t)^2}{2\sigma_{n,k}^2}}} \tag{9.7}$$

where $f_{X_k|H_1}(X_k)$ is the hypothesis of snore presence H_1 and $f_{X_k|H_0}(X_k)$ is the hypothesis of snore absence H_0 under the assumption of Gaussian distributed noise. The decision criteria is based on evaluating the geometric mean of the likelihood ratio for the individual frequencies and is given by

$$\log \Lambda_k = \frac{1}{K} \sum_{k=0}^{K-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9.8}$$

whre η denotes the threshold of decision.

The values of snore and background noise power spectrum have to be continuously tracked. In this case the method of (14), namely *Predicted Estimation (PD)* has been employed as described in Section 3.5.

9.5 Threshold Evaluation

To overcome problems from the bias introduced by the LRT towards snore detection H_1 (16) we use a modified version of the one proposed in Section 3.6. The underlying concept behind this threshold is to track continuously the mean likelihood ratio value for the snore absent intervals. In

this direction a buffer N_{buf} holding past values of $\log \Lambda_k$ is employed. Initially, for the first 1-2 sec. of operation the system assumes snore absence. Given the first likelihood values the computation of the threshold is performed by

$$\hat{\eta}(t) \equiv \left(\overline{N_{buf}} + 3 \cdot \sigma_{N_{buf}}\right) \tag{9.9}$$

where $\sigma_{N_{buf}}$ and $\overline{N_{buf}}$ are the standard deviation and mean of the values in N_{buf}. The buffer is updated with new values only within snore absence intervals and if those are smaller than $4 \cdot \sigma_{N_{buf}}$. For smoothing the threshold estimate, a forgetting factor $\lambda_\eta = 0.98$ is introduced

$$\hat{\eta}(t+1) = \lambda_\eta \cdot \hat{\eta}(t) + (1 - \lambda_\eta)\, \hat{\eta}(t+1) \tag{9.10}$$

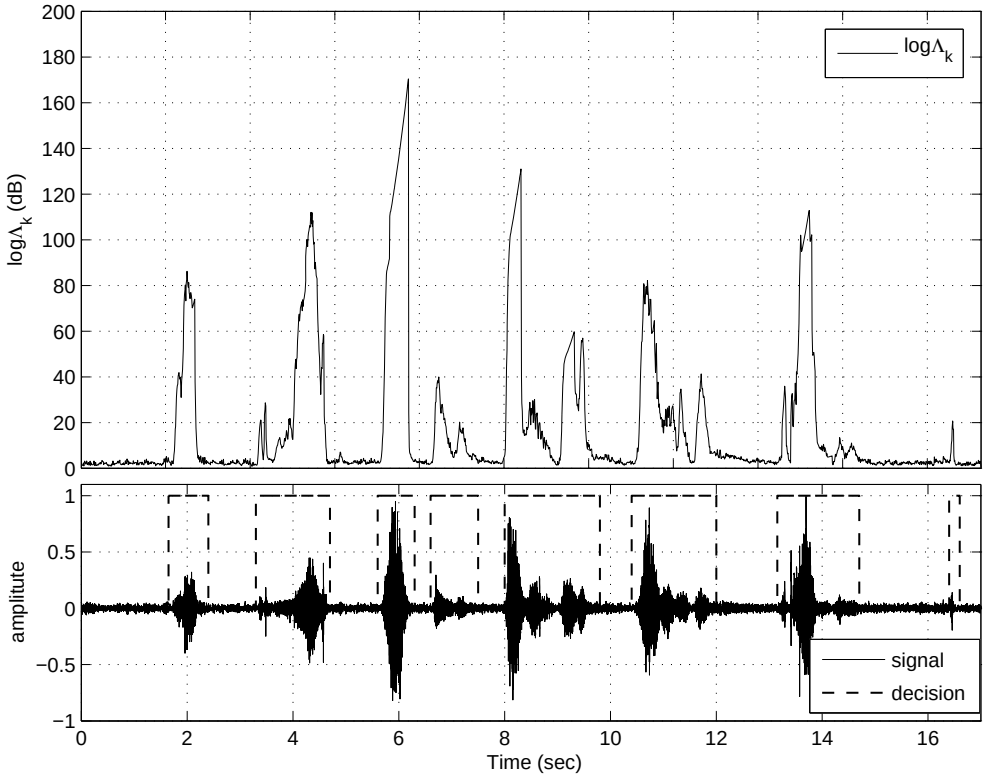

Figure 9.4: System response emerged for a sequence of snore events. Estimated geometric mean and snore presence/absence decision for the specific input.

9.5.1 Apnea Indication

Given the decision vectors, a snore absence / presence hangover state machine based on (15) that is able to track transitions from snore to silence intervals is defined. Time information elapsed

between phenomena of snore presence and absence can be stored into buffer and give the ability of indicating apnea condition.

