Audio segmentation based on transform domain

Moe Pwint

2005

Moe, P. (2005). Audio segmentation based on transform domain. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/4897>

<https://doi.org/10.32657/10356/4897>

Nanyang Technological University

Downloaded on 26 Apr 2023 20:54:34 SGT

Audio Segmentation Based On Transform Domain

Moe Pwint

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2005

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me and has not been submitted for a higher degree to any other University or Institute.

.....

Date

.....

Moe Pwint

*To my late parents,
for their encouragements and love.*

Acknowledgements

First, I would like to express my deepest gratitude to my supervisor Prof. Farook Sattar for motivating, directing, inspiring and supporting throughout all stages of my Ph.D study. Not only his guidance and co-operation are excellent but also his suggestions and advice are valuable. Whenever I was stuck on a problem, he always provided his technical knowledge and insightful perspectives. It was a great pleasure to work with him since he is kind, patient and understanding.

I would also like to thank Nanyang Technological University for giving me a chance to pursue Ph.D degree at Singapore and providing me three years scholarship. I would like to thank all my fellow graduate students and technicians in the Information System Research lab and Media Technology lab.

Thanks are due to all of my friends for their helps and words of encouragements when they are needed. I am also grateful to my fellow supervisor mates Yu Dan and Gao Xinting for their support, help, collaboration and friendship throughout my study in NTU.

Finally, I am indebted to my younger sisters Hninn Pwint and Zun Pwint for their encouragement, understanding, moral support, and love they have given me. Without their support, this project would never have been accomplished and this thesis would never have come into existence. I am also grateful to my youngest brother Ba Nyar for his love and encouragement.

This thesis is dedicated to the memory of my late parents, who always would have liked to see me pursuing Ph.D degree. They always encouraged and supported me to pursue more advanced education since I was in elementary school. Their solemn wish always has been a source of inspiration in pursuing this degree.

Contents

Acknowledgements	i
List of Figures	v
List of Tables	ix
List of Abbreviations	x
Summary	xi
1 Introduction	1
1.1 Motivation	2
1.2 Objectives	3
1.3 Thesis Contributions	4
1.4 Thesis Organization	6
2 Previous Work and Background	7
2.1 Previous Work	8
2.1.1 Speech and Non-speech Detection	8
2.1.2 Audio Segmentation	11
2.2 Background of Proposed Techniques	13
2.2.1 Walsh Transform	14
2.2.2 Wavelet Transform	17
2.2.3 Genetic Algorithms	18
2.2.4 Entropy Metric	21
2.2.5 Time Scaling	23

3	Speech/Non-Speech Detection	26
3.1	Introduction	27
3.2	Proposed Algorithm	29
3.2.1	Transformation of Signals	30
3.2.2	Determination of Minimal Walsh Basis Function	31
3.2.3	Decision Scheme	37
3.3	Experimental Study and Comparison	40
3.4	Receiver Operating Characteristics Analysis	44
3.5	Conclusion	46
4	Segmentation Based on Genetic Algorithm	49
4.1	Introduction	50
4.2	Determination of Number of Segments	52
4.2.1	Mean Difference Measure	53
4.3	Segmenting by a Genetic Algorithm	53
4.3.1	Sample Entropy	54
4.3.2	Genetic Algorithm	56
4.4	Experimental Results	61
4.4.1	Evaluation Metrics	63
4.5	Conclusion	68
5	Segmentation of Narrowly Spaced Noisy Audio Signals	69
5.1	Introduction	69
5.2	Time Scaling Based Segmentation	71
5.2.1	Vocoder Based Time Scaling	72
5.2.2	Experimental Study	77
5.2.3	Discussion	79
5.3	Segmentation Based on Residual Signal and Convexity Measure	80
5.3.1	Analysis-Synthesis Process	81
5.3.2	Detection Scheme	82
5.3.3	Convexity	83
5.3.4	Experiments and Evaluations	84

5.4	Conclusion	88
6	Segmentation of Noisy Multicomponent Audio Signals in Time-Scale Domain	90
6.1	Introduction	90
6.2	Signal Model	91
6.3	Assumptions	92
6.4	Segmentation Using Wavelet Transform	92
6.4.1	Morlet Wavelet Transform	93
6.4.2	Windowed Approximate Entropy	94
6.4.3	Algorithm	96
6.4.4	Evaluation	97
6.4.5	Application with Real Speech Signals	102
6.5	Segmentation Using Linear Binary Time-Scale Transform	103
6.5.1	Choice of the Transform	104
6.5.2	Basic Concept	105
6.5.3	Method	106
6.5.4	Evaluation	108
6.6	Discussion	110
6.7	Conclusion	111
7	Conclusions And Recommendation	114
7.1	Speech/Non-Speech Detection	114
7.2	Segmentation Based on Genetic Algorithm	115
7.3	Segmentation of Narrowly Spaced Noisy Audio Signals	116
7.4	Segmentation of Noisy Multicomponent Audio Signals in Time-Scale Domain	117
7.5	Recommendation	118
	Author's Publications	120
	Bibliography	122

List of Figures

2.1	A set of Walsh basis functions.	15
2.2	Schematic diagram of a phase vocoder.	25
3.1	Block diagram of the proposed speech/non-speech detection algorithm.	30
3.2	The distribution of the order of basis functions for the signals from clean to 0 dB.	34
3.3	The clean signal.	35
3.4	The noisy signal.	35
3.5	The modified signal using 0-order basis function ϕ_0	36
3.6	The modified signal using basis function ϕ_m	37
3.7	A clean signal.	38
3.8	The reconstructed signal using shifting operator CS	39
3.9	The reconstructed signal without using shifting operator CS	40
3.10	(a) Noisy speech at SNR = 0 dB in white Gaussian noise (b) manual detection (c) estimated detection.	41
3.11	Performance comparison for average non-speech detection rate of the proposed method and standard VADs (G.729, AMR1 and AMR2) in different backgrounds with varying SNRs.	43
3.12	Performance comparison for average speech detection rate of the proposed method and standard VADs (G.729, AMR1 and AMR2) in different backgrounds with varying SNRs.	44
3.13	Performance comparison for average error rate of the proposed method and standard VADs (G.729, AMR1 and AMR2) in different backgrounds with varying SNRs.	45
3.14	Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1 and AMR2 at 20 dB white Gaussian noise.	47

3.15	Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1 and AMR2 at 10 dB white Gaussian noise.	48
3.16	Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1 and AMR2 at 0 dB white Gaussian noise.	48
4.1	The analysis of sample entropy in a clean speech signal.	56
4.2	The analysis of sample entropy in a noisy speech signal.	56
4.3	(a) Noisy input signal at 5 dB SNR; (b) segmentation results shown in vertical lines together with the clean speech signal.	61
4.4	(a) Noisy input signal at 0 dB SNR; (b) segmentation results shown in dashed lines together with the clean speech signal.	62
4.5	Variation of the relative error with the number of generations.	64
4.6	Hit rate as a function of SNR in different noise backgrounds.	65
5.1	(a) Noise-free signal; (b) noisy signal; (c) time-scaled version of noisy signal using sinusoidal analysis-synthesis; (d) time-scaled signal using Walsh transform.	75
5.2	(a) Noise-free signal; (b) signal at 20 dB SNR; (c) modified time-stretched signal (d) signal at 0 dB SNR (e) modified time-stretched signal.	76
5.3	(a) Noise-free signal; (b) noisy signal; (c) time-scaled signal (d) segmentation results.	77
5.4	Modification of a music excerpt. (a) original signal; (b) modification without stretching; (c) modification with stretching factor $\alpha = 2$; (d) signal with gaps; (e) modification without stretching; (f) modification with stretching factor $\alpha = 2$	78
5.5	The analysis and synthesis process.	83
5.6	Segmentation result of a noisy signal (a) noise-free signal; (b) noisy signal; (c) segmentation result before convexity measurement (d) segmentation result after convexity measurement.	85
5.7	The recall rate for test data set I across five SNRs and five noise conditions.	86
5.8	The precision rate for test data set I across five SNRs and five noise conditions.	87
5.9	The error (difference between the real and estimated onset) of segments for 450 signals at SNR = 20, 15, 10, 5, 0 dB by investigating on three types of noise.	88

6.1	The results of our scale domain segmentation when the noise standard deviation is 1. (a) noise-free signal; (b) noisy signal; (c) highest absolute scales (for total number scales, $J=64$) vs. time ; (d) local maxima/minima; (e) selected local maxima; (f) masked/extracted noisy input signal; (g) approximate entropy; (h) binary segmentation function/estimated segment intervals and noisy gaps.	98
6.2	The results of segmentation are illustrated for low SNR (noise standard deviation=1.0) using the noisy signal in Fig. 6.2(c). Note that for segmentation in low SNR only the scale domain plot (see Fig. 6.2(b)) can be exploited, while time domain plots (see Figs. 6.2(d)-(e)) are difficult to exploit. (a) noise-free signal; (b) highest absolute scales (for total number scales, $J=64$) versus time; (c) noisy signal; (d) normalized average instantaneous energy; (e) estimated segments and noisy gaps. .	100
6.3	The results of segmentation are illustrated for high SNR (noise standard deviation = 0.2) using the noisy signal in Fig. 6.3(c). Note that the scale domain characteristics are the same for all components as we have used high frequency/scale wavelets. The high wavelet scales are able to separate in time domain. (a) noise-free signal; (b) highest absolute scales (for total number scales, $J=64$) versus time; (c) noisy signal; (d) normalized average instantaneous energy; (e) estimated segments and noisy gaps.	101
6.4	Error (difference between the real and estimated gap positions (Figs.6.4(a)-(c)) as well as the real and estimated gap durations (Figs.6.4(d)-(f)) of the segmentation results for the 750 simulated noisy signals at 14 dB (noise std=0.5), 5 dB (noise std=.75), and 0 dB (noise std=1) SNR (Figs.6.4(a) and (d):14 dB SNR; Figs.6.4(b) and (e): 5 dB SNR; Figs.6.4(c) and (f): 0 dB SNR). (X-axis=Trail, Y-axis=Mean error(in sample))	102
6.5	Segmentation results for real speech signals (The segmentation functions (scaled by 5) is plotted together with the input speech signal).	103
6.6	The results of segmentation are shown at low SNR when noise standard deviation is 0.5. (a) noise-free signal; (b) noisy signal; (c) segmentation using time-scale domain.	109
6.7	The results of segmentation are shown at low SNR when noise standard deviation is 1. (a) noise-free signal; (b) noisy signal; (c) segmentation using time-scale domain.	110
6.8	The results on segmentation are shown for the choice of number of coefficients ($a_o, a_1, \dots, a_{M-1}$) and the corresponding basis functions ($\phi_o, \phi_1, \dots, \phi_{M-1}$) to be used for segmentation; (a) using only a_o , (b) using a_o and a_1 , (c) using a_o , a_1 and a_2 , (d) using a_o , a_1 , a_2 and a_3 , (e) a_o , a_1 , a_2 , a_3 and a_4 (The noise standard deviation is 0.5 and the segments are powers of 2).	111

6.9	The results of segmentation are shown for the choice of number of coefficients $(a_o, a_1, \dots, a_{M-1})$ and the corresponding basis functions $(\phi_o, \phi_1, \dots, \phi_{M-1})$ to be used for segmentation; (a) using only a_o , (b) using a_o and a_1 , (c) using a_o, a_1 and a_2 , (d) using a_o, a_1, a_2 and a_3 , (e) a_o, a_1, a_2, a_3 and a_4 (The noise standard deviation is 1 and the segments are powers of 2).	112
6.10	Segmentation for a clean voiced speech; (a) clean voiced speech, (b) segmentation function describing the segments.	113
6.11	Segmentation for a noisy voiced speech; (a) noisy voiced speech, (b) segmentation function describing the segments.	113

List of Tables

3.1	Comparison of speech detection rates, non-speech detection rates and error rates of proposed method to standard methods, G.729, AMR1 and AMR2 for different levels of SNRs in various noisy environments. . . .	42
3.2	Comparison of the proposed method using basis functions ϕ_0 and ϕ_m for different levels of SNRs in various noisy environments.	46
4.1	Relative error at different SNRs.	63
4.2	Results of recall (RCL), precision (PRC) and F on different types of noise.	66
4.3	Classification accuracy of speech and non-speech frames.	67
4.4	Processing time.	67
5.1	Precision rate with different SNRs.	87
5.2	Recall rate with different SNRs.	88

List of Abbreviations

ABBREVIATIONS

FULL EXPRESSIONS

AMR	Adaptive Multi-rate Speech Coder
ApEn	Approximate Entropy
ASR	Automatic Speech Recognition
BIC	Bayesian Information Criterion
CWT	Continuous Wavelet Transform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DTX	Discontinuous Transmission
DWT	Discrete Wavelet Transform
ECG	Electrocardiogram
EMG	Electromyography
FFT	Fast Fourier Transform
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
KNN	K - Nearest Neighbour
LPCC	Linear Predictive Cepstral Coefficient
MFCC	Mel Frequency Cepstral Coefficient
PCA	Principal Component Analysis
PCG	Phonocardiogram
ROC	Receiver Operating Characteristic
SampEn	Sample Entropy
SNR	Signal-to-Noise Ratio
STFT	Short Time Fourier Transform
SVD	Singular Value Decomposition
VAD	Voice Activity Detector

Summary

With the increasing applications of audio and multimedia data, which are available on the Internet, audio segmentation becomes a challenging problem for many areas of speech and audio signal processing. The aim of this thesis is to design and develop audio segmentation techniques, which are efficient, robust, simple and applicable in real world situations. Segmentation techniques for noisy audio signals are investigated in four chapters of this thesis together with applications.

A new method to detect speech/non-speech components of a given noisy signal is presented employing the combination of binary Walsh basis functions and an analysis-synthesis scheme. The minimal number of Walsh basis functions to be applied is also analyzed. A framework for speech/non-speech detection is then established. The use of minimal Walsh basis functions makes the algorithm computationally efficient and easily implementable.

A novel approach to automatically determine the segment boundaries of noisy speech signals is described. Speech segmentation is formulated as an optimization problem and boundaries of speech segments are detected using a genetic algorithm (GA). Segmentation results are improved through the generations of GA by introducing a new evaluation function, which is based on sample entropy and a heterogeneity measure.

Segmentation of noisy audio signals which are closely spaced in time is also considered in this thesis. Two segmentation techniques are presented to address this issue. The first segmentation method proposed is primarily based on a time scaling approach. Utilizing the fast Fourier transform (FFT) and binary Walsh transform based time scaling approach, the locations and durations of noisy segments as well as

noise only segments of narrow intervals are determined. Employing the residual signal, which is developed by Walsh basis functions, the second segmentation algorithm has been proposed. The use of a convexity measure and lower order Walsh basis function has not only improved the accuracy of segmentation results but also the computational efficiency.

In addition, segmentation methods for multicomponent noisy audio signal in time-scale domain are presented in this thesis. A wavelet domain based segmentation approach is proposed. It is shown that the Morlet wavelet transform is useful for segmenting a noisy signal where the signal components are closely spaced. The usefulness of windowed approximate entropy is also investigated to estimate the locations and durations of narrow gaps and noisy segments. A linear binary time-scale transform based method is also developed for the multicomponents signal segmentation issue. This method is also applied for the detection of audio signal components which are closely spaced and for which the time interval between adjacent signal components is unknown.

Chapter 1

Introduction

Segmentation is the initial and crucial stage in many areas of audio processing. In general, audio segmentation is the task of partitioning an audio stream into its homogeneous units. For example, segmentation of audio data into speech, music, noise and silence has many applications such as audio content analysis, audio retrieval, speaker indexing and audio indexing. Furthermore, in an automatic speech recognition (ASR) system, the input signals are a continuous stream of speech signals to be fragmented. Segmentation of speech signals, therefore, into acoustically separated words or sub-word units is a prerequisite for ASR systems. When an ASR system has to deal with a large vocabulary, there is a high complexity in computation and requirement of large amount of memory space, which are inevitable. One way to reduce this computational burden is to pre-process the acoustic input through an effective segmentation algorithm. Several ASR systems have been designed in the literature and high levels of performance are achieved by these systems when they are operated in controlled situations [1]. However, they suffer substantial performance degradation when they are applied in noisy situations.

In addition to ASRs, segmented homogeneous regions of speech can be applied in

clustering of speech stream into speaker turns of broadcast news and meeting recordings [2]. The information of these speaker turns can also be useful for applications such as meeting and conference summarizations. With the advent of Internet technology and efficient coding schemes, large amount of digital audio and video are available on the web and more and more multimedia data are distributed over the Internet. From the multimedia technology point of view, segmentation is essential for many applications such as indexing of audio data, transcription of broadcast news and information retrieval where large amount of audio data are necessary to be handled. For example, to classify and index a large amount of audio recordings, which are available online over the web, the best indexing scheme can be realised only through an efficient segmentation method.

In audio content analysis for information retrieval and indexing of large audio databases, audio segmentation is the first step to search for a desired piece of audio data. Audio segmentation techniques are often employed to assist video scene analysis. As the audio tracks are mixture of different audio types, in this context, they are segmented into several classes such as speech, music, silence or ambient noise.

1.1 Motivation

There are several possible applications where segmentation is essential including ASR, managing of archived audio data and entertainment. For instance, separating the signal components into speech and non-speech segments is the primary task of any ASR. In reducing the requirement of large memory space in large vocabulary speech recognition system, segmentation also plays a vital role. For a long time, the segmentation of audio signals has been a major issue for an ASR. To handle large amount of audio and multimedia data available on the Internet, the use of segmentation can provide

fast information retrieval and indexing. However, manual segmentation is tedious, expensive and unrealistic for practical applications.

Although several segmentation algorithms have been proposed to address these issues [3], the systems which can meet the requirements of practical applications are not fully developed yet. We all know that noise is a well-known source of errors in many areas of signal processing. A lot of segmentation works have also been focused on the robustness against noise issue. However, the state-of-the-art of such systems cannot fully cover various types and levels of noise which are encountered in practical situations.

In this thesis, segmentation techniques for audio signals in noisy environments are suggested with its applications to segmentation of speech signals. However, they are applicable to other audio signals such as phonocardiogram (PCG) signals and electromyography (EMG) signals.

1.2 Objectives

The aim of this thesis is to explore audio segmentation techniques, which are efficient and applicable in real world situations. Segmentation techniques for audio signals in noisy conditions where the noise level is high and noise type is unknown are investigated for the following objectives:

- **Detection of Speech and Non-speech:** to propose an efficient segmentation algorithm in noisy environments, which would be applicable for detection of speech/non-speech in real time noise cancellation and speech enhancement systems.

- **Detection of Speech Segments Boundaries:** to propose a robust segmentation technique for determination of boundaries of speech segments in different environmental conditions. The proposed method would achieve robustness against noise without any preprocessing.
- **Segmentation of Narrowly Spaced Audio Signals:** to propose segmentation techniques for detection of signal components which are closely located in time and buried under high level of background noise. It is also assumed that time intervals between adjacent signal components are random and relatively smaller than the signal components.
- **Segmentation of Multicomponent Audio Signals in Time-Scale Domain:** to propose efficient segmentation techniques for multicomponent noisy audio signals in time-scale domain.