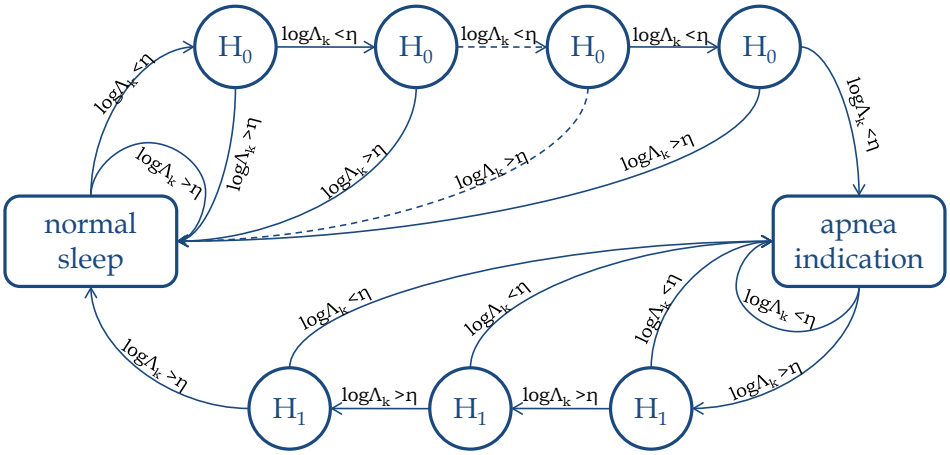

Figure 9.5: State Diagram of Sleep Breath Disorder Detection scheme.

The implementation of the hangover scheme as an apnea indicator is based on the idea that snores are highly correlated with time as generated with the function of breath. The hangover scheme is implemented as a state machine shown in Fig 9.5. Parameters H_1 and H_0 indicate snore presence and absence respectively, being triggered by the value of $\log \Lambda_k$. If the value of the geometric mean $\log \Lambda_k$ is greater than or equal to the threshold the snore event is detected otherwise snore absence is assumed. This slightly biases the system towards snore detection. Thus the value of $\log \Lambda_k$ is then used to determine which state H_1 or H_0 the machine should be in. As mentioned before an apnea incident occurs when patient breath is interrupted between snores for more than 6 seconds. Given that a new $\log \Lambda_k$ emerges every 20msec, a number of 50 consecutive snore absence detections should emerge by the system to indicate an 'apnea' incident. A set of 5 (100msec) consecutive snore presence indications are required to reset the state to 'normal' breathing.

Following the transitions in Fig. 9.5 let's assume that initially the system is at the 'normal' breathing state due to past sequential snore detection events. The value of $\log \Lambda_k$ is employed to determine how the hangover scheme should proceed. If it gets below the threshold η the state machine begins to progress through the transition states toward the 'apnea' state. At this point the incident of 'apnea' is not definite as the lower value might be a false by the snore detection algorithm not being able to detect a snore event. After 50 consecutive indications of snore absence is the hangover scheme will enter the 'apnea' state. The chain will remain in that state unless $\log \Lambda_k$ becomes greater that the threshold.

When this event occurs, the hangover scheme will begin to progress through transition states towards the 'normal' breath state. This done due to the uncertainty of snore presence indication which might be a false alarm. After five consecutive snore indications, the hangover scheme will return to the 'normal' state and wait till the value of $\log \Lambda_k$ drops below the threshold again.

9.6 Performance Evaluation

In order to evaluate the proposed technique for sleep sound analysis, a number of 9 recordings of individuals (6 males and 3 females) have been collected at Euagelismos Sleep Clinic, Medical School in University of Athens (85). Each sound sample corresponds to a complete sleep study (duration up to 6 hours) of patients that either suffered from sleep apnea or were examined for symptoms of sleep breath disorders. Snore sound events and apnea have been manually annotated by the Sleep Clinic experts. The experiments were conducted in three different room conditions. The first scenario involved recordings in a typical home environment during night with no apparent noise; the second was performed with additional noise from operating air-conditioning system and the third with additional urban noise from an open window. The noise intensity in the three recording scenarios has been restricted close to 15dB, since the recording environment had to be comfort for the patient. A recording sampling rate of 22.05kHz has been used. The evaluation metrics employed are:

- TP: True Positives (Percentage of correct indications of snore presence)
- FN: False Negatives (Percentage of incorrectly rejected snore events).
- TN: True Negatives (Percentage of correct indications of silence).
- FP: False Positives (Percentage of misclassified silence intervals).
- SENS: Sensitivity (=TP/(TP+FN)) (Tendency of system towards snore detection).
- SPCF: Specificity (=TN/(TN+FP)) (Tendency of system towards silence).

Table 9.1: Snore Detection Performance Results

Noise	SNR	TP%	FP%	FN%	TN%	SENS%	SPCF%
Clinic Sleep Room	15.30dB	98.31	1.69	6.18	93.82	94.02	98.2
Aircondition On	14.41dB	97.36	2.64	3.79	96.21	96.2	97.3
Open Window	12.85dB	92.07	7.93	4.81	95.19	95.0	92.3

Table 9.2: Apnea Indication Performance Results

Noise	SNR	TP%	FP%	FN%	TN%	SENS%	SPCF%
Clinic Sleep Room	15.30dB	79.03	20.97	4.08	95.92	95.09	82.06
Aircondition On	14.41dB	81.44	18.56	2.91	97.09	96.55	83.95
Open Window	12.85dB	76.62	23.38	3.82	96.18	95.25	80.44

Very few alternative methods for the classification of patients through snoring information have been published to date, like the one described in (98). Using a model based on the snoring pitch, a sensitivity of 91% at a specificity of 67% is obtained in the detection of Sleep Apnea patients.

Another method for the detection of Sleep Apnea patients based on the transient fluctuations of a logarithmic average of the respiratory sound intensity was recently validated (99). Subjects were classified with sensitivity of 93% and specificity 67%.

The method we proposed for snore detection performs similarly or better than those systems according to the results in Table 9.1 being at the same time computationally lightweight. When it comes to apnea indication, the simplicity of the hangover scheme we employed doesn't allow for optimum performance as FN had to be retained as low as possible so that the possibility that a healthy subject is indicated to suffer from apneic events is very low (Table 9.2). Nevertheless, given that apneic events occur multiple times within the time the patient is sleeping, it is very unlikely failing to indicate that a patient suffers from apnea. Nonetheless, any comparison must be taken with caution until a validation on a greater database is available.

9.7 Conclusions

In this Chapter a non-invasive system for automated Sleep Apnea detection utilizing snore sound analysis has been presented. Despite the fact that Obstructive Sleep Apnea is not widely known, it is a very common health issue with high potential implications and effects on patients health. The most common assessment method involves the overnight physiological sign monitoring of the patient in Sleep Clinics, and requires specific equipment and specialized personnel. Most widely used diagnosis technique of sleep breath disorder events rely completely on the manual scoring of physiological data by specialists, which is time consuming, costly and not readily available as well. Snore signals along with apneic events are captured by microphone. The core of the processing algorithm employed is based on the VAD developed in Chapter 3 with the scope of creating a computationaly lightweight system that could be potentially used to monitor patients at home improving this way the prognosis and treatment procedure and offering the maximum comfort to patients at same time. The system can only be utilized as a preliminary remote assessment method in case patients present both OSA and snoring. However, since snoring events are highly

related to OSA (100) the system can act as an indicator for further assessment of the patient using the standard PSG techniques. The indication of apnea is based on the fact long pauses in breaths or snores during sleep can indicate an apneic event according to the clinical protocol. A hangover state machine has been employed to continiously track transitions from snore to silence event indicating apneic events based on state transition times. Conducted experiments, using the proposed VAD architecture for detecting snore detection, under various conditions, have indicated significant accuracy in detecting snoring against background noise.

Chapter 10

Concluding Remarks

The performance of a VAD, like for most of speech processing system, is significantly downgraded when far-field (FF) microphones are used instead of the conventional close-talking (CT), due to reverberation effects, competitive sound sources, and speaker movement that can significantly alter the statistical characteristics of captured speech. The scope of this work was the development of robust VAD systems and algorithms, able to operate with one or more FF microphones within adverse environments. Thus, VAD systems had to be designed in such way so that they are able to cater for the varying energy of speech signals captured with FF sensors, reflections and highly-intensive interfering noises. The algorithms developed, as part of this the work, were designed in a real-time frame-by-frame processing basis to allow for their integration in modern telecommunication systems, smart rooms and other technologies.

In the first part of this work, speech distribution variability under external interferences was investigated. This study formed the basis for the development and design of an unsupervised VAD based in the convex combination of a set of prime distributions able to robustly operate within adverse conditions. The work continued with the design of a multi-microphone VAD that encapsulated spatial, apart from frequency and time, information. The multi-microphone VAD was later used in combination with a powerful data analysis method towards achieving optimal performance. Speech distribution information was also encapsulated in a supervised VAD scheme employing Hidden Markov Models the states of which are modelled using Gaussian Mixture Models to cater for the dynamics of captured speech. Additionally, a visual-VAD was developed and fused with audio one. Additionally, the effect of the developed VAD algorithms on other audio signal processing application fields was examined. Those included speaker tracking, noise reduction and acoustic event detection. Moreover, we worked on the involvement of those audio signal processing technologies by encapsulating observations, outcomes and techniques that demonstrated better performance for VAD. We now summarize the key results of our work and discuss open issues.

10.1 Summary of Main Results

The basis for this work was an extensive study on the effects of noise and reverberation on speech distribution at various intensity levels and conditions. We demonstrated that when FF microphones are used, competitive sound sources, and speaker movement can significantly alter the distribution of captured speech. In contrary to previous studies, we depicted that captured speech with FF microphones is not solely GD, LD, or ΓD distributed, given its non-stationarity in time and its dependence on external interferences. This way we justified why speech processing systems relying solely on the Gaussian or other fixed assumptions are expected not to perform adequately under varying conditions. Fixed distribution assumptions can be accurate only under specific conditions of reverberation and noise. Those outcomes, actually directed the whole research effort into the development dynamically adaptive systems able to overcome such environmental adversities and speech dynamics.