1.3 Thesis Contributions

Original contributions of this thesis can be briefly described as follows:

- A discrimination method for speech and non-speech segments of noisy speech signals is proposed using minimal Walsh basis functions based analysis-synthesis scheme. Unlike most of the previously reported speech/non-speech detection method, this method does not apply feature extraction and models of speech and non-speech. Thus, only a few parameters are to be optimized in this framework.
- The proposal for detection of segment boundaries of speech signals in varying background conditions is presented. Formulating the problem of segmentation as machine learning, locations of speech segments boundaries are determined. To the best of our knowledge, this is the first segmentation algorithm for speech

boundaries detection by means of a Genetic Algorithm (GA). Simulation study, which has been conducted on various speech signals with an evaluation metric shows the robustness of the proposed method.

- Time scaling techniques are usually employed to modify the time scale of an acoustic signal without losing the perceptual quality. A Walsh transform based time scaling scheme is introduced as a tool of segmentation method. The basic idea is to detect the small gaps and signal components from the reconstructed time-scale modified signal, which appears to be more useful than the original signal in segmentation. The usefulness of this approach is preliminarily investigated for narrowly spaced noisy speech and music signals.
- The proposal for identification of narrowly spaced speech segments using a residual signal, which is developed from Walsh basis functions is presented. A convexity measure is introduced to improve the segmentation results. Based on the experiments carried out using two sets of database, the proposed method achieves robustness against noise.
- To analyze the multicomponents of closely spaced noisy audio waveforms, a segmentation approach is developed exploiting wavelet transform and windowed approximate entropy. It has been shown that the proposed time-scale domain based method outperforms the Principal Component Analysis (PCA) based time-domain segmentation method.
- The proposal for detection of active intervals and quiet intervals of noisy audio signals using linear binary time-scale transform and minimal *a priori* information is also presented. The results obtained from synthetic signals are found promising.

1.4 Thesis Organization

The motivation and objectives of this thesis is introduced in Chapter 1. The rest of the organization of the thesis can be outlined as follows.

Chapter 2 reviews previous works on audio segmentation techniques in various application areas. Fundamentals of segmentation approaches presented in this thesis are also described in this chapter.

Chapter 3 presents a speech/non-speech detection technique based on minimal Walsh basis functions. An analysis to determine the minimal number of Walsh basis functions is also discussed in this chapter.

Chapter 4 introduces a novel segmentation technique based on a genetic algorithm (GA). Sample entropy, which is used in designing the evaluation function of GA is also briefly explained in this chapter.

Chapter 5 deals with the problem of closely spaced noisy audio segmentation. A time scaling approach based segmentation technique is proposed first. This chapter further describes a method to detect narrowly spaced audio signals based on a residual signal and a convexity measure.

Chapter 6 first propose a technique for segmentation of noisy multicomponent audio signals in wavelet domain. This chapter also discusses windowed approximate entropy which is used in this segmentation method. A Walsh transform based segmentation algorithm is also presented in this chapter.

Chapter 7 organizes the summaries of the works presented in this thesis, contributions and discussion for future directions.

Chapter 2

Previous Work and Background

During the past few years, there has been a growing amount of multimedia data such as audio and video that are available in digital format over the worldwide net with the availability of high speed and high capacity computers, Intranets and the Internet. Automatic audio indexing and retrieval systems are therefore essential to access a desired piece of information from the large amount of available audio data. Hence, efficient techniques to automatically segment or classify an audio stream into homogeneous regions depending on the specific applications become more and more important. Such methods are useful not only for audio content analysis and audio retrieval, but also applicable in speech recognition systems with large vocabularies. Recently, potentials of audio segmentation and classification have been investigated in video segmentation and video scene analysis since audio-based analysis systems are effective and require less computation time.

In this chapter, a review of the literature on audio segmentation in different areas of multimedia signal processing is described (Section 2.1). A brief introduction to the fundamentals of the segmentation techniques, which are proposed in this thesis, is also described in this chapter (Section 2.2).

2.1 Previous Work

The purpose of this review is to mention some of the works that have been presented in the literature for the issue of audio segmentation. Some of the basic concepts of audio segmentation used for voice activity detection (VAD), word boundary detection, ASR, indexing and content analysis are briefly mentioned.

2.1.1 Speech and Non-speech Detection

Several methods have been proposed in the literature related to speech and non-speech detection problem. In many fields of speech processing such as speech recognition [4], speech and audio coding [5], speech enhancement [6], wireless communication [7], efficient speech/non-speech segmentation algorithms are always necessary. For applications like mobile phones and packet networks, discontinuous transmission (DTX) mode based on an efficient speech and non-speech discriminator is usually required for lower bit-rate transmission speech coder. DTX is a crucial part of global system in mobile communications (GSM). It helps to increase the overall capacity of the system by minimizing co-channel interference. A VAD, which distinguishes between speech and noise of a signal, is the most important part of DTX. If some speech frames are identified as noise, intelligibility of speech can be decreased due to an unpleasant effect called *clipping*. On the other hand, if inactive (i.e. noise) frames are detected as active frames frequently, the quality of DTX becomes very poor in developing an efficient silence compression scheme. A VAD, which is standardized for the GSM cellular communication system is ETSI speech coder [8]. Based on spectral estimation and periodicity detection, this adaptive multi-rate speech coder (AMR) specifies two options for VAD to be used in DTX.

Noise is a well-known factor which degrades the quality and intelligibility of speech

in many applications areas. To reduce the noise level without affecting the quality of speech signals, a noise reduction algorithm is usually employed. Spectral subtraction is a widely used approach in practical noise suppression schemes. This scheme usually estimates the noise characteristics from the non-speech intervals of the signal. Therefore, identification of non-speech periods is an important and sensitive part of existing noise reduction schemes. In this context, accuracy and reliability of a VAD becomes critical in determining the performance of noise reduction algorithm. Most papers reporting on noise reduction refer to speech pause detection when dealing with the problem of noise estimation. Speech pause detectors are very sensitive and often limiting part of the systems for the reduction of noise in speech [9].

A speech pause detection algorithm based on an autocorrelation voicing detector algorithm was developed in [10]. The algorithm was designed for real-time system and implemented on a DSP platform for the application of speech enhancement for hearing aids. An adaptive Karhunen-Loève transform (KLT) tracking-based algorithm was also proposed for enhancement of speech degraded by additive color noise [11]. An algorithm, which detects the speech pauses by tracking the dynamics of the signal's temporal power envelope was proposed in [9]. Sometimes, detection algorithms were designed for specific applications such as noise suppression [12] and wideband coding [5]. Voice activity detection algorithms for cellular networks in the presence of babble noise and vehicular noise were presented in [13] by adopting the approach used in European digital mobile cellular standard [14]. Combining the geometrically adaptive energy threshold method (GAET) and least-square periodicity estimator (LSPE), conversational speech is separated from silence [15]. A fuzzy polarity correlation function is also applied to determine speech sections and background noise in the environment of telephone network [16].

VAD methods can be roughly categorized into rule-based and model-based or

classifier-based methods. In the rule-based methods, an acoustic feature vector is usually extracted first. Most of the widely used parameters for VAD algorithms are short-time energy, zero-crossing rate, mel frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC). Heuristically derived rules are then applied to separate speech from non-speech components of the signal [8, 12, 17]. Since a rule derived is specific to the extracted feature vector, different rules might be necessary for different combinations of features. Moreover, fine tuning is often needed to cope with the changing conditions of the data such as noise. Thus, it would be very difficult to obtain consistent and optimized results using the rule-based detectors.

In the classifier based methods, the problem of speech/non-speech detection is formulated as a classification problem assuming that speech and non-speech events are of different classes. Recently, statistical model based speech detectors have been developed. Different models such as Gaussian, complex Gaussian and Laplacian-Gaussian models are usually employed to model the distributions of speech and noise [18, 19, 20]. Based on these statistical models, a noisy speech is modelled with two hypotheses. Speech and non-speech events are then detected based on Hidden Markov Models (HMM). In order to reduce the miss detection errors, most of the rule-based and classifier-based methods often employ a hangover scheme assuming that speech events are highly correlated with time [8, 17, 18]. Although the accuracy of classification methods is more reliable than rule-based methods, they are usually complex in computation and have larger latency. The limitation of most of the existing methods is that they only work well in the case of high level of SNRs. Although there are some methods which can achieve good performance even at low SNRs, they are very complex and difficult to implement in real-time systems.

Taking the advantages of signal properties in a transformed domain, we have proposed a binary Walsh transform based speech/non-speech detection technique in the

presence of noise in Chapter 3.

2.1.2 Audio Segmentation

There are a number of works in automatic segmentation of audio signals. According to [3], the existing segmentation algorithms can be categorized into:

- Decoder-guided segmentation
- Model-based segmentation
- Metric-based segmentation

Decoder-guided segmentation first decodes the input audio stream using acoustic models. By chopping the input audio at the locations of silence generated from the decoder, the desired audio segments can be obtained heuristically. The gender change information from the decoder could also be utilized for segmentation. More details of this scheme can be found in [21, 22, 23].

Model-based segmentation builds different models such as Hidden Markov Models (HMM) or Gaussian Mixture Models (GMM) for a fixed set of acoustic classes such as speech, music from a training corpus. Then the incoming audio stream is classified by maximum likelihood selection over a sliding window. Segmentation occurs at the location where there is a change in the audio class. The HMM model for each acoustic class is trained where the model parameters can be estimated using a procedure such as Baum-Welch algorithm. Then the class HMMs are combined into segmentation network. Segmentation is achieved by using the Viterbi algorithm to determine the maximum likelihood sequence of states through the network [24].

Metric-based segmentation is based on the acoustic distance measurement. If the acoustic distance between two neighbouring windows is large, then the input audio

stream is to be segmented at that place.

Since the model-based segmentation and metric or rule based segmentation methods use most of the acoustic features applied in VADs methods, some thresholding schemes or heuristic rules are often needed in most of the audio segmentation approaches. Another limitation of the model-based methods is that they cannot be generalized to unknown acoustic conditions.

In ASR systems, the purpose of segmentation is to divide the continuous input speech stream into discrete units based on an acoustic similarity measure. Most of the existing segmentation algorithms are based on simple parameters such as energy, zero-crossing measure and MFCC. It is well known that one main source of errors in ASR is the imprecise detection of segment boundaries of the input utterance especially in noisy environments. A good segmentation algorithm should therefore provide accurate detection of speech and noise intervals with less misclassification for wide range of SNRs. Since energy and zero-crossing rates do not work well in noisy environments, sometimes entropy measure is combined with these features to achieve a certain level of robustness [25, 26].

Fuzzy logic and fuzzy rule based endpoint detectors and neural network endpoint detectors were proposed in [27, 28]. Junqua *et al.* [29] proposed a method where a parameter is derived from time and frequency features, which is especially beneficial in low SNR. In a word boundaries detection algorithm of [30], this time-frequency parameter is extended to extract useful frequency information using a neural fuzzy network. Despite small network size, high learning accuracy and high learning speed, this algorithm needs a longer CPU time.

A comparison and suggestion of features for speech/music segmentation has been reported in [31]. Using multiple features and models, a segmentation algorithm was

developed in [32] where a section of audio is considered as a semantically consistent sound segment if there is no significant changes in the features. This method achieves a strong agreement between the boundaries of audio and video scenes. Bayesian Information Criterion (BIC) is the one of the mostly used criterion in audio segmentation and classification process [33, 34]. Segmentation of audio stream into homogeneous regions according to speaker identity, environmental conditions and channel condition was presented in [3] using BIC. A method to retrieve the topics of broadcasted TV news based on the audio segmentation deploying BIC was also proposed in [35].

In the context of audio and multimedia content analysis, acoustic features like energy, zero-crossing rate and MFCC are commonly used for audio segmentation and classification. A review of the multimedia content analysis using audio and visual information can be found in [36]. In this review, audio features are categorized into two levels namely frame-level features and clip-level features. Some devised features such as cepstrum flux and block cepstrum flux parameters are introduced in [37] to fragment the audio for video streams. A video segmentation scheme with an audio analysis has been presented in [38]. In this analysis, discrimination of speech and non-speech is performed by a KNN (K - Nearest Neighbour) classifier and a GM-VQ (Gaussian Model-Vector Quantization) method. Based on the audio periodicity and other features, non-speech segments are further classified into music and environmental sound.

2.2 Background of Proposed Techniques

In this section, the background of the proposed segmentation techniques as presented in the following chapters are discussed.

2.2.1 Walsh Transform

The Walsh transform is a matrix product of a square set of data d and a matrix of basis vectors consisting of Walsh functions. The Walsh functions consist of trains of square pulses (with the allowed states being -1 and 1) such that transitions may only occur at fixed intervals of a unit time step, the initial state is always +1, and the functions satisfy certain other orthogonality relations [39]. For Walsh matrices W_n of order $N = 2^n$, $W_n \cdot W_n^T = W_n^T \cdot W_n = N \cdot I_n$ where I_n is the identity matrix of order 2^n . The choice of the Walsh transform rather than other orthogonal transforms in this thesis is due to its computational simplicity involving addition and subtraction only.

Walsh transforms are widely used in communications, signal and image processing and image compression [40, 41, 42]. The set of Walsh functions is generally classified into three groups. These groups differ from one another in that the order in which individual functions appear is different. Three orderings of Walsh functions in literature are Hadamard (natural ordering), Paley (dyadic ordering) and Walsh (sequency ordering) [43]. Figure 2.1 shows the first 8 Walsh functions arranged in dyadic order.

Let k_i and j_i be the i -th bit in the binary representations of integers k and j , respectively. The kernels of Walsh-Paley ($wal_{kj}^{(p)}$), Walsh-Hadamard ($wal_{kj}^{(h)}$) are given by

$$wal_{kj}^{(p)} = (-1)^{\sum_{i=0}^{n-1} k_{n-1-i} \cdot j_i} \quad (2.1)$$

$$wal_{kj}^{(h)} = (-1)^{\sum_{i=0}^{n-1} k_i \cdot j_i} \quad (2.2)$$

where $k, j = 0, 1, \dots, N - 1$ and $N = 2^n$. The elements of Walsh can also be generated by

$$wal_{kj}^{(w)} = (-1)^{\sum_{i=0}^{n-1} r_i(k) \cdot j_i} \quad (2.3)$$

Figure 2.1: A set of Walsh basis functions.

where

$$\begin{aligned}
 r_0(k) &= k_{n-1} \\
 r_1(k) &= k_{n-1} + k_{n-2} \\
 r_2(k) &= k_{n-2} + k_{n-3} \\
 &\vdots \\
 r_{n-1}(k) &= k_1 + k_0
 \end{aligned} \tag{2.4}$$

and $k, j = 0, 1, \dots, N-1$ and $N = 2^n$. When $N = 2^3$, matrices of Walsh-Paley (W_P), Walsh-Hadamard (W_H), and Walsh (W_W) have the following form. In this thesis we

use the ordering of Walsh, i.e. W_W .

$$W_P = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & -1 & -1 & +1 & -1 & +1 & +1 & -1 \end{bmatrix} \quad (2.5)$$

$$W_H = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 & -1 & +1 & +1 & -1 \end{bmatrix} \quad (2.6)$$

$$W_W = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & -1 & -1 & +1 & -1 & +1 & +1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \end{bmatrix} \quad (2.7)$$

2.2.2 Wavelet Transform

The wavelet transform has been used to analyze non-stationary signals in several research areas such as image processing, speech processing, subband coding, noise removal from time series and seismic geology. The wavelet transform provides high temporal resolution for high frequency components and high frequency resolution for low frequency components. Wavelets are generated from a single analyzing function called *mother wavelet*, $\Psi(t)$. The wavelet transform decomposes the signal into a set of basis functions, (i.e. wavelets), by scaling and translating of the mother wavelet. A family of wavelets can be constituted from the scaling parameter a and translating parameter τ as:

$$\Psi_{a,\tau}(t) = \frac{1}{\sqrt{a}} \Psi\left(\frac{t-\tau}{a}\right) \quad (2.8)$$

Wavelet transforms can be classified into discrete wavelets and continuous wavelets. For the continuous wavelet transform (CWT), $(a, \tau) \in R$ and $a \neq 0$ where the discrete wavelet transform (DWT) uses only a finite set of values of scale and translation factors.

The continuous wavelet transform of a signal $x(t)$ with respect to scale factor a and shifting or translation factor τ can be approximated by scaling and translating the mother wavelet $\Psi(t)$ as

$$CWT(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \Psi^*\left(\frac{t-\tau}{a}\right) dt \quad (2.9)$$

where $*$ denotes the complex conjugate form. In a discrete wavelet transform, the translation and the scaling factor are power of two such that $a = 2^j$ and $\tau = ka = k2^j$ where $j, k \in integers$. In [44], the discrete wavelet function is given as

$$\Psi_{j,k} = 2^{-j/2} \Psi\left(\frac{t - 2^j k}{2^j}\right) \quad (2.10)$$

The discrete wavelet transform (DWT) is defined as

$$DWT_{j,k} = \sum_{n=0}^{\infty} x(n) \Psi_{j,k} = \sum_{n=0}^{\infty} x(n) \frac{1}{\sqrt{2^j}} \Psi\left(\frac{n - k2^j}{2^j}\right) \quad (2.11)$$

A small scaling parameter corresponds to a compressed wavelet function. As a result, the rapidly varying features of the signal can be obtained from the wavelet transform by using a small scaling factor. On the other hand, low frequency features of signal $x(t)$ can be extracted by using a large scaling factor with a stretched wavelet function. A small scaling value is used for local analysis and a large scaling is used for global analysis.

2.2.3 Genetic Algorithms

Genetic Algorithms (GAs) are search algorithms based on the mechanics of natural selection and natural genetics [45]. They are part of a collection of stochastic optimization algorithms inspired by natural genetics and the theory of biological evolution [46]. Applying the principle of survival of the fittest, GAs operate on a population of individuals (chromosomes) which are the potential solutions to the problem in the search space. Usually, a predefined number of individuals are generated randomly and encoded into a suitable representation depending on the nature of the problem. The fitness of the individuals in the population are then calculated using an evaluation function. The evaluation function usually provides a score to each individual based on how well they perform in the problem domain.

In order to reproduce the next population, the fittest individuals from the current population are selected using a selection method such as roulette wheel or stochastic universal sampling. Selected individuals are then evolved in order to develop a new population with better individual fitness by applying the genetic operators such as crossover and mutation. In this way, the algorithm maintains the individuals with optimized fitness values and those with lower fitness values are discarded from the population. This process is repeated until the GA has converged to a good solution

or after a predetermined number of generations have passed. A top-level view of the conventional genetic algorithm cycle can be described as follows.

1. Random initialization of a population of chromosomes.
2. Evaluation of the fitness of each chromosome in the current population.
3. Creation of new chromosomes by mating current chromosomes, applying mutation and recombination as the parent chromosome mate.
4. Deletion of members of the population to make room for the new chromosomes.
5. Computation of the fitness of new chromosomes and insertion of them into the population.
6. Stop and return the best chromosome if the termination criteria is met; if not, go to step 3.