Moving a step further, a highly efficient statistical voice activity detector was developed, which relies on the modelling of the distribution of speech as a convex combination of a Gaussian, a Laplacian, and a two-sided Gamma distribution discussed in Chapter 3. The decision criterion of the proposed algorithm is the weighted sum of three likelihood ratio tests, each one corresponding to one of the fundamental core distributions. The computation of the corresponding weights has been based on the statistical distances of the instantaneous input samples from the Gaussian, the Laplacian, and the two-sided Gamma distribution, estimated using the Kolmogorov-Smirnov test. Experiments performed using artificially reverberated and contaminated with additive noise anechoic audio data revealed that the specific voice activity detector outperforms the existing systems in terms of error rate and that it produces reliable results even under adverse noise conditions and reverberation effects. The result justified our initial hypothesis, that speech distribution can be better modelled as linear combination of a set of primal distributions rather than any other single distribution approach.

In the next step, in Chapter 4, we considered the encapsulation of spatial information, embedded in signals captured by far-field microphone arrays, in a VAD scheme. The developed scheme is taking advantage of the spatial information provided by multiple sensors without the need of knowledge of direction-of-arrival estimates like previous approaches. Simulations performed demonstrated that the proposed system remains more robust than a set of related counterparts without imposing additional delay in the system or being subject to reverberation.

The multi-microphone VAD served as the platform to merge VAD with a very powerful analysis framework namely Empirical Mode Decomposition (EMD), presented in Chapter 5. This highly efficient signal decomposition method significantly enhanced the performance of VAD acting as a

speech enhancement technique prior to voice detection. The outcome of this procedure demonstrates significantly enhanced performance compared to single microphone approaches.

In the area of supervised voice activity detection, a system based on the modelling capabilities of hidden Markov models has been developed in Chapter 6 . Gaussian Mixtures modelling has been employed per model state to cater for the variable distribution of speech in respect to the outcomes of the research on speech distribution. Given the bi-modality of speech generation process, conveying both audio and visual information, an Audio-Visual VAD that combines the advantages of both modalities has been also considered. Although the developed system wasn't based on two optimal modalities, fusing of different VAD schemes showed that there is a noticeable increment in performance even under extremely adverse conditions.

Following the study plan we then concentrated on exploring applications of VAD. In Chapter 7 the performance benefits of combining the developed multi-microphone VAD with a direction-of-arrival (DOA) estimation scheme were demonstrated on the basis that speech emission is a discontinuous sound source. The employed DOA system was based on information theoretical TDE system. The optimization of this TDE scheme was also considered in the context of encapsulating speech shaped distributions in the underlying assumption of the speech model employed. Thus, we investigated how the performance of a robust information-theoretical TDE algorithm, changes as we switch between different underlying assumptions for the distribution of speech in respect to the instant input. The analysis performed, revealed a significant research outcome. The employed marginal MI criterion based TDE is not depended on the underlying assumption of the distribution of speech when that belongs to the family of Generalized Gaussian distribution, exploiting the invariance property of MI. To support the analysis, closed forms of the multivariate and univariate differential entropies for the Generalized Gaussian distribution were derived, that encapsulate the entropies of other well known distributions like Gaussian, Laplacian and Gamma.

Additionally, in Chapter 8, performance enhancement when using VAD in combination with noise reduction systems has been also documented in terms of residual suppression within silence intervals. For this scope an efficient noise reduction architecture has been developed based on cascading an one-pass scheme.

Finally, we steered our focus in acoustic event detection field. More specifically, an automated sleep apnea detection system utilizing snore sound analysis has been presented. The core of the processing algorithm employed was based partially on the convex combination of multiple statistical models VAD aiming to a computationally lightweight system that could be potentially used to monitor patients at home. This way prognosis, treatment procedure and offering the maximum comfort to patients is improved. Conducted experiments using the system in various conditions

have indicated increased accuracy in detecting snoring against background noise and indication of apneic events compared to other obtrusive methodologies.

10.2 Discussion and Future Directions

VAD accuracy is of paramount importance and plays a critical role in enhancing the performance of other speech processing systems. Thus, searching for methods to improve accuracy of VAD remains a highly tempting field. In this thesis we investigated the parameters that affect VAD performance towards designing better systems able to operate within adverse conditions.

Throughout the individual sections of this work some interesting research directions arose. The statistical VAD which relies on the modelling of the distribution of speech as a convex combination of a Gaussian, a Laplacian, and a two-sided Gamma distribution discussed in Chapter 3, outperformed existing systems in terms of error rate even under adverse noise conditions and reverberation effects. Although, in the case of intensive car noise (SNR<0 dB), which is Laplacian distributed, the scheme did not perform as expected. This is due to the assumption of Gaussian distributed noise. Thus, evaluating the combination of the three prime distributions used to model speech with other than Gaussian distribution for noise would be very interesting. The convex likelihood ratio would have to be modified so that it also encapsulates a combination scheme for noise distribution.

Encapsulating the specific convex combination scheme in the multi-microphone VAD approach, presented in Chapter 4, is indeed interesting to see in the direction of performance optimisation. Combining such an efficient modelling scheme with spatial information is expected to result in significantly increased performance.

Another important topic opens through combining the multi-microphone VAD with Empirical Mode Decomposition, presented in Chapter 5. This highly efficient signal decomposition method significantly enhanced the performance of VAD acting as a speech enhancement technique prior to voice detection. Although, the specific decomposition methodology requires significantly increased computation time in order for the signal to be decomposed into the corresponding intrinsic mode functions (IMF). Thus, it would be of great importance to investigate methods to boost execution times of EMD for VAD. Additionally an alternative adaptive way to the heuristic method of selecting the speech information bearing IMFs could be considered.

Future work in the field of supervised VAD described in Chapter 6 could explore more advanced facial feature extraction algorithms such as active appearance models aiming to the design of a more robust Visual VAD. Other decision fusing techniques could also be evaluated in the context of combining the two modalities of speech production.

Furthermore, in the field of time delay estimation, presented in conjunction with VAD in Chapter 7, optimisation in the context of encapsulating speech shaped distributions in the underlying assumption of the speech model employed can be reconsidered for different TDE methodologies. The analysis performed, as part of this work, revealed that the employed marginal MI criterion based TDE is not depended on the underlying assumption of the distribution of speech when that belongs to the family of Generalized Gaussian distribution, exploiting the invariance property of MI. Although, this property stands only for MI and thus it is of high interest to encapsulate speech shaped distribution in TDE for other methodologies for which this approach would be beneficial.

Additionally, in the field of noise reduction investigeted in Chapter 8, it would be very interesting working on improving the embedded VAD schemes that are responsible of accurately updating the estimation for the spectrum of noise and speech in many denoising systems. This would lead in more efficient noise reduction systems eliminating musical noise at the same time on the context of a more accurate noise spectrum estimation.

Finally, the apnea indication methodology presented in Chapter 9 could be further improved by developing more efficient classification techniques to detect the apneic events. Given the the system doesn't need to operate in real-time, advanced classification techniques such as deep-belief networks and support vector machines could be employed in the direction of increasing detection accuracy.

Bibliography

[1] H. Othman and T. Aboulnasr, "A semi-continuous state-transition probability HMM-based voice activity detector," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 1, pp. 821–824, 2007. 2

[2] D. Freeman, G. Gosier, C. Southcott, and I. Boyd, "The Voice Activity Detector for the Pan-european digital cellular mobile telephone service," *Proceedings of the IEEE, ICASSP*, vol. 1, pp. 369–372, May 1989. 2

[3] D. Vlaj, B. Kotnik, B. Horvat, and Z. Kacic, "A computationally Efficient Mel-filter bank vad algorithm for Distributed Speech Recognition Systems," *EURASIP Journal on Applied Signal Processing*, pp. 487–497, April 2005. 2

[4] A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, and J. 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," *Communications Magazine, IEEE*, vol. 35, no. 9, pp. 64–73, 1997. 2, 88

[5] R. Gemello, F. Mana, and R. Mori, "Non-linear estimation of voice activity to improve automatic recognition of noisy speech," in *Ninth European Conference on Speech Communication and Technology*, 2005. 2

[6] S. Ravindran, D. Anderson, and M. Slaney, "Low-power Audio Classification for Ubiquitous Sensor networks," *ICASSP Proceedings IEEE*, vol. 4, pp. 337–340, May 2004. 2

[7] R. Mangharam, A. Rowe, R. Rajkumar, and R. Suzuki, "Voice over Sensor Networks," *Real-Time Systems Symposium IEEE*, pp. 291–302, December 2006. 2

[8] K. Sakhnov, E. Verteletskaya, and B. Simak, "Dynamical energy-based speech/silence detector for speech enhancement applications," in *Proceedings of the World Congress on Engineering*, vol. 1, 2009. 2, 43

[9] R. Tucker, "Voice activity detection using a periodicity measure," in *Communications, Speech and Vision, IEE Proceedings I*, vol. 139, pp. 377–380, IET, 1992. 2

[10] E. Nemer, R. Gourban, and S. Mahmoud, "Robust Voice Activity Detection using higher-order in the LPC residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 217–231, 2001. 2

[11] S. Yang, Z. Li, and Y. Chen, "A fractal based voice activity detector for Internet telephone," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 1, pp. 808–811, April 2003. 2

[12] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 109–118, 2002. 2

[13] F. Talantzis and A. Constantinides, "Using information theory to detect voice activity," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009, IEEE International Conference on*, pp. 4613–4616, IEEE, 2009. 2

[14] J. Chang, N. Kim, and S. Mitra, "Voice activity detection based on multiple statistical models," *Signal Processing, IEEE Transactions on*, vol. 54, no. 6, pp. 1965–1976, 2006. 2, 3, 34, 39, 40, 41, 44, 46, 48, 49, 50, 57, 70, 131