The encoding scheme is a key issue in any GA because it can severely limit the window of information that is observed from the system [47]. Although binary encodings are the most commonly used, there is an increasing interest in other representation such as integer and real-valued encodings since they are more natural for some specific problems [48]. Proportionate selection, ranking selection and tournament selection schemes have been common selection methods in GA to emphasize the search in more promising regions of search space. In the proportionate method such as roulette wheel algorithm, members of the population are selected with a probability that is proportional to their fitness score. In ranking, all individuals are sorted by increasing values of their fitness. Then, each individual is assigned a probability of being selected from some prior probability distribution. Tournament selection scheme first selects a group of individuals. The individual with the highest fitness from the group is then selected

and the rest are discarded. Other selection methods include stochastic remainder and stochastic universal sampling.

The performance of selection algorithms are usually measured with bias and spread. Bias is defined as the absolute difference between an individual's actual and expected selection probability. Spread is the range in the possible number of times that an individual may achieve [49]. In stochastic universal sampling method, for N number of individuals to be selected, N pointers have to be equally spaced by $1/N$. The position of the first pointer is given by a randomly generated number in the range $[0, 1/N]$. If $s(i)$ is the actual number of times a particular individual i to be selected, then $s(i) \in \{\lfloor e(i) \rfloor, \lceil e(i) \rceil\}$ where $e(i)$ is the expected number of times of individual i to be selected. Thus, minimum spread is achieved by this selection method. In addition, as individuals are selected entirely on their positions in the population, stochastic universal sampling has zero bias. Stochastic universal sampling ensures a selection of individual which is closer to what is deserved than roulette wheel selection.

Crossover is an essential operator of GA for producing the new individuals. It is a reproduction operator which allows the information exchange between two parent individuals. In single point crossover, a crossover point is randomly generated within a chromosome and resulting subchromosomes are swapped to form two new offsprings. In multi-point crossover, m crossover points are chosen with no duplication. Then, the bits between successive crossover points are interchanged between the two parents. Uniform crossover make every locus a potential crossover point. For each bit of two offsprings, which parent contributes its bit value to which offspring is randomly decided. For each bit position on the two parents, which parent will contribute its value in that position from the other parent is indicated by a crossover mask. A review of the variations of crossover can be found in [49]. Mutation is another genetic operator that introduces variations into the individuals from one generation to the next by flipping

the value of single bits within individuals.

Parallel genetic algorithms have been developed to enhance the capability of GA. There exist three models of parallel genetic algorithms: migration, global and diffusion. In this thesis, migration GA with subpopulations approach is applied in order to search the solution space more effectively and to take full advantage of the intrinsically parallel architecture of GA. The migration GA divides the population into a set of subpopulations, which evolve independently as in the conventional GA. From time to time, a number of individuals are distributed between subpopulations (migration). The number of individuals to be exchanged and the frequency of migration (how many generations pass between successive migrations) are specified by migration rate and migration interval. The migration GA not only enhances the computational speed, but also improve the results when compared to a single population algorithm.

GAs have been widely applied in many fields of signal processing such as adaptive filtering, time-delay estimation, active noise control and so on [50, 51, 52]. Recently, image segmentation algorithms have been developed by adopting GAs [53, 54, 55, 56, 57]. GAs are also applied in some areas of speech and audio processing [58, 59, 60]. Moreover, in the context of video summarization, retrieval and browsing, there has been a growing interest in video segmentation techniques based on genetic algorithm [61, 62].

2.2.4 Entropy Metric

The approximate entropy ($ApEn$) is a family of statistics which was first proposed in [63] to measure the complexity of a time series. $ApEn(m, r, N)$ is approximately the negative natural logarithm of the conditional probability that a dataset of length N , having repeated itself within a tolerance r for m points, will also repeat itself for $m+1$

points [64]. Low value of $ApEn$ is found for a time series containing many repetitive patterns and for a process that is less predictable, a higher value of $ApEn$ is found.

Depending on two parameters of m and r , the approximate entropy of a given time sequence $ApEn(m, r, S_N)$, can be calculated for a time series S_N , consisting of N data points such that $T(1), T(2), \dots, T(N)$. Here, embedding dimension or length of the pattern is defined by m and the criterion of similarity or filter parameter is specified by r . To calculate $ApEn$ of this series, $N - m + 1$ vectors $p_m(i)$ for $\{i \mid 1 \leq i \leq N - m + 1\}$ are first constructed. Here, $p_m(i) = \{T(i+k) : 0 \leq k \leq m-1\}$ is the vector of m data points from $T(i)$ to $T(i+m-1)$. If the difference between any two vectors $p_m(i)$ and $p_m(j)$ in the series S_N is less than r , these vectors are assumed to be similar, i.e., if

$$|T(i+k) - T(j+k)| < r \quad \text{for } 0 \leq k \leq m-1. \quad (2.12)$$

If we consider all vectors of length m [i.e., $p_m(1), p_m(2), \dots, p_m(N-m+1)$] within S_N with the set P_m , then we define

$$C_{im}(r) = \frac{n_{im}(r)}{N - m + 1} \quad (2.13)$$

where $n_{im}(r)$ is the number of vectors in P_m that are similar to $p_m(i)$ within the similar criterion r and $C_{im}(r)$ is the fraction of vectors of length m that are similar to the vector of the same length. Then we define $C_m(r)$, the prevalence of repetitive vectors of length m in S_N , as the mean of $C_{im}(r)$ for all the vectors in P_m . Now we can determine the approximate entropy of S_N , for vectors of length m and within the tolerance r as

$$ApEn(m, r, S_N) = \ln \frac{C_m(r)}{C_{m+1}(r)} \quad (2.14)$$

This can be interpreted as logarithmic of the conditional probability of those vectors, which are close in m points, are also close for $m+1$ points. Usually, the embedding dimension m cannot be large. $ApEn$ strongly depends on the filter parameter r . The

filter parameter is set to be some percentage of the standard deviation of the time series because then $ApEn$ does not depend on the absolute variability or the unit of the signal. If r is very small, noise in the signal affects the $ApEn$ values. When very large r values are used, essential properties of the signal remain undetected. In [65], r is recommended to be chosen as a fixed percentage (often 15% or 20%) of the standard deviation of the sequence with the values for $m = 1$ and $m = 2$.

$ApEn$ has been extensively used in many physiological time series analysis [66, 67, 68]. Since the computational load of the original formulation of $ApEn$ is high to characterize a finite sequence from start to end, windowed version of $ApEn$ is considered in [69] to make it computationally tractable in motion detection problem. In [70], a segmentation method for noisy audio signal is presented using windowed $ApEn$. Furthermore, $ApEn$ is heavily dependent on the length of sequence and lacks consistency. That is, if $ApEn$ of one data set is higher than that of another, it should, but does not, remain higher for all conditions tested [71]. In $ApEn$, to avoid the occurrence of $\ln(0)$ in the calculations, the comparison between the template vector and the rest of the vectors also includes comparison with itself (i.e. self-matching). As a result, a lower value of $ApEn$ is achieved and thus the analyzing sequence is interpreted as more regular than it actually is.

To reduce these shortcomings of $ApEn$, a new and related measure, $SampEn$ has been developed in [72] with a small computational difference. The details of the formulation of $SampEn$ is described in Section 4.3.1. The usefulness of $SampEn$ in noisy speech signal segmentation is also investigated in Chapter 4 of this thesis.

2.2.5 Time Scaling

Time scaling is the process of stretching or compressing a signal. Time scaling techniques are usually employed to shorten or lengthen the time duration of the audio and speech signals (i.e. to accelerate or slow the audio signal). There are a number of applications for time-scale modification of signals in the areas of speech and audio processing such as interactive real time audio processors and PC - based sound editors. Time scaling techniques can also be applied to a musical instrument or an audio effects processor to make effects in film, games, music, DVD and DTV [73, 74].

Time scaling techniques can be classified into time domain and frequency domain based techniques. Time domain based techniques usually suffer artifacts when they are applied to complex signals, or when a large modification factor is used. To avoid these drawbacks, the frequency domain based phase-vocoders are usually applied in the time scale modification of audio and speech signals. Using phase vocoders, time stretching techniques are designed to stretch sound files in their duration without affecting their pitch. Phase vocoder consists of two main parts: an analysis and a synthesis which are usually based on short time Fourier transform. Implementations of phase vocoder have been extensively described in the literature [75, 76, 77]. Fundamental elements of a phase vocoder [78] can be presented as a schematic diagram as shown in Fig. 2.2.

The short time Fourier transform (STFT) of the input signal $x(n)$ (with an analysis window of N samples and hop size of M at the time index n and in the frequency index k , is given by

$$X(n, k) = X_R(n, k) + jX_I(n, k) = |X(n, k)|.e^{j\varphi(n, k)} \quad (2.15)$$

where $|X(n, k)|$ represents the magnitude and $\varphi(n, k)$ is the phase of the time varying spectrum. After transformation by the function F , the following is obtained:

$$\hat{X}(n, k) = \hat{X}_R(n, k) + j\hat{X}_I(n, k) = |\hat{X}(n, k)|.e^{j\hat{\varphi}(n, k)} = F(X(n, k)) \quad (2.16)$$

Figure 2.2: Schematic diagram of a phase vocoder.

where $|\hat{X}(n, k)|$ and $\hat{\varphi}(n, k)$ represent the modified magnitude and phase functions. The function F is a transformation process where a modification of the parameters resulting from the analysis stage can be performed. For example, time scaling can be performed by modifying the derivative of the phase function before synthesis. Lastly, the output signal $y(n)$ is obtained by the inverse Fourier transform (IFFT).

Chapter 3

Speech/Non-Speech Detection

This chapter presents a new method to detect speech/non-speech components of a given noisy signal. Employing the combination of binary Walsh basis functions and an analysis-synthesis scheme, the original noisy speech signal is modified first. From the modified signals, the speech components are distinguished from the non-speech components by using a simple decision scheme. Minimal number of Walsh basis functions to be applied is determined using singular value decomposition (SVD). The main advantages of the proposed method are low computational complexity, less parameters to be adjusted and simple implementation. It is observed that the use of Walsh basis functions makes the proposed algorithm efficiently applicable in real world situations where processing time is crucial. Simulation results indicate that the proposed algorithm achieves high speech and non-speech detection rates while maintaining a low error rate for different noisy conditions.

3.1 Introduction

Speech/non-speech detection is simply the task of discriminating noise-only frames of a signal from its noisy speech frames. In the literature, this process is usually known as voice activity detection (VAD) and it becomes an important problem in many areas of speech processing such as real-time noise reduction for speech enhancement, speech recognition, digital hearing aids and modern telecommunication systems. In multimedia communications, silence compression algorithms are usually applied to reduce the average transmission rate during silence periods of speech. These compression algorithms are also based on speech/silence detection and they allow the speech channel to be shared with other information so that the capacity of channel can be improved. Furthermore, VAD is an essential component in variable rate speech coders to achieve efficient bandwidth reduction without a speech quality degradation. Several methods that trade off the accuracy, delay, perceptual quality and computational complexity have been proposed in the literature to deal with the problem of speech/non-speech detection.

A silence compression speech communication system with VAD was standardized by ITU-T Recommendation G. 729 Annex B [17]. It uses a feature vector consisting of four differential parameters: full-band energy, low-band energy, zero-crossing rate and a spectral measure for the multiboundary decision. The multi-boundary initial voice activity decision is obtained based on the difference between each parameter and its respective long-term average. The initial voice activity decision is set to 1 if one of the 14 boundary decisions in the 4-dimensional space is true. Final decision is made by smoothing the initial decision in four stages. (i.e. hangover scheme) A voice detection algorithm based on a pattern recognition approach and fuzzy logic was proposed for wireless communications in noisy environments [7]. This algorithm uses

the same acoustic parameters adopted by G.729 for feature extraction. Furthermore, in [79], the support vector machine is applied together with the set of parameters used by G.729 to classify the active and inactive frames of speech signal.

A VAD standardized for the GSM cellular communication system is the ETSI speech coder [8]. This adaptive multi-rate speech coder (AMR) specifies two options for VAD to be used in DTX based on spectral estimation and periodicity detection. In AMR option 1, the input signal is divided into sub-bands and the level of signal in each band is calculated. The VAD decision is made by using the outputs from pitch detection, tone detection, complex signal analysis modules and signal level. A hangover scheme is also added before the final decision is made. On the other hand, in AMR option 2 the input signal is first converted into frequency domain. Then, based on the channel energy estimator, channel SNR estimator, spectral deviation estimator, background noise estimator, peak-to-average-ratio module and voice metric calculation module, the VAD decision is made.

Apart from classical voice activity detection methods, as most of which are based on parameters of speech, model based VADs have been introduced recently. Formulating the problem of speech pause detection into a statistical decision theory, two detectors based on maximum a posteriori probability (MAP) and Neyman-Pearson test were described in [80]. A Gaussian statistical model which assumes that the Discrete Fourier Transform (DFT) coefficients of speech and noise are asymptotically independent Gaussian random variables was proposed [18, 19]. Assuming the distributions of speech and noise signals to be Laplacian and Gaussian models, a soft voice activity detector was developed in [20] by decomposing the speech signal into Discrete Cosine Transform (DCT) components.

In this chapter, a method to discriminate the active and inactive periods of speech signals corrupted by unknown type and unknown level of noise is presented. It is

assumed that intervals of the inactive segments can be short as well as long (i.e. while some active segments are located very closely, some active segments may be separated by longer periods). Taking the simplicity of binary Walsh transform as an advantage, the proposed speech/non-speech detection algorithm is developed. First, the signal to be classified is modified employing binary Walsh basis functions. The minimal number of basis functions to be applied is determined by using a technique for the selection of wavelet decomposition at natural scale. Using the statistics of the modified signals, which are highly informative about the characteristics of noisy speech frames as well as noise only frames, classification is performed with a simple decision scheme.

Unlike other VAD methods, in which the decision is made on a frame-by-frame basis, the proposed method instantaneously obtains the sets of consecutive frames as speech and non-speech segments. The effectiveness of the proposed method is evaluated by conducting the objective performance on different types of noise with varying SNRs using the criteria: error rate, speech/non-speech detection rates and false alarm rate. ROC analysis have been used to compare with the standardized algorithms including G.729 and AMR Option 1 and Option 2. Experimental results show that the detection accuracy of the proposed algorithm is high for both speech and non-speech frames regardless of noise levels.

3.2 Proposed Algorithm

The block diagram of the proposed speech/non-speech detection algorithm based on the binary Walsh basis functions is depicted in Fig. 3.1. First, the signal is represented using fast Fourier transform (FFTs). These representations are then modified by Walsh basis functions before reconstructing. The number of basis functions to be applied is determined using SVD. Finally, speech/non-speech periods are detected from the

modified signals utilizing a decision scheme. Details of the algorithm are explained in the following sections.

Figure 3.1: Block diagram of the proposed speech/non-speech detection algorithm.

3.2.1 Transformation of Signals

The noisy input signal is reconstructed as a modified sequence based on an analysis/synthesis scheme described in [77]. Firstly, the input signal $x(n)$ is multiplied by a Hann window to yield successive windowed segments of $x_s(n)$. These window segments are then transformed into the spectral domain by using FFTs. In this manner, a time varying spectrum $X_s(n, k) = |X_s(n, k)|e^{j\varphi(n, k)}$ with $n = 0, 1, \dots, N - 1$ and $k = 0, 1, \dots, N - 1$ for each windowed segment is computed. Here, $X_s(n, k)$ denotes the spectral component of the noisy input signal at frequency index k and time index n . Before synthesis, each s -th windowed segment is modified as the weighted sum of the

magnitude $|X_s(n, k)|$ using binary Walsh basis functions. Using Walsh basis functions, the number of parameters to track the variations between active and inactive regions of the noisy signal can be reduced. In this context, SVD is used to determine the minimal number of Walsh basis functions to be applied. The detailed procedure for the identification of the minimal number of Walsh basis functions is described in the next section. Applying the i -th basis function ϕ_i , a modified sequence, $y_s(n)$, for each windowed segment can be obtained [81] as

$$y_s(n) = \sum_{k=0}^{N-1} |X_s(n, k)| \phi_i(k) \quad (3.1)$$

All the modified segments are then concatenated to obtain an output signal $y(n)$ showing as a time-varying magnitude response.

$$y(n) = \sum_{s=0}^{S-1} y_s(n - sN) \quad (3.2)$$

3.2.2 Determination of Minimal Walsh Basis Function

The Walsh transform is a matrix consisting of a complete orthogonal function set having only two values +1 and -1 over their definition intervals. The motivation for using Walsh transform rather than other transforms is its computational simplicity giving a realistic processing time. The Walsh function of order N can be represented as

$$g(x, u) = \frac{1}{N} \prod_{i=0}^{q-1} (-1)^{b_i(x)b_{q-1-i}(u)} \quad (3.3)$$

where $u = 0, 1, \dots, N - 1$, $N = 2^q$ and $b_i(x)$ is the i -th bit value of x . In this context, Walsh functions are arranged into sequency order, (i.e. the number of zero crossings of Walsh function per definition interval), to obtain a set of basis functions. The number of zero crossings increases with the order of basis functions.

It is important to select the proper basis functions so that the variations between

the dynamics of speech and non-speech can be captured more precisely. A method to select the global natural scale in discrete wavelet transform [82] is adopted to determine the required number of basis functions. This method adaptively detects the optimal scale using SVD technique while decomposition is being carried out.

Consider an input noisy speech signal x of length $\mathcal{V}$,

$$x \triangleq \{x(1), x(2), \dots, x(\mathcal{V})\} \quad (3.4)$$

with $\nu = 1, 2, \dots, \mathcal{V}$; and $y_d(\nu)$ be its modified sequence obtained by applying the basis functions of order d into Eqs. (3.1) and (3.2). Modified sequences $\{y_d(\nu)\}_{d=0}^{D-1}$ can be represented by a matrix P of size $(D \times \mathcal{V})$. To determine the order of basis functions with dominant eigenvalues, the SVD of matrix P is calculated adaptively starting with the first two orders (i.e., ϕ_0 and ϕ_1) while adding the higher orders. In order to determine the number of basis functions to be applied, we study the probability distributions of the order of basis functions with respect to SNRs.

In this analysis, speech signals from TIDIGITS database spoken by male and female speakers were used. If there exist long inter-word silences, they were removed first. Silence segments of different sizes were then inserted to get varying intervals between active regions. To generate the noisy signals, the commonly used white Gaussian noise was artificially added with SNR levels of 20, 10, 5 and 0 dB where SNR is defined as

$$\text{SNR} = 10 \log_{10} \left\{ \frac{\sum_{n=1}^{N_s} s^2(n)}{\sum_{n=1}^{N_n} v^2(n)} \right\} \quad (3.5)$$

where s is speech, v is noise and N_s and N_n are the lengths of speech and noise signals, respectively.

Fig. 3.2 displays the probability distributions of basis function orders, termed as coverage at different levels of SNR. It is observed that dominant eigenvalue is located

only at the first few basis functions. In particular, the minimal order for highly noisy signals of 5 dB and 0 dB is found to be 1. And for the signals at high SNR of 20 dB, 10 dB and clean speech, the dominant eigenvalue is found when the order of basis function is three. Hence, the lower order basis functions of Walsh transform matrix are highly informative and they should be used in modification process. Moreover, it is found that higher order coefficients carry less weight in terms of their magnitude and may not be evident to interpret a large Walsh kernel [83].

In practice, it is not possible to obtain *a priori* information about noise level and noise type. Hence, the proposed algorithm defines the minimal order of basis functions $N_{min} = 3$ throughout the experiments. In the original algorithm of [82], optimal scale is defined as the average of the details from the first level to the natural scale, the level associated with the dominant eigenvalue. However, this averaging may introduce clipping effect for the signals with low level of speech. To avoid this effect, a shifting operator which swaps the right and left halves of the basis function is applied first. Then a good estimate of the binary Walsh basis function at the dominant eigenvalue is defined as

$$\phi_m = \frac{\phi_0 - \sum_{i=1}^{N_{min}} CS(\phi_i)}{\max\{|\phi_0 - \sum_{i=1}^{N_{min}} CS(\phi_i)|\}} \quad (3.6)$$

where $N_{min} = 3$ is the largest order with most prominent eigenvalues and $CS(\cdot)$ is the shifting operator. This new basis function ϕ_m provides sharper representation and higher discriminating features. It is also found that identification between noisy speech periods and noise only narrow intervals become more apparent in the modified sequence developed by ϕ_m .

For length N , the function ϕ_m consists of 1's for $n = 0, \dots, N/2 - 1$ followed by -1's for $n = N/2, \dots, 3N/4 - 1$ and 1's for $n = 3N/4, \dots, N$, where n is the sample

Figure 3.2: The distribution of the order of basis functions for the signals from clean to 0 dB.

index. Substituting the values of ϕ_m in (3.1), we find

$$y_s(n) = \sum_{k=0}^{N/2-1} |X_s(n, k)| + \left(\sum_{k=3N/4}^{N-1} |X_s(n, k)| - \sum_{k=N/2}^{3N/4-1} |X_s(n, k)| \right) \quad (3.7)$$

Since we know that signal information lies between 0 – 2 kHz for the speech spectrum when the sampling frequency is 8 kHz. Therefore, in (3.7), N corresponds to the maximum frequency of 4 kHz, $N/2$ corresponds to 2 kHz and so on. From the expression in (3.7), it can be noticed that the quantity (which may be termed ‘short-term area under the magnitude spectrum’) will be large for 0-2 kHz band (first term) compared to 2-4 kHz band (second term) in the case of noisy speech signal. On the other hand, the corresponding quantity will be small for the noise only case due to the spreadness of the noise spectrum.

In order to compare ϕ_m with ϕ_o , we replace ϕ_i with ϕ_o and rewrite (3.1) as

$$y_s(n) = \sum_{k=0}^{N-1} |X_s(n, k)| \quad (3.8)$$

Using (3.8), the difference between the ‘short-term area under the magnitude spectrum’

Figure 3.3: The clean signal.

for the noisy speech case and the noise only case (specially for white Gaussian noise) will be less due to the sum taken over the whole 0-4 kHz frequency band. Based on the expressions of (3.7) and (3.8), we notice that the discrimination between speech and non-speech segments will be higher using ϕ_m compared to ϕ_o .

Figure 3.4: The noisy signal.

To demonstrate the effectiveness of the proposed modification presented above, an

Figure 3.5: The modified signal using 0-order basis function ϕ_0 .

example is also shown in Figs. 3.3-3.6. A clean signal and the corresponding noisy signal in white Gaussian noise at 5 dB SNR are shown in Figs. 3.3 and 3.4. The modified version of the noisy signal using 0-order basis function ϕ_0 and optimal basis function ϕ_m are also shown in Figs. 3.5 and 3.6, respectively. It is observed from Figs. 3.5 and 3.6 that discriminating ability of the modified signal y_m as obtained using ϕ_m is higher for the speech and non-speech frames due to its deeper and sharper waveform. It seems that the function ϕ_m is more efficient to capture the *intra-segment* variation between noisy speech and noise only segments of narrow interval. Noted that we have illustrated, in this section, the advantage of using the ϕ_m function through a simulation example. In Section 3.3, the performance of basis functions ϕ_0 and ϕ_m will be compared for different types and levels of noise.

Without shifting operation(CS) in Eq. 3.6, the basis function ϕ_m of length N will consist of -1's for $n = 0, \dots, N/4 - 1$ followed by 1's for $n = N/4, \dots, N$ where n is the sample index. Figs. 3.7 - 3.9 show the effect of shifting the basis functions in reconstructing a signal. First, a clean signal is shown in Fig. 3.7. Reconstructed versions of this signal at 5 dB SNR employing ϕ_m with and without shifting operator

Figure 3.6: The modified signal using basis function ϕ_m .

are shown in Figs. 3.8 and 3.9. While the locations of segment boundaries are revealed in Fig. 3.8, mid speech clipping is found in the reconstructed signal shown in Fig 3.9. In these figures it can be seen that clipping can be avoided by this shifting operation, thereby reconstruction of signal using shifted basis functions can improve the discriminability between speech and non-speech.

3.2.3 Decision Scheme

First, 0-order basis function, ϕ_0 is used to produce a modified sequence, $y_0(\nu)$, to get the global information of the original noisy signal. This modified sequence is used as a reference or pilot signal as in the area of telecommunication. In telecommunication, a pilot signal is usually transmitted over a communication system for supervisory, control or reference purposes. Carrying the local characteristics, another modified signal, $y_m(\nu)$, is formed using the new basis function ϕ_m . From this sequence, locations and durations of speech active and inactive periods can be captured more closely. In this way, the approximate locations of active and inactive frames are first determined from

Figure 3.7: A clean signal.

the modified signal, $y_0(\nu)$. Then, the accuracy of these reference decisions are improved by using the second modified signal, $y_m(\nu)$, containing the detailed information. Applying the reconstructed signals y_0 and y_m , the procedure of detection scheme can be described as follows.

- Extract the series of local minima, $\{\alpha_{0i}\}_{i=1}^L$ and $\{\alpha_{mi}\}_{i=1}^L$, where L is the number of frames, from every 4 ms frame of $y_0(\nu)$ and $y_m(\nu)$ for which it is assumed that the initial 200 ms consists of noise only periods.
- Set the thresholds for each waveform by using statistics: $\tau_0 = \mu_0 - \kappa\delta_0$ and $\tau_m = \mu_m - \kappa\delta_m$, where μ_0 and δ_0 are the mean and the standard deviation of the first local minimum set and μ_m and δ_m are those of the second local minimum set and κ is a positive value. After examining the modified waveforms of a number of clean as well as noisy speech signals, κ is set to be 0.75.
- A frame can be considered as inactive if either $\alpha_{0i} < \tau_0$ or $\alpha_{mi} < \tau_m$. In this way, the indices of non-active frames are obtained from $y_0(\nu)$ and $y_m(\nu)$ as $\mathcal{R}$ and $\mathcal{T}$:

$$\mathcal{R} = \{r_1, r_2, \dots, r_P\} \quad (3.9)$$

Figure 3.8: The reconstructed signal using shifting operator CS .

$$\mathcal{T} = \{t_1, t_2, \dots, t_Q\} \quad (3.10)$$

where r_i, t_i are the inactive frames detected and P and Q are their total number of frames that are detected as inactive from the first and second modified signals, respectively.

- Combine the two initial boundary decisions as follows:

$$\mathcal{C} = \mathcal{R} \cap \mathcal{T} \quad (3.11)$$

where $\mathcal{C} = \{c_1, c_2, \dots, c_J\}$ is the set of elements common to $\mathcal{R}$ and $\mathcal{T}$. By concluding the members of $\mathcal{C}$ as inactive frames, final decisions for the speech and non-speech frames are acquired.

In the above, we assume that there exist inactive regions when some or all of the potential local minima found in the first modified signal coincide with the local minima found in the second modified signal. The detected frames, which are not contained in both modifications, are attributed to be noise spikes and discarded.

Figure 3.9: The reconstructed signal without using shifting operator CS .

3.3 Experimental Study and Comparison

In this section, the results and objective evaluation of the proposed method is presented. The detection result for a noisy speech signal is illustrated in Fig. 3.10 as an example. This signal is embedded in white Gaussian noise at 0 dB SNR. The results estimated by the proposed detection method are shown together with real speech and non-speech which were determined manually. It is seen that the detection accuracy of the proposed method is high for both speech and non-speech periods. An thus a good performance level is achieved.

To evaluate the efficiency of the proposed method, its performance was compared with G.729 VAD and AMR Options 1 and 2. For comparison purpose, the speech signals from eleven speakers of TIDIGITS database were extracted. Three signals from each of these male and female speakers were concatenated to generate the signals of 8 s to 11 s long. The sampling frequency was 8 kHz. Silence or pause segments of varying intervals were then introduced between the active segments as described in Section 3.2.2. Test sequences composed of about 70% of active speech components

Figure 3.10: (a) Noisy speech at $\text{SNR} = 0$ dB in white Gaussian noise (b) manual detection (c) estimated detection.

and 30% of inactive speech components. The silence segments of very short as well as long durations are included in the test sequences. For reference decisions, active and inactive frames of all clean signals were marked manually. Five types of noise: white Gaussian, babble, car, street and train were added to the original signals with different SNRs (20 dB, 10 dB and 0 dB).

As the performance criteria, speech detection rate, non-speech detection rate and error rate were employed. Speech and non-speech detection rates are defined as correctly classified speech frames to the total number of speech frames and correctly classified non-speech frames to the total number of non-speech frames. Error rate is defined as the ratio of incorrectly classified frames to the total number of frames. In Table 3.1, speech/non-speech detection rates and error rates of the proposed method are compared to the standardized VADs: G.729, AMR Options 1 and 2 under different noise sources and SNR levels.

Speech detection accuracy of ITU G.729, ETSI AMR1 and AMR2 decreases with increasing noise levels in all noise types. Proposed binary Walsh transform based

Table 3.1: Comparison of speech detection rates, non-speech detection rates and error rates of proposed method to standard methods, G.729, AMR1 and AMR2 for different levels of SNRs in various noisy environments.

		Speech Detection $DS(\%)$				Non-Speech Detection $DNS(\%)$				Error Rate $E(\%)$			
Noise	SNR	Proposed	G.729	AMR1	AMR2	Proposed	G.729	AMR1	AMR2	Proposed	G.729	AMR1	AMR2
White	20dB	89.20	96.79	96.26	97.07	95.48	31.51	61.09	48.21	9.81	20.85	12.41	15.56
	10dB	88.48	90.42	93.03	92.01	95.13	42.21	45.11	52.52	10.53	22.74	18.68	18.12
	0dB	87.07	67.09	81.32	60.57	81.26	62.37	56.98	77.97	15.97	34.72	24.72	35.49
Car	20dB	88.76	97.65	97.84	98.06	96.40	19.19	62.61	45.95	9.76	23.55	11.04	15.62
	10dB	88.01	95.42	96.36	93.64	92.47	17.04	51.21	50.31	11.74	25.59	15.30	17.76
	0dB	87.37	91.55	81.02	64.46	70.35	16.57	55.53	70.50	18.28	28.67	26.10	34.92
Babble	20dB	88.34	97.02	98.20	97.82	95.45	19.60	56.84	42.51	10.33	23.84	12.32	17.17
	10dB	89.11	93.85	98.44	95.28	84.44	18.58	29.09	40.81	13.48	26.91	19.81	19.69
	0dB	86.19	90.46	90.85	85.87	56.32	14.44	31.02	37.46	22.74	29.99	25.04	27.23
Street	20dB	88.55	96.41	97.33	98.37	95.20	21.85	66.16	47.36	10.31	23.90	10.45	15.07
	10dB	89.60	92.49	97.36	93.12	83.51	17.28	45.98	51.95	12.95	27.75	15.85	17.79
	0dB	84.51	88.81	86.80	69.22	65.61	13.75	46.46	67.87	21.55	31.26	23.89	31.71
Train	20dB	88.86	97.22	97.20	98.66	95.85	23.47	67.40	50.69	9.84	22.91	10.20	13.85
	10dB	88.10	93.47	96.44	96.08	92.40	25.50	60.08	54.42	11.50	24.66	12.68	14.81
	0dB	84.83	90.92	86.10	78.88	82.87	14.22	62.16	70.22	16.75	29.65	19.91	23.89

method can consistently detect the speech frames with almost constant rate regardless of noise types and levels. Considering the non-speech detection rates, G.729 is the worst with an accuracy of less than 20% for most of the times. Although AMR1 and AMR2 yield better detection rate than G.729, the proposed method is found to be the best one in the problem of non-speech detection for all noise conditions. Moreover, the proposed method can detect both speech and non-speech frames with least error probabilities for all levels of SNRs in all environments. Furthermore, most of noise only segments of short durations are only detected by the proposed algorithm.

The results of the performance comparisons for average rates of speech detection, non-speech detection and error of the proposed method to ITU G.729, AMR Options 1 and 2 in five background noise (white, babble, car, street and train) and SNR ranging from 20 dB to 0 dB are provided in Figs. 3.11, 3.12 and 3.13. Average speech detection

Figure 3.11: Performance comparison for average non-speech detection rate of the proposed method and standard VADs (G.729, AMR1 and AMR2) in different backgrounds with varying SNRs.

rates of the proposed method is nearly constant for varying SNRs of 20 dB, 10 dB and 0 dB having respective values of 88.74%, 88.66% and 85.99%. Although the speech detection rates of above standardized methods are high in 20 dB, their performance is decreased with decreasing SNRs. In terms of non-speech detection rates, G.729 yields the lowest rates followed by AMR1. The non-speech hit rates of the proposed algorithm are the highest although AMR2 achieves improved rates over G.729 and AMR1. The proposed method yields the significant lowest error rates (10.00%, 12.04% and 19.05%) for SNRs of 20 dB down to 0 dB. Error rates of AMR2 is found to be dependent on the noise levels although it offers moderate non-speech detection rates over G.729 and AMR1.

Moreover, we have studied the discrimination ability of the basis functions ϕ_0 and ϕ_m for speech dataset described above. Table 3.2 compares the proposed method by using the basis functions ϕ_0 and ϕ_m individually in terms of speech hit rate, non-speech hit rate and error rate. At low level of SNR, non-speech detection rate of ϕ_m is superior to ϕ_0 for almost all types of noise. However, at high level of SNRs, the

Figure 3.12: Performance comparison for average speech detection rate of the proposed method and standard VADs (G.729, AMR1 and AMR2) in different backgrounds with varying SNRs.

difference between non-speech detection rate obtained by ϕ_m and ϕ_0 is not significant. For speech detection, nearly equal rates are obtained by both basis functions for all levels of SNRs in most background noises. From the Tables 3.1 and 3.2, it is found that VAD results of the proposed method improve the speech detection accuracy with the lowest error rates when both basis functions are employed. Again, non-speech detection rate of the proposed method is higher than other standardized methods.

3.4 Receiver Operating Characteristics Analysis

In this section, the detectability and discriminability of the proposed method is verified in terms of receiver operating characteristics (ROC) analysis. In the signal detection area, the relationship between detection and false alarm probabilities are often characterized by ROC curves. Only the subset of speech database in white Gaussian noise as described in Section 3.3 has been used in this ROC analysis. Figs. 3.14, 3.15 and 3.16 show the result of ROC analysis at 20 dB, 10 dB and 0 dB SNRs. For each noise

Figure 3.13: Performance comparison for average error rate of the proposed method and standard VADs (G.729, AMR1 and AMR2) in different backgrounds with varying SNRs.

level, non-speech hit rate (non-speech detection rate) and false alarm rate ($1 - \text{speech detection rate}$) are determined over the proposed method, G.729, ETSI AMR1 and AMR2. Since each algorithm produces a binary decision, only a single operating point is generated in ROC space. The operating points of G.729, AMR1 and AMR2 shift to the right in ROC space with decreasing SNRs. However, the operating point of the proposed method can maintain an almost constant false alarm rate.

False alarm rates of G.729 increases with decreasing SNR although it achieves a lower non-speech hit rates across varying SNRs. Among these standard VADs, ETSI AMR1 maintains the lowest false alarm rate. However, it has poor non-speech hit rate. While working with lower false alarm rate than G.729, detectability of non-speech in AMR2 is better than AMR 1. Obviously, the proposed method yields 10-40% improvement in non-speech hit rate over other methods with the lowest false alarm rates in all instances. For a given non-speech hit rate, the proposed scheme can detect the signal with the lowest false alarm rate. In addition, for a given false alarm rate, the highest non-speech hit rate can be yielded by our method. From this objective

Table 3.2: Comparison of the proposed method using basis functions ϕ_0 and ϕ_m for different levels of SNRs in various noisy environments.

Noise	SNR	ϕ_0			ϕ_m		
		$DS(\%)$	$DNS(\%)$	$E(\%)$	$DS(\%)$	$DNS(\%)$	$E(\%)$
White	20 dB	86.42	97.98	10.78	85.95	98.07	11.11
	10 dB	84.92	97.63	11.97	84.67	97.32	12.24
	0 dB	82.63	88.21	16.51	81.35	92.17	16.21
Car	20 dB	86.33	98.03	10.86	85.59	98.22	11.35
	10 dB	86.65	94.85	11.67	84.87	96.82	12.32
	0 dB	90.52	58.07	19.62	85.18	77.24	17.79
Babble	20 dB	86.84	97.88	10.55	86.10	98.14	10.99
	10 dB	86.90	91.90	12.32	85.85	89.94	13.62
	0 dB	83.51	60.79	22.95	83.03	61.83	23.09
Street	20 dB	87.11	97.44	10.49	86.16	97.21	11.29
	10 dB	87.44	87.53	13.11	87.01	83.81	14.45
	0 dB	77.71	70.87	24.14	79.22	68.96	23.51
Train	20 dB	86.74	97.98	10.58	85.65	98.22	11.30
	10 dB	85.11	97.47	11.95	84.14	96.64	12.91
	0 dB	80.91	77.41	20.65	81.31	84.72	18.25

evaluation, it can be concluded that discriminability of the proposed method in speech with noise is superior to all of these standardized methods.

3.5 Conclusion

In this chapter, a speech/non-speech detection problem in the presence of noise is considered. A method, which is based on the binary Walsh basis functions is developed. The basic idea is to reconstruct the noisy speech signal as modified sequences from which speech and non-speech frames are detected. The main advantage of this method is its very low computational complexity. The Walsh basis functions make the proposed algorithm efficient, simple, fewer parameters to be optimized and faster in implementation. Thus the algorithm is applicable in practical situations where processing time is critical. Experimental results indicate that the proposed method can detect speech

Figure 3.14: Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1 and AMR2 at 20 dB white Gaussian noise.

as well as non-speech frames with lower error rates across different types of noise with varying SNRs. ROC analysis also shows that the proposed method consistently outperforms G.729, AMR1 and AMR2 in terms of discriminability between speech and noise. Since the computational complexity of the algorithm is relatively low, the algorithm can be applied in the areas such as real time noise cancellation systems and noise reduction for enhancement of speech signals.

Figure 3.15: Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1 and AMR2 at 10 dB white Gaussian noise.

Figure 3.16: Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1 and AMR2 at 0 dB white Gaussian noise.

Chapter 4

Segmentation Based on Genetic Algorithm

A new approach to automatically segment speech signals in noisy environments is described in this chapter. Speech segmentation is formulated as an optimization problem and boundaries of the speech segments are detected using a genetic algorithm (GA). The number of segments present in a segmenting signal is initially estimated from the reconstructed sequence of the original signal using the minimal number of binary Walsh basis functions. A multi-population GA is then employed to determine the locations of segment boundaries. The segmentation results are improved through the generations of GA by introducing a new evaluation function, which is based on the sample entropy and a heterogeneity measure. The results of the experiments, which have been carried out on TIDIGITS database demonstrate that the proposed genetic segmentation algorithm can accurately detect the noisy speech segments as well as noise only segments under various types and levels of noise.