[15] A. Davis and R. Togneri, "Statistical Voice Activity Detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 412–424, March 2006. 2, 43, 79, 80, 83, 132

[16] J. Shin, H. Kwon, S. Jin, and N. Kim, "Voice activity detection based on conditional MAP criterion," *Signal Processing Letters, IEEE*, vol. 15, pp. 257–260, 2008. 2, 42, 131

[17] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999. 2, 32, 49, 56

[18] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian - Gaussian model," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 498–505, 2003. 2, 46, 47

[19] H. Kim and S. Park, "Voice activity detection algorithm using radial basis function network," 2

[20] J. Shin, J. Chang, and N. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language, Elsevier*, vol. 24, no. 3, pp. 515–530, 2010. 2

[21] E. Rentzeperis, C. Boukis, A. Pnevmatikakis, and L. Polymenakos, "Combining finite state machines and LDA for voice activity detection," *Artificial Intelligence and Innovations 2007: from Theory to Applications, Springer*, pp. 323–329, 2007. 3

[22] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/Fuzzy voice activity detectors," *Signal Processing Letters, IEEE*, vol. 9, pp. 85 –88, march 2002. 3

[23] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. 249–253, IEEE, 2002. 13, 15, 34, 94, 128

[24] S. Gazor and W. Zhang, "Speech probability distribution," *Signal Processing Letters, IEEE*, vol. 10, no. 7, pp. 204–207, 2003. 13, 15, 33, 34, 94, 128

[25] Electronics Department, "Aalborg University Small Anechoic Room." `http://doc.es.aau.dk/labs/acoustics/facilities/anechoicroomsmall/`, March 2012. 15, 45

[26] A. Glen, L. Leemis, and D. Barr, "Order statistics in goodness-of-fit testing," *Reliability, IEEE Transactions on*, vol. 50, no. 2, pp. 209–213, 2001. 18

[27] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustical Society of America, JASA*, vol. 65, no. 4, pp. 943–950, 1979. 21, 45, 51, 59, 71, 106, 119

[28] H. Kuttruff, *Room Acoustics*. Elsevier Applied Science, 1990. 21

[29] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984. 32, 40, 57, 112

[30] J. Shin, J. Chang, H. Yun, and N. Kim, "Voice activity detection based on generalized gamma distribution," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 781–784, 2005. 33, 44, 45, 46, 47, 48, 49

[31] J. Shin, J. Chang, and N. Kim, "Statistical modeling of speech signals based on generalized gamma distribution," *Signal Processing Letters, IEEE*, vol. 12, no. 3, pp. 258–261, 2005. 33

[32] G. Almpanidis and C. Kotropoulos, "Voice activity detection with generalized gamma distribution," in *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 961–964, 2006. 33

[33] A. Varga and H. 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, Elsevier*, vol. 12, no. 3, pp. 247–251, 1993. 45, 59, 71, 115, 118

[34] A. Sehr, E. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010. 51

[35] I. Potamitis and E. Fishler, "Speech activity detection and enhancement of a moving speaker based on the wideband generalized likelihood ratio and microphone arrays," *Journal of the Acoustical Society of America, JASA*, vol. 116, p. 2406, 2004. 55

[36] K. Ishizuka, S. Araki, and T. Kawahara, "Statistical speech activity detection based on spatial power distribution for analyses of poster presentations," in *INTERSPEECH'08*, pp. 99–102, 2008. 55

[37] J. Ramírez, J. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *Signal Processing Letters, IEEE*, vol. 12, no. 10, pp. 689–692, 2005. 55, 57

[38] National Institute of Standards and Technology, "The NIST Mark-III microphone array." `http://www.nist.gov/smartspace/mk3_presentation.html`, August 2012. 58

[39] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The Empirical Mode Decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998. 65, 66, 67, 73

[40] X. Tan, J. Gu, H. Zhao, and Z. Tao, "A noise robust endpoint detection algorithm for whispered speech based on Empirical Mode Decomposition and entropy," in *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on*, pp. 355–359, IEEE, 2010. 66, 71, 73

[41] Z. Lu, B. Liu, and L. Shen, "Speech endpoint detection in strong noisy environment based on the Hilbert-Huang transform," in *Mechatronics and Automation, 2009. ICMA 2009. International Conference on*, pp. 4322–4326, IEEE, 2009. 66, 71

[42] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. 80, 81, 82

[43] A. Stergiou, A. Pnevmatikakis, and L. C. Polymenakos, "Enhancing the performance of a GMM-based speaker identification system in a multi-microphone setup," in *CLEAR '07 Eval. Camp. and Workshop - Class. of Events, Act. and Relat.*, pp. –1–1, 2007. 81

[44] S. Lloyd, "Least squares quantization in PCM," *Information Theory, IEEE Transactions on*, vol. 28, no. 2, pp. 129–137, 1982. 81

[45] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977. 81

[46] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Communication, Elsevier*, vol. 49, no. 7-8, pp. 667–677, 2007. 84