4.1 Introduction

Segmentation is the process of partitioning a signal into its homogeneous units. Automatic segmentation is a crucial part in many areas of speech and audio processing such as automatic speech recognition(ASR), transcription, classification of audio visual data, indexing, segmenting of broadcast news and speech/music discrimination. Usually, the input data to these systems consist of continuous audio stream which is needed to be segmented for further level of processing. Since manual segmentation is tedious and time consuming, it is unrealistic for real applications where the systems have to deal with larger amount of data.

In general, speech segmentation can be divided into two main categories; phoneme unit segmentation and syllabic unit segmentation. Phonemes are basic unit of the human speech having varying acoustic values while syllables are perceptually more meaningful units having relatively constant values in a word. Syllable units are used as basic components in speech recognition [84, 85] and syllabic segmentation algorithms are proposed in [86, 87]. Approaches for automatic phonetic segmentation of spoken speech are described in [88, 89, 90].

Various speech segmentation algorithms for word boundary detection, syllable segmentation and phoneme segmentation have been addressed in the literature. According to [91], these algorithms can be broadly categorized into rule-based or metric based segmentation and classifier-based segmentation. Rule-based methods usually employ the acoustic properties of the signal. Energy and zero-crossing rates are most widely used parameters of the signal [92, 93]. Mel frequency cepstrum coefficients (MFCC) are also applied in rule-based segmentation algorithms [25]. Following the feature extraction, rules relating to these acoustic properties are usually derived heuristically.

Formulating speech segmentation as a classification problem, classifier-based meth-

ods use two separate models for speech and non-speech events. Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) are commonly used models [94, 95]. In the rule-based methods, fine tuning is always necessary for the rules and the parameters applied to these rules. In general, the segmentation accuracy of classifier-based methods is better than rule-based methods. However, classifier-based segmentation methods are computationally complex and they need extensive training time to achieve reasonable good results for different acoustic conditions. For example, GMM-based algorithms are very efficient when training condition matches with testing condition.

Noise is one of the most prominent source of errors in speech processing applications such as ASR and telecommunication. Usually, these application systems are designed to work in controlled conditions. Thus acceptable performances are obtained when they are applied in noise free situations. However, these performances degrade dramatically when they are applied in noisy environments. When a speech processing system is to be applied in noisy environments, it must be robust against various noise types in wide range of SNRs. Although there have been many segmentation methods described, segmentation algorithms which are efficient and robust to different types and levels of noise with low computational load are still required.

In this chapter, a method to detect spoken words of a speech signal which are corrupted by unknown level and type of noise is considered. The segmentation method described in this chapter is based on the genetic algorithm, which is a stochastic global search method. Using the ability of GA which can utilize parallel exploration of the search space, the proposed method can reduce the possibility of being stuck in local optima. The algorithm of proposed GA based speech segmentation can be divided into two stages. The algorithm first reconstructs the original signal into a modified sequence using minimal binary Walsh basis function. Obtaining the local and global variations from the reconstructed signal, the number of segments (i.e. words) present

in the segmenting signal is determined using *mean difference measure* [96].

When the signals are contaminated by high level of background noise, the problem of segmentation becomes hard and thus it is difficult to locate the boundaries of speech and non-speech events accurately. To deal with this problem, a genetic algorithm is employed in the second stage to segment the noisy speech signal into its homogeneous regions. In GAs, the search directions are influenced by the evaluation function and its corresponding fitness value. Hence, an evaluation function measuring the regularity and homogeneity of speech segments is designed using a metric called *sample entropy* [72] and *heterogeneity* [97]. Through the generations of GA, locations of segment boundaries are optimized. The performance of proposed GA based speech segmentation method is evaluated on the signals from TIDIGITS database with a sampling frequency of 8000 Hz. The experimental results show that the proposed method can detect the boundaries of speech and non-speech events with high accuracy in various noise conditions.

4.2 Determination of Number of Segments

In this speech segmentation algorithm, the number of segments present in a signal is estimated from its reconstructed sequence. In order to modify the original signal, an analysis and synthesis scheme and a set of basis functions are employed. By using the basis functions, the identification of a highly varying signal such as speech can be controlled more properly. Moreover, the number of parameters in tracking along the variations between active and inactive regions of noisy signal can be lessened.

In this context, binary Walsh basis functions are selected for modification since they are computationally simple and they can be implemented quickly. The analysis and synthesis scheme and optimal basis function (ϕ_m) which have been described in Section 3.2.1 and Section 3.2.2 are used to reconstruct the original signal. The approx-

imate number of segments present in a given signal is obtained from this modified or reconstructed signal applying the *mean difference measure*.

4.2.1 Mean Difference Measure

From the modified sequence as obtained above, the input to GA (i.e. the number of segments) is determined using mean difference measure described in [96]. In this approach, two adjacent windows of equal length are moved through the modified signal. At each position, the magnitude of the difference of the means within each window is calculated. This mean differences sequence is thresholded to determine the local maxima of a certain value. Local maxima having a width greater than a specified value are taken as segments. All the other maxima which do not meet these conditions are discarded. Finally, the total number of significant local maxima which satisfy both requirements are assumed to correspond to the number of segments present in the segmenting signal. Here, we have used the width of window as 62.5 ms (500 samples).

4.3 Segmenting by a Genetic Algorithm

As second step, the locations of the segment boundaries are detected using a genetic algorithm. Depending on the total number of segments estimated from the first step as discussed in the previous section, an initial population is randomly generated. To guide the search space of GA, a new evaluation function is introduced. First the irregularity in the time series (i.e. modified sequence) is investigated using *sample entropy* (*SampEn*). Measuring the *homogeneity* and *heterogeneity* of the candidate segments, the fitness of the evaluation function is designed. Through the generations of GA, the locations of segment boundaries are optimized.

4.3.1 Sample Entropy

In this GA based segmentation method, the boundaries of speech segments (spoken words) are determined using a similarity measure of time series, i.e, the values of sample entropy. The changes in speech signals are therefore determined from sample entropy sequence. Sample entropy (*SampEn*) is developed in [72] to measure the complexity and regularity of clinical and experimental time series data similarity.

The origin of sample entropy is the *approximate entropy* (*ApEn*), which is introduced in [71] to measure the regularity in time series. $ApEn(m, r, N)$ is defined as the negative natural logarithm of the conditional probability that a data set of length N , having repeated itself within a tolerance r for m points, will also repeat itself for $m + 1$ points. Small values of *ApEn* indicates a high regularity in time series while large values of *ApEn* implies that the time series is irregular. Recently, *ApEn* is applied to analyze the time series of clinical data [98, 99]. Since the *ApEn* algorithm counts self-matching, it (i) lacks of relative consistency and is (ii) heavily dependent on the record length.

To reduce the bias and inconsistent results caused by self-matching, sample entropy (*SampEn*) that does not count self-matches is developed in [72]. $SampEn(m, r, N)$ is also defined as the negative natural logarithm of the conditional probability that a data set of length N , having repeated itself within a tolerance r for m points, will also repeat itself for $m + 1$ points, without allowing self-matches. Thus, a low value of *SampEn* reflects a high degree of self-similarity in a time series. In [64], the dynamics of neonatal heart rate variability is investigated using *SampEn*.

For an input signal u of length N , $\{u(j) : 1 \leq j \leq N\}$ forms the $N - m + 1$ vectors $x_m(i)$ for $\{i | 1 \leq i \leq N - m + 1\}$, where $x_m(i) = \{u(i + k) : 0 \leq k \leq m - 1\}$ is the vector of m data points from $u(i)$ to $u(i + m - 1)$. In this context, only the first $N - m$

vectors of length m are considered to ensure that, $x_m(i)$ and $x_{m+1}(i)$ are defined for $1 \leq i \leq N - m$. Let $B^m(r)$ be the probability that two sequences will match for m points and $A^m(r)$ be the probability that two sequences will match for $m + 1$ points. $B_i^m(r)$ is defined as $(N - m - 1)^{-1}$ times the numbers of vectors $x_m(j)$ within r of $x_m(i)$, where $1 \leq j \leq N - m$, and $j \neq i$ to exclude self-matches. Then $B^m(r)$ is defined as

$$B^m(r) = (N - m)^{-1} \sum_{i=1}^{N-1} B_i^m(r) \quad (4.1)$$

Similarly, $A_i^m(r)$ is defined as $(N - m - 1)^{-1}$ times the numbers of vectors $x_{m+1}(j)$ within r of $x_{m+1}(i)$, where $1 \leq j \leq N - m$ and $j \neq i$. Then set $A^m(r)$ as

$$A^m(r) = (N - m)^{-1} \sum_{i=1}^{N-1} A_i^m(r) \quad (4.2)$$

Finally, sample entropy ($SampEn$) is calculated by

$$SampEn(m, r, N) = -\ln \frac{A^m(r)}{B^m(r)} \quad (4.3)$$

$SampEn$ measures the regularity of data sequence. A low value of $SampEn$ indicates that the sequence is regular. With increasing irregularity, a larger value of $SampEn$ is obtained. Fig. 4.1 (b) illustrates a sequence of sample entropy calculated for a clean speech shown in Fig. 4.1 (a). For non-speech portions, the value of $SampEn$ is the smallest. Then $SampEn$ increases for speech segments and decreases during the appearance of non-speech regions. Fig. 4.2 also shows the analysis of $SampEn$ for a speech signal in a noisy background. The plots of the clean signal and the noisy signal at 5 dB SNR are shown in Fig. 4.2 (a) and (b). The sample entropy sequence calculated for noisy signal is shown in the Fig. 4.2 (c). The value of $SampEn$ decreases at the locations of non-speech segments as in the clean speech case. From these results, it is observed that $SampEn$ of speech signal is increased before noisy speech components and decreased quickly thereafter. Hence, the dynamics of segmenting speech signal

can be investigated through the sample entropy sequence. And sample entropy can be applied as a useful tool to determine the locations of the noisy segments as well as noise only segments for a noisy speech data.

Figure 4.1: The analysis of sample entropy in a clean speech signal.

Figure 4.2: The analysis of sample entropy in a noisy speech signal.

4.3.2 Genetic Algorithm

Genetic algorithms (GAs) are numerical optimization algorithms inspired by both natural selection and natural genetics [100]. Instead of single solution, GAs operate on a

population of chromosomes, that is, a group of potential solutions of a problem. To measure how good or bad the solutions within the population, fitness of each chromosome is calculated applying an evaluation function. At each generation, a new set of solutions are produced by selecting the fittest chromosomes in the problem domain and through the application of the genetic operators such as crossover and mutation. A review of the fundamental operations of a simple GA can be found in [101]. The procedure of a simple GA can be described as follows, where the population of candidate solutions at time t is represented by $P(t)$:

```

begin
     $t = 0$ ;
    initialize  $P(t)$ ;
    while not termination criteria do
        begin
             $t = t + 1$ ;
            select  $P(t)$  from  $P(t-1)$ ;
            reproduce pairs in  $P(t)$ ;
            evaluate  $P(t)$ ;
        end
    end
end

```

Initial Population

The number of potential local minima, determined from Section 4.2 is considered as the number of segments present in a given signal. In order to detect both start and end locations of each segment, a population of GA is generated with chromosomes

whose length is two times the total number of segments obtained above. Although the binary-coded GAs are the most commonly used representation, a more natural real-valued representation is used in this system to increase the efficiency of GA. Using the real-valued chromosomes, there is no need to convert chromosomes to phenotypes before fitness function evaluation and thus it would be faster in computation.

Evaluation Function

In GAs, an evaluation function or fitness function is usually used to evaluate the performance of the chromosomes in the problem domain. In order to obtain accurate boundaries of each segment, the evaluation function is designed using the heterogeneity measure and sample entropy. This function simultaneously maximizes the homogeneity within the segments and heterogeneity among different segments using sample entropy.

In this context, *SampEn* of the original segmenting signal is calculated first to investigate the dynamics. To prevent the requirement of heavy computational time (i.e. to obtain a feasible computation time and to make the proposed algorithm tractable), *SampEn* is calculated on each data set of length 80 (i.e. $N=80$) within a tolerance r of $0.1 \times SD$ for 1 point (i.e. $m=1$). Here, SD is the standard deviation of the data set. Let H_w be the total within-segment heterogeneity and H_b denotes the total between-segment heterogeneity, a segmentation evaluation function is defined as

$$H = \frac{H_b + 1}{H_b + H_w + 1} \quad (4.4)$$

where total within-heterogeneity H_w is defined as

$$H_w = \frac{\sum_{i=1}^S L_i \sigma_i^2}{L} \quad (4.5)$$

where L is the total length of the segmented signal, L_i is the length of the i -th segment, σ_i^2 is the variance of the sample entropy of i -th segment and S is the number of segments

in the segmented signal. The between-segment heterogeneity, H_b , is defined as the average Euclidean distance between the mean value of the sample entropy of any two adjacent segments.

$$H_b = \frac{\sum_{(i,j) \in \text{adjacent}, i \neq j} \|\mu_i - \mu_j\|^2}{ns} \quad (4.6)$$

where ns is the total number of the adjacent segments in the segmented signal, μ_i and μ_j are the mean value of the sample entropy of the i -th and j -th segments. H reaches one when the internals of all segmented speech are completely homogeneous.

Evolution Procedure

In order to effectively search the solution space, and to take advantage of the parallelism of GAs, the proposed algorithm applies a multiple subpopulations approach provided by [49] for the evolutionary process. A problem with simple or sequential GA is premature convergence on a suboptimal solution. Parallel GAs can search in parallel different subspaces of the search space, thus making it less likely to become trapped by low-quality subspaces. Multiple populations GA is a widely used parallel GA model where multiple subpopulations evolve independently toward different optima. More diverse subpopulations can be maintained by exchanging genetic materials between subpopulations. The premature convergence effect of simple GA can then be mitigated by this approach. To reduce the required computational time, it is implemented through the use of high-level genetic operator functions and exchanging individuals between subpopulations as described in Section 2.2.3.

Over the generations, each subpopulation is evolved as in traditional simple genetic algorithm (SGA) using the basic operators: *crossover* and *mutation*. Depending on the migration interval (i.e. the number of generations between successive migration) and the migration rate (i.e. the number of individuals to be migrated from one sub-

population to another), individuals from one subpopulation migrate to another from time to time. The initial population is created using 8 subpopulations containing 20 individuals each. At each generation, 90% of the individuals with higher fitness values within each subpopulation are selected for breeding using a *stochastic universal sampling* function which has minimum spread and zero bias.

In GAs, a recombination operator is usually used to produce the new offsprings. By applying *discrete recombination crossover*, a uniform crossover for real-valued representation, the new offsprings within each subpopulation are produced. Normally, offsprings are mutated after recombination to prevent the population from converging to local minima. And new possible solutions can be introduced to the population by mutating the offsprings. In this system, a mutation rate of $1/nvar$ is used, where $nvar$ is the length of an individual.

When the offsprings produced are less than the size of the original population, the new offsprings have to be reinserted into the population to maintain the size of the original population. Similarly, when not all the offsprings are to be used at each generation, or if the offsprings produced are more than necessary, a reinsertion scheme must be used. This scheme determines which individuals should be replaced by the offsprings produced and which individuals should be inserted into the new population. In this segmentation method, offsprings are inserted into the appropriate subpopulations depending on *fitness-based reinsertion* with a rate of 0.9.

In this multi-population GAs, migration of individuals between subpopulations is performed at every 20 generations with a migration rate of 0.2. After GA iterates for $maxgen$ times (here $maxgen=80$), the evolution of this GA stops. The best individual with the maximum fitness value presents the optimized solution for boundaries of the segments of the segmented signal.

4.4 Experimental Results

To evaluate the performance of the proposed GA based segmentation, the experiments are carried out on the speech signals of 20 speakers (10 male and 10 female) from the TIDIGITS database. 10 digit strings of lengths varying from 3 to 7 digits have been considered from each speaker. If there exist long inter-word silences, they are extracted first. For reference decisions, the boundaries of each segment are manually determined. White noise, babble noise, car noise, street noise and train noise are then added to obtain the corrupted signals at different SNRs (20 dB, 15 dB, 10 dB, 5 dB and 0 dB). Therefore, a total of 5000 signals sampled at 8 kHz are applied to the proposed algorithm.

Figure 4.3: (a) Noisy input signal at 5 dB SNR; (b) segmentation results shown in vertical lines together with the clean speech signal.

Figs. 4.3 and 4.4 demonstrate the segmentation results of input noisy speech signals. The speech signals in white Gaussian noise at 5 dB and 0 dB SNR are shown in Figs. 4.3(a) and 4.4(a), respectively. The corresponding clean signals are displayed in Figs. 4.3(b) and 4.4(b). The detected boundaries of each segment are also shown in Figs. 4.3(b) and 4.4(b) by the vertical dashed line. It is found that the difference between

the deviation from the boundaries of segments in signal of 5 dB SNR and from that of the signal in 0 dB SNR is not significant. This indicates that the proposed GA based segmentation algorithm is robust against noise and can yield satisfactory results.

Figure 4.4: (a) Noisy input signal at 0 dB SNR; (b) segmentation results shown in dashed lines together with the clean speech signal.

When there is no inter-word silences between consecutive digits, the proposed method can detect them as one segment only. To tackle this problem, durational information of digits, studied in [87], is applied. When a detected segment has a duration greater than 390 ms, the mean duration of digits spoken by male speakers, further segmentation is performed to find another segment within this interval. Similarly, if the duration of a segment is less than 150 ms, it is merged to one of its neighboring segments having shorter duration. The overall improvement is achieved by including the durational information although it becomes the limitation of the proposed algorithm. We have also investigated the algorithm on the TIMIT database without using durational information. The algorithm can detect most of the boundaries except for some spurious boundaries. Here, the difference between the actual and estimated number of boundaries may be due to the algorithm, which defines each segment with start and end boundaries. However, there might exist common boundaries for some of the

Table 4.1: Relative error at different SNRs.

	Noise Types				
SNR	White	Car	Babble	Street	Train
20 dB	0.0858	0.0983	0.1077	0.0963	0.0983
15 dB	0.1088	0.1194	0.1314	0.1228	0.1165
10 dB	0.1378	0.1512	0.1737	0.1523	0.1417
5 dB	0.1753	0.1883	0.1990	0.1936	0.1800
0 dB	0.2092	0.2332	0.2553	0.2316	0.2222

connected words.

4.4.1 Evaluation Metrics

Relative Error

As a performance measure, *relative error* of a segment RE is calculated as follows:

$$RE = \frac{|AD - ED|}{AD} \quad (4.7)$$

where AD is the duration of the segment which is determined manually and ED is duration of the segment estimated by the proposed method. In Table 4.1, the results for the performance measure of the proposed segmentation algorithm for different types of background noise and varying SNR are given. From this table, it can be seen that nearly 96% of the segments are detected within 30 ms of manually determined boundaries. Note that a high degree of accuracy is achieved for speech segments in varying environmental conditions.