[47] P. A. Viola and M. J. Jones, "Rapid object detection using a Boosted Cascade of simple features," in *IEEE Comp. Soc. Conf. on Computer Vision and Pat. Recog. CVPR (1)'01*, pp. 511–518, 2001. 84

[48] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Systems, Man and Cybernetics*, vol. 9, pp. 62–66, 1979. 84

[49] A. Adjoudani and C. Benoit, "Audio-visual speech recognition compared across two architectures," in *Fourth European Conference on Speech Communication and Technology*, 1995. 86

[50] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998. 86

[51] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Proc. EUSIPCO*, 2008. 86

[52] F. Talantzis, A. Constantinides, and L. Polymenakos, "Estimation of direction of arrival using information theory," *Signal Processing Letters, IEEE*, vol. 12, pp. 561 – 564, aug. 2005. 93, 94, 95, 98

[53] F. Talantzis, "An acoustic source localization and tracking framework using particle filtering and information theory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1806–1817, 2010. 93

[54] J. Li and R. Wu, "An efficient algorithm for time delay estimation," *Signal Processing, IEEE Transactions on*, vol. 46, no. 8, pp. 2231–2235, 1998. 93

[55] G. Carter, "Time delay estimation for passive sonar signal processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, pp. 463 – 470, June 1981. 93

[56] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, pp. 320 – 327, aug 1976. 94, 95

[57] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Verlag, 2001. 94, 95

[58] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 549 – 557, nov. 2003. 94

[59] J. Benesty, Y. Huang, and J. Chen, "Time Delay Estimation via Minimum Entropy," *Signal Processing Letters, IEEE*, vol. 14, pp. 157 –160, march 2007. 94, 95, 99, 100, 102, 105

[60] F. Wen and Q. Wan, "Robust time delay estimation for speech signals using information theory: A comparison study," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, p. 3, 2011. 95, 100, 105

[61] F. Talantzis, A. Pnevmatikakis, and A. Constantinides, *Audio-Visual Person Tracking: A Practical Approach*, vol. 1. World Scientific Publication Company, Imperial College Press, 2011. 95

[62] T. Cover, J. Thomas, J. Wiley, *et al.*, *Elements of information theory*, vol. 6. Wiley Online Library, 1991. 95, 97, 98

[63] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, pp. 994–1008, 2009. 99

[64] G. Verdoolaege and P. Scheunders, "On the geometry of multivariate generalized Gaussian models," *Journal of Mathematical Imaging and Vision, Springer*, pp. 1–14, 2011. 101, 102

[65] K. Fang and Y. Zhang, *Generalized multivariate analysis*. Science Press Beijing, 1990. 102

[66] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E, APS*, vol. 69, no. 6, p. 066138, 2004. 105

[67] T. Hager, S. Araki, K. Ishizuka, M. Fujimoto, T. Nakatani, and S. Makino, "Handling speaker position changes in a meeting diarization system by combining doa clustering and speaker identification," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC-2008)*, 2008. 106

[68] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "Speaker indexing and speech enhancement in real meetings/conversations," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 93–96, IEEE, 2008. 106

[69] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, pp. 29–32, IEEE, 2008. 106

[70] M. Swartling, B. Sallberg, and N. Grbic, "Direction of arrival estimation for speech sources using fourth order cross cumulants," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pp. 1696–1699, IEEE, 2008. 106

[71] R. Martin, "Spectral subtraction based on minimum statistics," *EUSIPCO-94*, vol. Edinburgh, Scotland, pp. 1182–1185, September 1994. 111, 112, 114, 115, 118

[72] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, 2001. 111, 112, 114

[73] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, 2006. 111, 112

[74] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, vol. 4, pp. 208–211, IEEE, 1979. 112, 114

[75] I. R. P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," 2000. 119

[76] J. Shepard Jr, "Cardiopulmonary consequences of obstructive sleep apnea.," in *Mayo Clinic proceedings. Mayo Clinic*, vol. 65, p. 1250, 1990. 125, 126, 128

[77] A. Kales, A. Caldwell, R. Cadieux, A. Vela-Bueno, L. Ruch, and S. Mayes, "Severe obstructive sleep apnea–II: Associated psychopathology and psychosocial consequences," *Journal of chronic diseases, Elsevier*, vol. 38, no. 5, pp. 427–434, 1985. 126, 128

[78] K. Wei and T. Bradley, "Association of obstructive sleep apnea and nocturnal angina," *Am Rev Respir Dis*, vol. 145, no. 4 pt 2, p. A443, 1992. 126, 128

[79] C. Guilleminault, S. Connolly, and R. Winkle, "Cardiac arrhythmia and conduction disturbances during sleep in 400 patients with sleep apnea syndrome," *The American journal of cardiology, Elsevier*, vol. 52, no. 5, pp. 490–494, 1983. 126

[80] J. Hung, E. Whitford, D. Hillman, and R. Parsons, "Association of sleep apnea with myocardial infarction in men," *The Lancet, Elsevier*, vol. 336, no. 8710, pp. 261–264, 1990. 126