When the SNR is 20 dB, the average segmentation error is almost less than 10% of the duration of segments for all noise types. The segment boundaries determined by the proposed algorithms deviate from 10% and 13% of the manual segmentations

when SNR level is 15 dB. For the signals at 10 dB and 5 dB, the relative segmentation errors lie between (13-17)% and (17-19)%, respectively. The highest segmentation errors occur at 0 dB SNR with the values between 20% and 25%. Considering the different noisy environments, it is found that white noise has the lowest segmentation errors at all levels of SNRs, while the babble noise has the highest relative error for all instances. On average, the smallest relative error for SNR levels ranging from 20 dB down to 0 dB is observed at white noise followed by train noise, car noise, street noise and babble noise.

Fig. 4.5 shows the variations of the relative error with respect to the number of generations allowed. This convergence curve is generated by testing on the signals of one speaker in white Gaussian noise. From this figure, it can be seen that the relative error of the proposed method decreases rapidly up to 60 generations. When more generations were performed, it was found that no further improvement is obtained after 80 generations.

Figure 4.5: Variation of the relative error with the number of generations.

Hit Rate

Assuming that the difference between the durations of segments detected by the presented method and those of segments manually determined of 25 ms is acceptable, *hit rate* is defined as the fraction of all manually determined segments that are detected by the proposed algorithm within this error range. The hit rate is also examined across varying SNR levels in five types of background noise. In Fig. 4.6, the hit rate obtained as a function of SNRs is presented. It can be seen that hit rates achieved by the proposed algorithm are acceptable. Moreover, in each level of SNRs, almost the same hit rate is obtained for different noise types.

Figure 4.6: Hit rate as a function of SNR in different noise backgrounds.

Precision and Recall

The error of segmentation results can also be categorized into insertion error and deletion error. If a detected segment has the relative error of greater than 25ms, we assumed that segment is missed by the proposed method and a deletion error occurs. If a detected segment boundaries is not found in the manual detection, an insertion error

Table 4.2: Results of recall (RCL), precision (PRC) and F on different types of noise.

	RCL					PRC					F				
Noise	20dB	15dB	10dB	5dB	0dB	20dB	15dB	10dB	5dB	0dB	20dB	15dB	10dB	5dB	0dB
White	0.90	0.85	0.79	0.67	0.56	0.79	0.75	0.69	0.59	0.50	0.83	0.79	0.73	0.62	0.51
Car	0.90	0.86	0.78	0.67	0.54	0.84	0.79	0.72	0.62	0.50	0.86	0.82	0.75	0.64	0.50
Babble	0.87	0.82	0.74	0.64	0.52	0.82	0.78	0.71	0.62	0.50	0.84	0.80	0.73	0.62	0.51
Street	0.89	0.83	0.76	0.64	0.53	0.84	0.79	0.72	0.60	0.53	0.86	0.81	0.73	0.62	0.51
Train	0.88	0.82	0.76	0.64	0.52	0.84	0.79	0.73	0.62	0.53	0.85	0.80	0.74	0.63	0.52

occurs. In the literature of information theory, two measures are often used, namely precision (PRC) and recall (RCL). Corresponding to the deletion and insertion error with precision and recall respectively, they are defined as

$$RCL = \frac{\text{number of hit segments}}{\text{total number of detected segments}} \quad (4.8)$$

$$PRC = \frac{\text{number of hit segments}}{\text{total number of reference segments}} \quad (4.9)$$

Combing the precision and recall, a performance criteria F -measure is often used. It is defined as

$$F = \frac{2 \times PRC \times RCL}{PRC + RCL} \quad (4.10)$$

The value of F lies between 0 and 1, where larger values are desirable. The results of the proposed method conducted at various environments are presented in the Table 4.2. It is observed that the proposed algorithm can maintain high recall rates as well as high precision rates at different SNR levels.

Speech and Non-speech Frames Classification

The performance of proposed method is also evaluated in terms of speech and non-speech detection accuracy. For this evaluation, we have tested the speech signals in white Gaussian noise of the above database. Table 4.3 shows the result for the detection

Table 4.3: Classification accuracy of speech and non-speech frames.

SNR	$E_r(\%)$	$D_n(\%)$	$D_s(\%)$
20 dB	6.13	96.10	91.55
15 dB	7.47	95.53	89.66
10 dB	9.29	95.00	86.72
5 dB	11.49	93.60	83.67
0 dB	13.55	92.21	81.18

Table 4.4: Processing time.

Mean	20.34 s
Median	18.98 s
Minimum	14.11 s
Maximum	30.93 s
25th Percentile	15.96 s
75th Percentile	23.68 s

accuracy of proposed method. D_s and D_n are the ratios of correct speech decisions to the total number of manually marked speech frames and correctly detected non-speech decisions to manually determined non-speech frames, respectively. E_r is the ratio of total false decisions to total number of frames. In this comparison, we have used 10 ms frame size. Under different levels of SNR, our method achieves high performance in both speech and non-speech detections with small error rates.

Computational Complexity

The proposed GA based speech signal segmentation is implemented using Matlab programming language. The computational cost of our algorithm is measured in terms processing time. The processing time needed to complete the whole process running on Pentium IV 3 GHz PC is summarized in Table 4.4.

4.5 Conclusion

A scheme for the segmentation of speech signals in different noisy backgrounds is presented in this chapter. The problem of segment boundary detection is formulated as an optimization problem. Based on the genetic algorithm (GA), the start and end points of the segments are determined. Using sample entropy (i.e., a regularity measure of time series) and heterogeneity measure, an evaluation function of GA is designed. The optimal locations of the boundaries are achieved through the evolution of genetic algorithm. The efficiency and simplicity of Walsh basis functions also make the proposed scheme to be effective to determine the segments present in the segmenting signals. The performance of the method is evaluated using speech signals in different noise conditions. The results show that the proposed method can precisely detect the speech segments as well as noise only segments.

Chapter 5

Segmentation of Narrowly Spaced Noisy Audio Signals

In this chapter, the segmentation of narrowly spaced noisy speech signals is addressed. The purpose of this segmentation is to estimate the locations and durations of noise only narrow gaps which are located among the signal components in order to segment the whole noisy input signal. To deal with this problem, two segmentation techniques have been proposed.

5.1 Introduction

Nowadays, web-audio technology is developing with the increasing demand of multimedia applications, such as audio, video conferencing, voice over IP and distance learning over the Internet. However, quality of service (QoS) for real time multimedia applications has not been completely provided yet. Yet there are challenges for real-time transmission due to the constraints such as bandwidth, packet loss and delay jitters. For example, signal drop outs and/or extraneous noise can occur as a result of

transmission error through the digital communication channels. Drop outs can result in small intervals where the audio signal from the transmitter becomes noisy or lost completely. Signal loss can also occur due to spurious noises such as clicks, pops, and crackles. In general, these types of noise have shorter durations and are often present in data restoration of old disk records. In the presence of signal loss, the quality of perceived audio usually degrades significantly. In many application areas of speech and audio signal processing, detection and reconstruction algorithms for missing audio data is always necessary. Much work on detection and reconstruction of missing audio therefore have been reported in the literature [102, 103, 104]. An overview of packet loss recovery techniques for streaming audio can be found in [105].

In the previously reported segmentation algorithms, the algorithms are based on the acoustic parameters such as energy, entropy, zero-crossing rate, LPC coefficients, mel cepstrum coefficients (MFCC) [106, 107]. Extraction time of these parameters to describe the process is quite long in all these methods. Often complex strategies are needed to achieve the desired level of accuracy.

In contrast to those previously described methods, the first segmentation method proposed in this chapter is mainly based on a time scaling approach (Section 5.2). Utilizing FFT and binary Walsh transform, the signal to be segmented is time scaled first. Employing this time scaled modified signal, the locations and durations of noisy segments as well as noise only narrow segments are determined. Although the computational complexity of this algorithm is relatively low, the preliminary segmentation results are accurate without *a priori* information about signal or noise.

The second segmentation algorithm has been developed exploiting the convexity measure and the residual signal (Section 5.3). Depending on the potential local minima of the residual signal, narrowly spaced audio segments are detected. Using an analysis by synthesis scheme and binary Walsh transform, the original input signal

is reconstructed as the residual signal. Boundaries of noisy and noise only segments are initially determined from the residual signal. The results for the location of segment boundaries are then improved by adopting the *convexity* measurement. As in the time scaling based segmentation algorithm, the determination of narrow gaps is accomplished while detecting the noisy segments. To evaluate the efficiency of this approach, experiments have been carried out on noisy speech signals which are created using five types of noise from the NOISEX-92 database and three types of noise from the database in [108].

5.2 Time Scaling Based Segmentation

Most of the segmentation techniques that have been described in the literature are metric based approaches. Usually, they are implemented using the combination of two or more acoustic features extracted from the audio signal to be segmented. In this section, a new method to segment a noisy speech signal based on time scaling technique is proposed. It is assumed that noise only gaps buried between signal components are relatively smaller than signal components. Using the time stretched or time scaled modified version of the original noisy signal, the presence of narrow gaps as well as noisy speech components are determined. The use of binary Walsh transform for the segmentation of noisy audio signals is also investigated.

Time scaling techniques are usually employed when we need to modify the time scale of a speech or audio waveform without losing the perceptual quality. They usually shorten or stretch out the length of a sound file with a desired scaling factor. Several methods have been addressed for the issue of time scaling of acoustic signals [109, 110] as they are applicable in a number of occasions.

5.2.1 Vocoder Based Time Scaling

The basic idea of this method is to enhance the input signal with time scale expansion from which locations and periods of the gaps and segments are to be obtained. By stretching the audio signal, the reconstructed signal becomes more meaningful than the original signal since the algorithm can more easily differentiate the changes between speech segments and noise gaps. In this system, a phase vocoder based time scaling approach is employed with a sequence of analysis, modification, and resynthesis. The standard phase vocoder technique for time scaling is described in [76]. The main algorithm of time scaling is to define an analysis hop factor which is different from the synthesis hop factor. Here, we denote the analysis hop factor as R_a and synthesis hop factor as R_s . The scaling factor α is then defined as

$$\alpha = \frac{R_a}{R_s} \quad (5.1)$$

Traditionally, phase vocoder based time stretching algorithms keep the magnitude unchanged and modify the phase using phase unwrapping by preserving the instantaneous frequencies. Finally, the signal can be reconstructed as the weighted sum of cosines using the filter bank approach or IFFT approach.

First, we have modified the signal as weighted sum of cosine as described in [77, 76]. This time scaling algorithm first extracts a series of sequential blocks by setting analysis time instants along the input signal for successive integer values. For each block, FFT is computed over a Hann windowed portion of the original signal to obtain a magnitude and phase representation for every R_a samples as follows.

$$X(t_a^u, \Omega_k) = \sum_{n=-\infty}^{\infty} h(n)x(t_a^u + n)e^{-j\Omega_k n} \quad (5.2)$$

Here, STFT representation of the input signal x , $X(t_a^u, \Omega_k)$, is obtained for the analysis time instants t_a^u for set of successive integer values u starting at 0. $h(n)$ is the

analysis Hann window, $\Omega_k = \frac{2\pi k}{N}$ is the centre frequency of the k th vocoder channel, N is the size of Fourier transform and $t_a^u = uR_a$.

At each synthesis time instants t_s^u , a short-time signal $y_u(n)$ is obtained for $t_s^u = R_s u$ by taking the inverse Fourier transform of the synthesis STFT as $Y(t_s^u, \Omega_k)$.

$$y_u(n) = \frac{1}{N} \sum_{k=0}^{N-1} |Y(t_s^u, \Omega_k)| e^{j\Omega_k n} \quad (5.3)$$

where interpolated magnitude values are the same as the analysis magnitude values.

$$|Y(t_s^u, \Omega_k)| = |X(t_a^u, \Omega_k)| \quad (5.4)$$

When all the short-time signals are concatenated together, the output signal $y(n)$ is obtained as:

$$y(n) = \sum_{n=-\infty}^{\infty} y_u(n - t_s^u) \quad (5.5)$$

In this context, to calculate the phase of $Y(t_s^u, \Omega_k)$, the phase increment per sample is also necessary to be computed as the hop size for the synthesis is different from the analysis.

$$\Delta\Phi_k^u = \angle X(t_a^u, \Omega_k) - \angle X(t_a^{u-1}, \Omega_k) - R_a \Omega_k \quad (5.6)$$

The instantaneous frequency $\hat{w}_k(t_a^u)$ is then derived after taking the principal value of the phase increment, $\Delta_p\Phi_k^u$.

$$\hat{w}_k(t_a^u) = \Omega_k + \frac{1}{R_a} \Delta_p\Phi_k^u \quad (5.7)$$

After the instantaneous frequency at time t_a^u is estimated, the phase of the time-scaled STFT at time t_s^u is set using the following formula.

$$\angle Y(t_s^u, \Omega_k) = \angle Y(t_s^{u-1}, \Omega_k) - R_s \hat{w}_k(t_a^u) \quad (5.8)$$

An example of this scheme is given in Fig. 5.1. A clean speech signal is depicted in Fig. 5.1(a), where the narrow gaps are located between samples 2300 and 4450. When

this signal is corrupted by noise, it is difficult to detect the location of the gaps as shown in Fig. 5.1(b). The time scaled version of this noisy speech signal using phase vocoder approach is displayed in Fig. 5.1 (c). Here, the signal is expanded by a scaling factor $\alpha=2$. Obviously, the time stretched signal as sum of sinusoids cannot give any indicative representations for the locations of speech segments, or noise only narrow gaps.

As we are less concerned about perfect reconstruction, the binary Walsh transform is replaced at the resynthesis phase to reconstruct the signal as a weighted sum of Walsh coefficients as follows.

$$y_u(n) = \sum_{k=0}^{N-1} |X(t_a^u, \Omega_k)| \cdot \phi(k) \quad (5.9)$$

where ϕ represents a binary Walsh basis function of Walsh ordering. The resultant modified sequence is shown in Fig. 5.1(d).

It is observed that each time a gap or boundary of a segment appears in the processing signal, sharp and sudden changes always occur in the reconstructed or time scaled signal. In other words, the magnitude of the modified signal may reach its minimum in accordance with the scaling factor. On the other hand, the changes in the remaining parts of the expanded signal are not noticeable. Thus the significant information for the detection of narrow gaps in noisy speech waveforms can be achieved by simply applying this Walsh transform based synthesized signal.

Furthermore, computational efficiency can be achieved by using binary Walsh transform. For example, Walsh-Hadamard matrices of order N can be generated using the recursive relationship $H_N = H_{N/2} \otimes H_2$ where N is the power of 2. In that case, ' $\otimes$ ' denotes the Kronecker product and the lowest order H_2 is of the following form:

Figure 5.1: (a) Noise-free signal; (b) noisy signal; (c) time-scaled version of noisy signal using sinusoidal analysis-synthesis; (d) time-scaled signal using Walsh transform.

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad (5.10)$$

Since waveform characteristics of the proposed time scaled modified signal is highly informative, we only set a threshold to determine the location of gap boundaries. We have investigated many time scaled speech waveforms of speech signals in different noise levels using the proposed time scale modification technique. An example of a clean signal including narrow gaps is given in Fig. 5.2 (a). The waveform of this signal at 20 dB SNR is shown in Fig. 5.2 (b). Time-scale modification of this signal is shown in Fig. 5.2 (c), where the abrupt changes occur according to the locations of gaps. In Fig. 5.2 (e), the time-stretched waveform of the signal at 0 dB (Fig. 5.2 (d)) is also shown. Here, magnitude of the time stretched modified signal decreases significantly at the locations of narrow gaps as in Fig. 5.2 (c) despite high noise level.

Figure 5.2: (a) Noise-free signal; (b) signal at 20 dB SNR; (c) modified time-stretched signal (d) signal at 0 dB SNR (e) modified time-stretched signal.

In general, after time scaling of the input signal, a gap occurs with decrease variation in the stretched waveform. Thus, we set a global threshold from the magnitude of the reconstructed signal. Here, the median of the magnitude of the reconstructed signal is set as threshold. First, the consecutive locations of the reconstructed signal whose magnitude is less than, or equal to this threshold is assumed as noise only gaps. Start and end locations of gaps are then determined from the interval of the candidate gaps. Here, the beginning and local minima of this interval are taken as start and end locations of the gaps. If an estimated interval of a gap is less than 4 ms (for sampling frequency of 8 kHz), we declare that it is a segmentation boundary.

Figure 5.3: (a) Noise-free signal; (b) noisy signal; (c) time-scaled signal (d) segmentation results.

5.2.2 Experimental Study

In order to evaluate the described time scaling based segmentation method, simulations were performed using music and noisy speech signals. Each of speech signals consist of speech segments as well as gaps of various sizes and they are assumed to be buried under white Gaussian noise. Fig. 5.3 illustrates the result of segmentation of a noisy speech utterance. Narrow gaps are manually allocated between the signal components as in Fig. 5.3(a). However, the presence of these gaps are diminished when it is corrupted by the noise as it can be seen in Fig. 5.3(b). This noisy speech waveform is scaled and reconstructed by the proposed Walsh transform based time-scaling method. We have found that if the scaling factor applied is grater than 2, the phase shifting is occurred at the scaled signal. Therefore, we set the scaling factor as $1 \leq \alpha \leq 2$ through out the simulation study. The time scaled signal with the scaling factor of 2 is shown in Fig. 5.3(c). The segmentation results are displayed by the vertical solid lines in Fig. 5.3(d).

According to these lines, we can detect the segments of speech. Every two solid lines between two arrows estimates the location a gap and a single solid line may be start or end point of a segment. From there we can determine in which part of the signal there exist noise only components and where the components of combining speech and noise are.

Figure 5.4: Modification of a music excerpt. (a) original signal; (b) modification without stretching; (c) modification with stretching factor $\alpha = 2$; (d) signal with gaps; (e) modification without stretching; (f) modification with stretching factor $\alpha = 2$.

As another possible application, segmentation of music signals for onset detection is considered in this section. A music excerpt of 3.25s is extracted from “Around the World” by ATC. The waveform of the music signal is shown in Fig. 5.4(a) where

music events are closely located in time. The signal is first converted into modified waveform with a scaling factor $\alpha = 1$ (i.e. no stretching). The resulting waveform revealing sudden changes for music events is shown in Fig. 5.4 (b). In other words, the onset of music segments can be detected by tracking the abrupt variations in the modified waveform. In Fig. 5.4 (c), the signal is modified with scaling factor $\alpha = 2$ (i.e. time stretching). Comparing the modified waveforms of Fig. 5.4 (b) and Fig. 5.4 (c), the variations in this stretched modified waveform is more visible with sharper local minima. In general, a segmentation point can be identified whenever a sharp decrement of the modified waveform occurs.

The advantage of modification with time stretching is also shown when the signal contains narrow gaps as in Fig. 5.4(d). Modification of this waveform with stretching factors 1 and 2 are shown in Figs. 5.4 (d) and 5.4 (e), respectively. The resulting modified waveforms provide local minima at the onsets of music segments. However, due to more sharp representation of the stretched waveform, the narrow gaps and music segment boundaries can be detected more easily (Fig. 5.4 (e)).

5.2.3 Discussion

Segmentation of closely spaced noisy speech signal is described in this section. Based on a new time-scaling technique applying linear binary Walsh transform, a segmentation algorithm for closely spaced noisy audio signal is presented. Applying the reconstructed or time stretched signal, segmentation of speech components and detection of gaps are performed. Moreover, estimated durations of the narrow gaps can also be determined. The result of the experiments are accurate in the absence of *a priori* knowledge. According to the experiments conducted on several speech signals, the scaling factor, α should lie within 1 and 2 to capture the dynamics of the detecting signal. However,

modification of the algorithm is required to improve the efficiency of this method. By using the variable scaling factors at the transitions of signal components is still under study.

5.3 Segmentation Based on Residual Signal and Convexity Measure

In this section, we propose a fast and efficient method to detect the audio segments in noisy situation where noise level is high and noise type is unknown. Instead of extracting the features from the input signal, the noisy audio segments are identified from a residual signal, which is derived from the reconstructed versions of the segmenting signal. As in the previous section, we assumed that adjacent signal components are closely located and the time interval between them are unknown. While determining the presence of the signal components having signal plus noise information, the durations of those noise only narrow gaps are also estimated.

The process of the proposed method can be divided into three major stages. The input signal is firstly reconstructed as a residual signal using a modified analysis/synthesis scheme and binary Walsh transform. A threshold scheme is then applied to the resulting reconstructed signal to obtain the baseline segments. Finally, the locations of baseline segments are validated with a measurement called *convexity*. As stated in the previous chapters, the choice of Walsh transform rather than other transform is its simplicity in computation, involving only addition and subtraction operations. We observed that higher order Walsh functions yield smaller amplitudes and they cannot provide the discriminative features for the locations of segments and gaps. Therefore, we have used only the lower order of Walsh functions having localization property. The

performance of the algorithm is also enhanced due to its low computation complexity. The speech signals with signal-to-noise ratio as low as 0 dB in five different noise conditions have been tested for performance evaluation.

5.3.1 Analysis-Synthesis Process

The main idea behind the proposed method is to reconstruct the noisy input signal as a residual signal, from which the desired segmentation function can be derived. This is implemented by using an analysis-synthesis scheme of [77] and Walsh functions of lower order sequences. The purpose of this analysis-synthesis scheme is to modify the original signal before resynthesis.

In the analysis phase, the input signal $x(n)$ is multiplied by a sliding window of length N , which provides successive windowed signal segments. These windowed signal segments are then transformed to the frequency domain by using FFTs. In this way, a time-varying spectrum $X(n, k) = |X(n, k)|e^{j\varphi(n, k)}$ with $k = 0, 1, \dots, N - 1$ is calculated for each windowed segment.

In the synthesis phase, only the magnitude portion of this short-time spectrum is taken out and reconstructed into two synthesized sequences using two different Walsh functions. Output segments are then overlapped and added accordingly to yield the two output signals. In this method, only the Walsh functions of sequence order 0 and 2 are used as they can provide more detailed local variations of the input signal. Walsh function ϕ_2 is shifted prior to the application to move the zero frequency components to the center of the spectrum. Walsh function ϕ_2 consists of 1's for $n = 0, \dots, N/4$, followed by -1's from $N/4 + 1, \dots, 3N/4$ and 1's for $n = 3N/4, \dots, N$. After shifting, Walsh function ϕ_2 will contain -1's for $n = 0, \dots, N/4$, followed by 1's from $N/4 + 1, \dots, 3N/4$ and -1's for $n = 3N/4, \dots, N$. Since most of the speech energy is present in the lower

frequencies from 0 up to 2 kHz for the sampling frequency of 8 kHz, we can correspond N to maximum frequency of 4 kHz, $N/2$ as 2 kHz and so on. Thus, shifted ϕ_2 contains more significant energy comparing to ϕ_2 . As a result of this shifting, the residual signal becomes more informative and more indicative having sharper waveform with higher magnitudes.

Since the first few functions of the Walsh matrix(i.e. ϕ_0, ϕ_2) are necessary in this implementation, the other functions can be discarded using a proper algorithm. This also reduces the computational time drastically. Finally, the output residual signal is reconstructed by taking the difference between these two synthesized signals. Fig. 5.5 illustrates the block diagram of the proposed modified analysis-synthesis scheme.

5.3.2 Detection Scheme

While the signal components have both signal and noise information, gaps contain noise only. Thus, the amplitude of the residual signal drops significantly at the locations of gaps. Observing that, a sequence of local minima, $\{\gamma_a\}_{a=1}^L$, are detected from the residual signal after being smoothed by a moving average filter. A frame is considered as a noise only frame if its local minimum is below a preset amplitude threshold, th . This threshold is set as $th = \mu_\gamma - 0.5\sigma_\gamma$, where μ_γ and σ_γ are mean and standard deviation of γ_a .

After applying the threshold, th , positions of narrow gaps and baseline segments can be preliminarily determined. At this stage, fine-tuning is required to reduce the over and under segmentation rate. For this purpose, we empirically set the thresholds η_1 and η_2 on temporal distance. The segments whose temporal distance is smaller than η_1 are to be merged with a neighboring segment. And if the temporal distance of a baseline segment exceeds η_2 , it needs to be split up. In this case, a local minimum is

Figure 5.5: The analysis and synthesis process.

searched between two consecutive local maxima within this segment and set this local minimum as a segmentation point.

5.3.3 Convexity

We have mentioned that the amplitude of the residual signal falls significantly at the intervals of narrow gaps and it increases thereafter. When the signal is at high SNR, this property is more prominent and we can easily determine the presence of the boundaries by using the predefined thresholds th , η_1 and η_2 . However when the signal of interest

is at low SNR, there are more potential boundaries than real segment boundaries. In this case, justifying the positions of the boundaries with a bound, *convexity*, may be appropriate. The main reason of why we are interested in convexity is that *over the interval of convexity there is only one minimum*. Here, we use the definition given in [111].

The set A in $\mathbb{R}^n$ is said to be convex if whenever it contains two points, it also contains the line segment joining them. Algebraically, A is convex if $\lambda x + \mu y \in A$ whenever $x, y \in A$ and $\lambda, \mu \geq 0$ with $\lambda + \mu = 1$.

In this context, we search a set of points which are convex with a minimum within every two consecutive baseline segment boundaries obtained after preliminary detection. Then we define the points of this convex set as the interval of narrow gaps. Fig. 5.6 illustrates how the convexity measurement can improve the base line segmentation results. Fig. 5.6(a) and (b) are the clean signal and noisy signal, respectively. The reference segments and gaps acquired before the convexity measurement are shown in Fig. 5.6(c). It can be seen that the duration of the detected second gap becomes wider than the actual gap size. However, after the convexity measurement, this widening effect is reduced to an acceptable level as illustrated in Fig. 5.6(d).

5.3.4 Experiments and Evaluations

Test Data Set I

The described segmentation method has been evaluated with several simulations tested on the speech signals corrupted by high levels of noise. This test database contains 172 speech utterances of 10 speakers (5 males and 5 females) from TIDIGITS. In order to obtain the closely spaced signals, small gaps of varying durations were randomly inserted among the signal components before the addition of noise. Mean duration of

the gaps is 56.98 ms and the standard deviation of the duration is 39.06 ms. Five types of typical noise from NOISEX-92 were considered: white, volvo, babble, F-16 cockpit and factory floor. To generate the contaminated signals, the noise signals were added to the clean speech signals with different SNRs including 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB. Thus, a total of 4300 utterances were used in this experiment.

Figure 5.6: Segmentation result of a noisy signal (a) noise-free signal; (b) noisy signal; (c) segmentation result before convexity measurement (d) segmentation result after convexity measurement.

For this test data set I, the performance evaluation is described in terms of *recall* and *precision*. Assuming that manually determined segment boundaries are the best segmentation points, these two performance indices are calculated with respect to the total number of speech segments which are determined manually as follows.

$$Precision = \frac{\text{number of correctly detected segments}}{\text{total number of detected segments}} \quad (5.11)$$

$$Recall = \frac{\text{number of correctly detected segments}}{\text{total number of segments}} \quad (5.12)$$

Figs. 5.7 and 5.8 show the results as a function of SNRs in different noise types. It is observed that the proposed method can detect the speech segments with high accuracy in adverse environments. When SNR is between 5 dB and 20 dB, 90% of precision and

Figure 5.7: The recall rate for test data set I across five SNRs and five noise conditions.

recall rates can be achieved. The average lowest rates of precision and recall are found in babble noise followed by F-16, factory floor, volvo and white Gaussian noises.

Test Data Set II

This data set also consists of speech signals from TIDIGITS database which are different from Test Data Set I. In this case, 150 speech signals spoken by 10 speakers (5 males and 5 females) were extracted. Before adding the noise, narrow gaps of different sizes were randomly inserted among the audio components. Then noisy signals of SNRs 0 dB, 5 dB, 10 dB, 15 dB and 20 dB were generated by adding noise signals to each clean signal. Three types of noise signals from [108] were considered: white Gaussian, car and street. Therefore a total of 2250 noisy audio signals were conducted in this simulation. Table 5.1 and Table 5.2 list the results using the evaluation metric: precision and recall. As high precision and high recall rates are achieved at different noisy situations, the method is robust against noise. In this experiment, an evaluation index which is presented in [112] is also considered. We have measured the difference between the real and estimated onset on the same set of noisy audio signals. Different

Figure 5.8: The precision rate for test data set I across five SNRs and five noise conditions.

Table 5.1: Precision rate with different SNRs.

SNR	Noise Types		
	White	Car	Street
0 dB	92.69	88.50	90.98
5 dB	95.79	94.69	93.96
10 dB	96.62	95.31	96.62
15 dB	98.49	97.19	97.86
20 dB	98.85	98.50	98.89

distributions of audio segments in each simulated signal, different levels of SNRs and different types of noise lead to different errors in detection as shown in Fig. 5.9. A positive error corresponds to an anticipated detection and a negative error corresponds to a delayed estimate. The bias of the onset estimate is lower than 10 ms and the standard deviation of the onset estimate also reduces when the range of SNR varies.

Table 5.2: Recall rate with different SNRs.

SNR	Noise Types		
	White	Car	Street
0 dB	94.15	91.53	94.67
5 dB	96.17	94.92	95.06
10 dB	96.26	95.74	96.64
15 dB	97.84	97.06	97.59
20 dB	98.48	97.99	98.52

Figure 5.9: The error (difference between the real and estimated onset) of segments for 450 signals at SNR = 20, 15, 10, 5, 0 dB by investigating on three types of noise.

5.4 Conclusion

Segmentation is a prerequisite for a number of application in speech and audio processing. Performance of segmentation methods usually degrades when the signals are contaminated by noise. Therefore, the aim of this chapter is to address the issue of noisy audio segmentation. In particular, segmentation of audio signal with narrow interval gaps is considered. This type of algorithms are suitable for the detection and reconstruction of missing audio data for the streaming data.

In this chapter, we have intended to propose the segmentation methods which are simple and efficient for different noisy conditions. For this purpose, first we have investigated the potential of using Walsh transform in a time scaling approach. It is shown that Walsh transform based time stretched signal is highly informative to be applied in detecting the segments of noisy audio signal.

Based on an analysis-by-synthesis scheme and Walsh basis functions, a residual signal has been developed in the second method. The algorithm has been tested with noisy speech signals consisting of various size of narrow gaps and segments. The experimental results indicate that the proposed methods are robust and can be applicable in different environmental conditions. Moreover, high precision and high recall rates are achieved across the varying SNRs while the overall onset error was relatively low. From these results, it can be concluded that Walsh transform based time scaled signal and residual signal can be utilized as a useful tool in audio segmentation.

Since the proposed algorithms are simple and computationally inexpensive, it can be applied in many areas of audio processing including online systems. In this study, we have tested the efficiency of proposed methods only on the speech signals. Extension of our proposed segmentation algorithm can include the detection of signals such as electromyograph (EMG) signals and phonocardiogram (PCG) signals.

Chapter 6

Segmentation of Noisy Multicomponent Audio Signals in Time-Scale Domain

This chapter proposes two efficient segmentation methods of multi-component noisy signals using Morlet wavelet transform (Section 6.4) and a linear binary time-scale transform (Section 6.5), when the signal components are closely spaced and the time interval between adjacent signal components are unknown. The segmentation problem is formulated using the paradigm of estimating the locations and durations of noisy narrow gaps of the input noisy signals.

6.1 Introduction

Segmentation of audio components separated by narrow gaps are considered in this chapter. The segmentation problem is formulated using the paradigm of estimating the locations and durations of noisy narrow gaps of the input noisy signals. All the

signal components are clubbed together to form one set containing signal plus noise, while the noise components residing in the inactive time period between the signal components form another set containing only the noise.

Segmentation of signal components involves exploiting the time domain, spectral and spatial parameters. The key time, scale/frequency and spatial domain parameters are the time location and duration, center frequency and bandwidth, and spatial location of the waveform. In this chapter, we restrict to our study time and scale parameters. Adjacent signal components are distinguished from another signal component in terms of their time and scale domain parameters. To exploit the time characteristic, high frequency (low-scale) components are employed to separate signal component from noise while to separate adjacent signal components low frequency (high-scale) are employed.

6.2 Signal Model

A multi-component signal is formed of L signal components, which may overlap both in time and in frequency. We may express the signal, $s_o(k)$, as follows:

$$s_o(k) = \sum_{i=1}^L \alpha_i s_{oi}(k - k_i) \quad (6.1)$$

where k is the time index, $s_{oi}(k)$ is the i -th component signal, α_i is its weight and k_i is its time delay. The measured signal, $y_o(k)$, is corrupted by noise, $v_o(k)$.

$$y_o(k) = s_o(k) + v_o(k) \quad (6.2)$$

It is assumed that the duration of the signal segment is much larger than that of the noisy gap. In another way, the activation intervals in the signal are much longer than the noisy silence/quiet intervals.

6.3 Assumptions

- y_o is a slowly time-varying signal so that it may be assumed to be block-wise stationary.
- The noise, v_o , is assumed to be a zero-mean white Gaussian noise process and is uncorrelated with the signal components.
- The bandwidth of the overall signal, s_o , (and not the individual components, $\{s_i, i = 1, \dots, L\}$) is either known *a priori*, or may be determined from the Fourier transform of the measurement, y_o . Thus we need to estimate the bandwidth of s_o and not the bandwidth of the individual components, $\{s_i, i = 1, \dots, L\}$.

6.4 Segmentation Using Wavelet Transform

A new segmentation method of multi-component noisy signals using wavelet transform is proposed in this section. It is shown that the Morlet wavelet transform is useful for segmenting a noisy signal, when the signal components are closely spaced. A wavelet scale sequence comprising of the highest absolute scales for each time instant is employed as test statistics for segmentation. A number of selected local maxima obtained from the wavelet scale sequence correspond to the position of the noisy gaps. Finally, windowed approximate entropy is calculated for the masked noisy signal to estimate the locations and durations of the narrow noisy gaps as well as the noisy segments.

6.4.1 Morlet Wavelet Transform

The objective is to use the features (e.g. the scale) of the wavelet transform, $\{\phi(k, m)\}$, so that the given noisy signal can be segmented efficiently. The choice of the wavelet transform is problem dependent, that is, it depends upon the requirement of the time-frequency/scale resolutions. If the signal components are closely spaced in time but overlap in frequency, then time resolution is important. If the signal components are closely spaced in frequency but overlap in time, then frequency resolution is important. The wavelet transform can handle a wide class of problems including those requiring high time resolution, and those requiring high frequency/scale resolution by choosing appropriate wavelet parameters.

There are many choices within the wavelet family. In our case, the Morlet wavelet transform, [113, 114], is employed because of its ability to provide different window lengths for signals composed of different frequencies/scales, and the time-bandwidth product is minimal providing thereby a good time and frequency resolution. The relative bandwidth, which is the ratio of the bandwidth, Δf , to the center-frequency of the input signal, given by, $\Delta f/f_o$, is constant for all frequency. The time resolution becomes very good at high frequencies, while the frequency resolution becomes very good at low frequencies. Thus the Morlet wavelet is ideal for commonly occurring signals, which contain components of low frequencies with longer duration, and components of high frequencies components with short duration. The Morlet wavelets form a non-orthogonal basis function with considerable spectral overlap among the basis functions. The Morlet wavelet transform [113, 114] is given by

$$\phi(k, m) = \pi^{-1/4} e^{j2\pi f_o k/m} e^{-k^2/2m^2} \quad (6.3)$$

The Fourier transform of the Morlet wavelet is given by

$$\Phi(\omega, m) = \pi^{-1/4} e^{-(\omega - \omega_o/m)^2/2}, \quad \omega \geq 0 \quad (6.4)$$

6.4.2 Windowed Approximate Entropy

A statistic called Approximate Entropy ($ApEn$) was introduced by Pincus [63] to measure the regularity or predictability of time series. A time series containing many repetitive patterns has a relatively small $ApEn$; a less predictable process has a higher $ApEn$. The original formulation of the $ApEn$ by Pincus [63, 115] which has been described in Section 2.2.4 is a computationally expensive measurement which characterizes the irregularity of a finite sequence from start to end of the sequence. Therefore, a windowed version of the $ApEn$ is considered which processes only M signal values at a time for a window of length M .

Consider an input signal $y(n)$ of length N such that $1 \leq n \leq N$,

$$\mathbf{y} = [y(1), y(2), \dots, y(N)]^T \quad (6.5)$$

where T denotes transpose. First, $(N - M + 1)$ number of windowed sequences of length M are constructed from the input signal. Here, a windowed sequence w_l of length M is of the following form

$$w_l = [y(l), y(l + 1), \dots, y(l + M - 1)] \quad (6.6)$$

where $l = 1, 2, \dots, N - M + 1$.

To compute the $ApEn$ of each windowed sequence, $(M - m + 1)$ blocks of length m are then constructed. A block x of length m is

$$x_{l,i} = [w_l(i), w_l(i + 1), \dots, w_l(i + m - 1)] \quad (6.7)$$

where $i = 1, 2, \dots, M - m + 1$.

To measure the similarity between the two blocks, the difference between these two blocks is calculated. The distance measure between two blocks, $x_{l,i}$ and $x_{l,j}$, of length

m is defined as [115]

$$d(x_{l,i}, x_{l,j}) = \max_{k=1, 2, \dots, m} |x_{l,i}(k) - x_{l,j}(k)| \quad (6.8)$$

Two blocks, $x_{l,i}$ and $x_{l,j}$, are similar when the distance between these two blocks is less than a threshold T , which is defined as

$$T(l, r) = r\sigma \quad (6.9)$$

where $r < 1$ and $\sigma = \text{std}[w_l]$. ‘std’ refers to standard deviation.

Considering all the blocks of length m within a windowed sequence, we find the number of blocks in that sequence that resemble to a block $x_{l,i}$ using the threshold $T(l, r)$.

$$C_{l,i}(m, r) = \frac{(\text{number of } j \exists d(x_{l,i}, x_{l,j}) \leq T(l, r))}{M - m + 1} \quad (6.10)$$

where $j \leq M - m + 1$ and $C_{l,i}(m, r)$ is the probability that any block of $x_{l,j}$ is within the threshold $T(l, r)$ of $x_{l,i}$. Then the average of the natural logarithms of the functions $C_{l,i}(m, r)$ is calculated.

$$\Theta_l(m, r, M) = \frac{1}{M - m + 1} \sum_{i=1}^{M-m+1} \ln C_{l,i}(m, r) \quad (6.11)$$

The windowed ApEn is finally defined as

$$\text{ApEn}(m, r, l, M) = \Theta_l(m, r, M) - \Theta_l(m + 1, r, M) \quad (6.12)$$

where $m \geq 1$.