[81] M. Partinen and C. Guilleminault, "Daytime sleepiness and vascular morbidity at seven-year follow-up in obstructive sleep apnea patients.," *Chest, American College of Chest Physicians*, vol. 97, no. 1, pp. 27–32, 1990. 126

[82] M. Aldrich, "Automobile accidents in patients with sleep disorders.," *Sleep: Journal of Sleep Research & Sleep Medicine, American Academy of Sleep Medicine*, 1989. 126

[83] T. Young, P. Peppard, and D. Gottlieb, "Epidemiology of obstructive sleep apnea," *American Journal of Respiratory and Critical Care Medicine, Am. Thoracic. Soc.*, vol. 165, no. 9, pp. 1217–1239, 2002. 126

[84] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," *New England Journal of Medicine, Mass Medical Soc.*, vol. 328, no. 17, pp. 1230–1235, 1993. 126

[85] Sleep Disorder Clinic, "University of Athens, Medical School, Euagelismos Hospital." http://www.hypnos.gr/, August 2012. 126, 134

[86] V. Kapur, D. Blough, R. Sandblom, R. Hert, J. de Maine, S. Sullivan, and B. Psaty, "The medical costs of undiagnosed sleep apnea.," *New England Journal of Medicine*, vol. 22, no. 6, pp. 749–755, 1999. 126

[87] K. Pang and D. Terris, "Screening for obstructive sleep apnea: an evidence-based analysis," *American journal of otolaryngology, Elsevier*, vol. 27, no. 2, pp. 112–118, 2006. 127

[88] M. Mendez, A. Bianchi, M. Matteucci, S. Cerutti, and T. Penzel, "Sleep apnea screening by autoregressive models from a single ECG lead," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 12, pp. 2838–2850, 2009. 127

[89] T. Sugi, F. Kawana, and M. Nakamura, "Automatic EEG arousal detection for sleep apnea syndrome," *Biomedical Signal Processing and Control, Elsevier*, vol. 4, no. 4, pp. 329–337, 2009. 127

[90] D. Morillo, J. Ojeda, L. Foix, and A. Jiménez, "An accelerometer-based device for sleep apnea screening," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 2, pp. 491–499, 2010. 128

[91] K. Watanabe, T. Watanabe, H. Watanabe, H. Ando, T. Ishikawa, and K. Kobayashi, "Non-invasive measurement of heartbeat, respiration, snoring and body movements of a subject in bed via a pneumatic method," *Biomedical Engineering, IEEE Transactions on*, vol. 52, no. 12, pp. 2100–2107, 2005. 128

[92] D. Van Brunt, K. Lichstein, S. Noe, R. Aguillard, K. Lester, *et al.*, "Intensity pattern of snoring sounds as a predictor for sleep-disordered breathing.," *Sleep*, vol. 20, no. 12, p. 1151, 1997. 128

[93] J. Fiz, J. Abad, R. Jane, M. Riera, M. Mananas, P. Caminal, D. Rodenstein, and J. Morera, "Acoustic analysis of snoring sound in patients with simple snoring and obstructive sleep apnoea," *European Respiratory Journal*, vol. 9, no. 11, pp. 2365–2370, 1996. 128

[94] A. Ng, T. Koh, E. Baey, and K. Puvanendran, "Diagnosis of obstructive sleep apnea using formant features of snore signals," in *World Congress on Medical Physics and Biomedical Engineering 2006*, pp. 967–970, Springer, 2007. 128

[95] A. Ng, T. Koh, E. Baey, T. Lee, U. Abeyratne, and K. Puvanendran, "Could Formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea?," *Sleep medicine, Elsevier*, vol. 9, no. 8, pp. 894–898, 2008. 128

[96] Z. Yu and W. Ser, "Kalman smoother and its application in analysis of snoring sounds for the diagnosis of obstructive sleep apnea," in *World Congress on Medical Physics and Biomedical Engineering 2006*, pp. 1041–1044, Springer, 2007. 128

[97] A. Ng, T. San Koh, K. Puvanendran, and U. Abeyratne, "Snore signal enhancement and activity detection via translation-invariant wavelet transform," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 10, pp. 2332–2342, 2008. 128

[98] U. Abeyratne, A. Wakwella, and C. Hukins, "Pitch jump probability measures for the analysis of snoring sounds in apnea," *Physiological measurement, IOP Publishing*, vol. 26, p. 779, 2005. 135

[99] H. Nakano, M. Hayashi, E. Ohshima, N. Nishikata, T. Shinohara, *et al.*, "Validation of a new system of tracheal sound analysis for the diagnosis of sleep apnea-hypopnea syndrome," *Sleep New York Then Westchester, American Academy of Sleep Medicine*, vol. 27, no. 5, pp. 951–958, 2004. 135

[100] M. Cavusoglu, T. Ciloglu, Y. Serinagaoglu, M. Kamasak, O. Erogul, and T. Akcam, "Investigation of sequential properties of snoring episodes for obstructive sleep apnoea identification," *Physiological measurement, IOP Publishing*, vol. 29, p. 879, 2008. 136