It is noted that during the software implementation of the windowed *ApEn*, the evaluated time instant is placed in the center of the odd length window (or one of the center for the case of an even length window) instead of at the trailing edge of the window as shown in (6.6).

6.4.3 Algorithm

The proposed algorithm is based on two stages. The first stage is the time-scale transformation while the second stage is the implication of a decision strategy for segmentation based on scale information along time. This segmentation algorithm, thereby, consists of the following steps (step 1 corresponds to stage 1, whereas steps 2 to 6 corresponds to stage 2):

1. The input noisy signal is transformed into time-scale domain from the time-domain using Morlet wavelet. Here, the Morlet's wavelet filter bank divides input signal spectrum into J number of octave scales (e.g, $J = 64$ is used in our examples).
2. Then a wavelet scale sequence containing the highest absolute scale (i.e. the scale with the highest absolute value) at each time instant is obtained from the time-scale representation of the input signal. This scale sequence will be then used as a test statistics for segmentation using the following steps.
3. A number of local maxima/minima for the wavelet scale sequence are detected due to the sharp transitions in the scale sequence. It is found that the presence of the noise gaps can be realized from the sharp variations in scale values with respect to time instants.
4. In order to reduce the false detection of the noisy narrow gaps, the local maxima/minima are selected according to the following rules:
 - Select only the local maxima and discard the local minima since it is observed that the local maxima represent mostly the locations of the noise gaps.
 - Remove the local maxima which are close to the start and end boundaries of the signal. It is found that those maxima are appeared due to the truncation effect during the processing.

- Also discard the maxima which are close together and consider the maxima which are apart from each other. Since we assume that the noise gaps are quite narrow compared to the length of each signal segment.
5. Then the input noisy signal is masked by considering only the signal values around the selected maxima by multiplying with a binary function. The binary function is set ones for the neighbouring samples of the input noisy signal around each of the selected maxima and zeros elsewhere.
6. Finally, the windowed approximate entropy is measured for the masked noisy signal to estimate the locations and durations of the narrow noisy gaps as well as the noisy segments. In this context, a binary segmentation function is obtained from the $ApEn$ measurements. We estimate the gap durations as half the total number of $ApEn$ values higher than the mean $ApEn$ value. The binary segmentation function thus shows the estimated locations and durations of the noise gaps as well as the durations of the signal segments.

6.4.4 Evaluation

The proposed scheme is evaluated on a number of simulated examples for the synthetic signals buried in noise. In the following, we present results for a synthetic signal consisting of closely spaced four different sinusoids buried in zero-mean white noise. The period of the wavelet must be chosen to be small (i.e. large scale/high frequency) compared to the narrow time-gap between the components which is of 4 samples (See Fig. 6.1(a)). In fact, the gap sizes, which are chosen to be equal in this example, can be considered to be varied. Using a larger wavelet period (i.e. small scale/low frequency) may miss the narrow gaps. We have chosen the total number of octave scales $J = 64$, where the scale, j , varies from 1 to J . For the windowed $ApEn$ calculation, we have

used $r = 0.5$, $M = 6$ and $m = 1$ as considered in other methods.(e.g. in [116]) and the window is chosen as rectangular window. Fig. 6.1 illustrates the results of segmentation for the synthetic signal. Fig. 6.1 also demonstrates how effectively segmentation is performed by exploiting both the time and scale domain properties using Morlet wavelet. The results are shown for a low SNR (e.g. 4 dB).

Figure 6.1: The results of our scale domain segmentation when the noise standard deviation is 1. (a) noise-free signal; (b) noisy signal; (c) highest absolute scales (for total number scales, $J=64$) vs. time ; (d) local maxima/minima; (e) selected local maxima; (f) masked/extracted noisy input signal; (g) approximate entropy; (h) binary segmentation function/estimated segment intervals and noisy gaps.

Fig. 6.1(a) and (b) show the noise-free and noisy input signals. Fig. 6.1(c) shows the variations of the highest absolute scale values with respect to time. The local maxima/minima obtained from Fig. 6.1(c) are shown in Fig. 6.1(d). The selected local maxima selected according to step 4 in Section 5 are plotted in Fig. 6.1(e). Fig. 6.1(f) shows the extracted noisy sequence using a narrow rectangular window at the positions of the selected maxima. The positions of the selected maxima estimate the locations of the noisy gaps. The windowed $ApEn$, which are then measured for the extracted noisy sequence in Fig. 6.1(f), is plotted in Fig. 6.1(g). The duration of the noisy narrow gaps are estimated from the $ApEn$ as shown in Fig. 6.1(h). In Fig. 6.1(h), the binary segmentation function (solid line) is showing the estimated noisy gaps as well as the signal segments. Comparing Fig. 6.1(h) with 6.1(a), it is found that the estimated location of the first noisy gap is correct while the estimated duration is 6 samples which is larger than the actual duration of 4 samples and the estimated locations for the second and third gaps are slightly shifted although their gap intervals are estimated correctly, i.e. 4 samples.

Figs. 6.2 and 6.3 demonstrate the effectiveness of the time-domain and our scale-domain segmentation method in terms of noise level or signal-to-noise ratio (SNR). According to Fig. 6.2(b) the narrow gaps at low SNR (with $\sigma = 1$) can be detected using scales, i.e. by exploiting scale domain features. It is not possible to do segmentation in the time-domain properties due to the presence of spurious noise peaks in the time domain plot (see Figs. 6.2(c)–(e)). For the signal with higher SNR (with $\sigma = 0.2$) (see Fig. 6.3(c)) the time domain characteristics are able to separate as shown by Figs. 6.3(d)–(e), while the scale domain characteristics, i.e. scales are constant for all components and unable to separate the signal (Fig. 6.3(b)). It is noted that the time-domain method is based on the energy of the wavelet transform outputs, while the test statistics and the threshold are estimated by the following principle component

analysis (PCA) method.

Figure 6.2: The results of segmentation are illustrated for low SNR (noise standard deviation=1.0) using the noisy signal in Fig. 6.2(c). Note that for segmentation in low SNR only the scale domain plot (see Fig. 6.2(b)) can be exploited, while time domain plots (see Figs. 6.2(d)-(e)) are difficult to exploit. (a) noise-free signal; (b) highest absolute scales (for total number scales, $J=64$) versus time; (c) noisy signal; (d) normalized average instantaneous energy; (e) estimated segments and noisy gaps.

Some statistical evaluations are shown for the uniform noisy gaps of 4 samples in Fig. 6.4. For each SNR value, the results of 750 realizations are shown using generated synthetic signals corrupted by additive white Gaussian noise. Fig 6.4 shows the mean errors of the segmentation results taken over the three uniform noisy gaps for the 750 simulated signals at SNR of 14 dB, 5 dB and 0 dB. The errors for both the gap location and gap duration are calculated as the differences between the real and estimated gap position as well as the real and estimated gap duration.

Figure 6.3: The results of segmentation are illustrated for high SNR (noise standard deviation = 0.2) using the noisy signal in Fig. 6.3(c). Note that the scale domain characteristics are the same for all components as we have used high frequency/scale wavelets. The high wavelet scales are able to separate in time domain. (a) noise-free signal; (b) highest absolute scales (for total number scales, $J=64$) versus time; (c) noisy signal; (d) normalized average instantaneous energy; (e) estimated segments and noisy gaps.

Figs. 6.4(a)-(c) show the errors for the gap locations while Figs. 6.4(d)-(f) show the errors for the gap durations. The bias of the mean error for Figs. 6.4(a)-(c) are found to be 1.02, 0.96, 0.92, while the standard deviation (std) of the mean error are obtained as 0.027, 0.34, 0.35, respectively. Similarly, the bias for Figs. 6.4(d)-(f) are found to be 2.16, 2.14, 0.37, while the standard deviation (std) of the mean error are obtained as 2.63, 4.02, 5.99, respectively. According to the above statistics, it is remarkable that bias of the error is lowered as the SNR is decreased although it is found the probabilities of false detection and false rejection are increasing with the decrease of SNR. One can

then expect to achieve better performance using the scale-domain method instead of time-domain method in the high noise condition.

Figure 6.4: Error (difference between the real and estimated gap positions (Figs.6.4(a)-(c)) as well as the real and estimated gap durations (Figs.6.4(d)-(f)) of the segmentation results for the 750 simulated noisy signals at 14 dB (noise std=0.5), 5 dB (noise std=.75), and 0 dB (noise std=1) SNR (Figs.6.4(a) and (d):14 dB SNR; Figs.6.4(b) and (e): 5 dB SNR; Figs.6.4(c) and (f): 0 dB SNR). (X-axis=Trail, Y-axis=Mean error(in sample))

6.4.5 Application with Real Speech Signals

In this section, we have shown the segmentation results for the recorded speech signals when the sampling rate is 8,000 Hz. Figs. 6.5(a) and 6.5(b) illustrate the segmentation

results for the multicomponent speech signals. The results are promising since the signals are segmented into speech segments and non-speech segments correctly.

Figure 6.5: Segmentation results for real speech signals (The segmentation functions (scaled by 5) is plotted together with the input speech signal).

6.5 Segmentation Using Linear Binary Time-Scale Transform

This section proposes an efficient segmentation method of narrowly spaced noisy audio signals using a linear binary Walsh transform. It is shown that Walsh transform is appropriate for segmenting a noisy waveform. A subset of Walsh functions are chosen to cover principally the noise subspace such that the resulting linear combination of the selected basis functions captures the features that can discriminate between signal and noise. In the absence of *a priori* information about the signal and noise statistics, the proposed scheme is based on the linear combination of those basis functions which must be able to identify the adjacent components. It is not necessary that the basis functions reconstruct the noise-free versions of the signal components. The only restrictions is

that the segment length should be some integer power of 2 for the most accurate segmentation.

6.5.1 Choice of the Transform

The objective is to choose the basis functions, $\{\phi(k, m)\}$, or equivalently the transform matrix, $\mathbf{W}_k$, so that the given waveform can be segmented efficiently. The choice of the basis functions is problem dependent, that is, it depends upon the requirement of the time, scale and time-scale resolutions. The basis functions are to be selected to give a compact and efficient representation—fewer number of terms to capture the distinguishing features of the signal and not the noise components.

The binary transform can handle the segmentation efficiently when the received signal is highly noisy and narrowly separated by choosing the appropriate binary functions. There are many choices within the binary transform family. In this sequel, Walsh transform given in [117] is employed because of its ability to detect the sharp changes from signal plus noise case to noise only case.

The shape of the binary transform, e.g. Walsh transform can match the narrow gaps more appropriately compared to the non-binary transforms. In order to do that large binary functions having good frequency resolutions can be used to separate signal components from noise.

Moreover, any scale transform attenuates the noise components. A signal component, which is transient, requires a fewer number of large binary functions, while the noise requires a very large number of tiny scale components to represent effectively. Hence by neglecting the coefficients related to tiny scale basis functions, the noise can be filtered out.

6.5.2 Basic Concept

This segmentation approach is based on extracting merely the features, which are relevant for waveform segmentation. The time instants when feature change are estimated. The Walsh transform is employed to extract features relevant for segmentation. An estimate of the noise-free waveform is expressed as a linear combination of Walsh basis functions. The weights of the basis functions are obtained by minimizing the least-squares estimation error.

The Walsh basis function is binary assuming values of -1 and 1 over sub-intervals. The number of sub-intervals increases with the order of the basis function; the first-order basis function has only one sub-interval, which is the entire data-record duration, the second-order basis function has two equal sub-intervals and i th order has 2^i sub-intervals. For segmentation the period of the highest order Walsh function should be such that duration of each of the signal components should be integer multiple of the period of the highest order Walsh functions.

The least-squares estimate of the weight takes the form of a sum of additive and subtractive combinations of averages of the noisy waveform over sub-intervals. The averaging operation attenuates zero-mean white noise corrupting the waveform. The estimate of the waveform as a weighted linear combination show a jump at the transition region between adjacent signal components. The waveform is segmented from the locations of the jump. For accurate segmentation, the duration of each of the signal components must be equal to some integer power of 2 so as to match the transition intervals of the basis functions.

6.5.3 Method

The proposed method is based on representing a binary function (staircase type) for segmentation of noisy signals when the segment intervals are determined by the constant intervals and the narrow gaps are located by jumps.

Let $y(k)$ be the estimate of the waveform $y_0(k)$ given as a linear combination of Walsh functions $\{\phi_i(k), i = 0, 1, \dots, M-1\}$

$$y(k) = \sum_{i=0}^{M-1} a_i \phi_i(k) \quad (6.13)$$

Expressing in vector-matrix form, we get

$$\mathbf{y} = \mathbf{W}^T \mathbf{a} \quad (6.14)$$

where superscript T indicates transpose, $\mathbf{a}$ is an $(M \times 1)$ vector, $\mathbf{y}$ is $(N \times 1)$ and $\mathbf{W}$ is $(N \times M)$ matrices given by

$$\begin{aligned} \mathbf{a} &= [a_0 \ a_1 \ \cdots \ a_{M-1}]^T \\ \mathbf{y} &= [y_0 \ y_1 \ \cdots \ y_{N-1}]^T \end{aligned} \quad (6.15)$$

$$W = [\ \phi_0 \ \phi_1 \ \cdots \ \phi_{M-1}] \quad (6.16)$$

where ϕ_i is the i th basis function. The coefficient vector $\mathbf{a}$ is obtained from the least squares minimization

$$\min_{\mathbf{a}} \|\mathbf{y}_0 - \mathbf{y}\|^2 \quad (6.17)$$

Substituting for $\mathbf{y}$, we get

$$\min_{\mathbf{a}} \|\mathbf{y}_0 - \mathbf{W}^T \mathbf{a}\|^2 \quad (6.18)$$

The optimal least squares estimate [118] is given by

$$\mathbf{a} = (\mathbf{W}\mathbf{W}^T)^{-1} \mathbf{W}\mathbf{y}_0 \quad (6.19)$$

Exploiting the binary level property of Walsh basis functions, the optimal coefficients $\{a_i, i = 0, 1, \dots, (M - 1)\}$ takes the following form:

$$a_0 = \frac{1}{N} \sum_{k=0}^{N-1} y_0(k) \quad (6.20)$$

$$a_1 = \frac{1}{N} \left\{ \sum_{k=0}^{N/2-1} y_0(k) - \sum_{k=N/2}^{N-1} y_0(k) \right\} \quad (6.21)$$

$$a_2 = \frac{1}{N} \left\{ \sum_{k=0}^{N/4-1} y_0(k) - \sum_{k=N/4}^{N/2-1} y_0(k) + \sum_{k=N/2}^{3N/4-1} y_0(k) - \sum_{k=3N/4}^{N-1} y_0(k) \right\} \quad (6.22)$$

$$a_j = \sum_i \sum_{k \in I_{ij}} b_{ij} y_0(k) \quad (6.23)$$

where b_{ij} are the coefficients of Walsh transform with $i = 0, 1, \dots, M - 1$, $j = 0, 1, \dots, N - 1$ and I_{ij} represents the time interval. The coefficients $\{a_i, i = 0, 1, \dots, (M - 1)\}$ capture the features of the waveform in a hierarchical form. The coefficient a_0 captures merely the average value of the waveform, a_1 captures the difference between the averages of the waveform over two equal time intervals, a_2 captures the features obtained from averages over four equal intervals, a_3 captures the features obtained from averages over eight equal intervals, a_j captures the features obtained from averages over 2^j equal intervals, and finally a_{M-1} captures the features obtained from averages over 2^{M-1} equal intervals. Thus the information about the waveform is captured progressively from a macroscopic detail to microscopic detail: a_0 gives the macroscopic detail while a_{M-1} gives the complete microscopic detail.

The averaging operation involved in the coefficient calculation attenuates the zero-mean uncorrelated error, v_0 corrupting the measured waveform y_0 . Therefore, the choice of the number of Walsh basis functions employed in estimating the given waveforms must be large enough to capture the features of the waveform, but small enough to attenuate the noise. a_0 gives the least information, while a_{M-1} will contain detailed

information about the waveform but it is affected by the noise as the averaging interval gets progressively reduced from the basis function ϕ_0 to ϕ_{M-1} .

6.5.4 Evaluation

The proposed method is evaluated on a simulated example of a the synthetic signals buried in noise. Here we have illustrated results for a synthetic waveform consisting of four different closely spaced sinusoids (i.e, $L=4$) buried in zero-mean white noise. The signal is assumed to be unknown except that the spectral bandwidth of the entire signal, s_o , is assumed to dominate the noise over its dominant frequencies.

Figs. 6.6 and 6.7 demonstrate how effectively segmentation is performed by exploiting both the time and scale properties. According to Figs. 6.6(c) and 6.7(c) the narrow-gaps at low SNR can be detected using scales, i.e. by exploiting scale domain features. For signal $\mathbf{y}_0$ of length 256 in Figs. 6.6(b) and 6.7(b), we have used the Walsh transform matrix of order 8, providing $2^8 = 256$ binary functions of length 256 (2^8). The first four of the lowest frequency (highest scale) Walsh functions, $\phi(k, m), k = 0, \dots, 255; m = 0, \dots, 3$ and the corresponding coefficients a_0, a_1, a_2 and a_3 are linearly combined and used as the detection index for segmentation defined by *segmentation index*, as shown in Figs. 6.6(c) and Figs. 6.7(c). Note that the segmentation examples are illustrated in Figs. 6.6 and 6.7, when the noise standard deviations are 0.5 and 1.0, respectively. From Figs. 6.6(c) and 6.7(c), it seems that the segmentation index functions are seemed robust against noise level and therefore, is able to separate both the noisy waveforms.

Figs. 6.8 and 6.9 show the results of segmentation with respect to the choice of total number of coefficients ($a_0, a_1, \dots, a_{M-1}$) and their corresponding basis functions ($\phi_0, \phi_1, \dots, \phi_{M-1}$). In Figs. 6.8 and 6.9, the results are shown for the signal in 6.6(a)

Figure 6.6: The results of segmentation are shown at low SNR when noise standard deviation is 0.5. (a) noise-free signal; (b) noisy signal; (c) segmentation using time-scale domain.

with $\sigma_v=0.5$ and 1.0, respectively. According to Figs. 6.8 and 6.9, the period of the highest order basis function should be less than each of the segment intervals. Besides as we can see the interclass variation (i.e. the changes due to narrow gaps between two adjacent segments) is larger than the intra-class variation (i.e. the changes in the same segment without any narrow gaps). These might provide a method to select the total number of coefficients as well as Walsh functions to be used for segmentation providing attenuation of the noise.

An application for segmenting the voiced speech is shown as it is needed for speaker identification/verification. Fig. 6.10(a) shows a voiced signal (phoneme signal), which is taken from the TIMIT data base. The segmentation result is shown in Fig. 6.10(b), where the signal is segmented into four small segments. In Fig. 6.11(a), the noisy voiced speech is shown which is found by adding Gaussian noise to the clean signal shown in Fig. 6.10(a). The segmentation is successful in the high noisy case to be seen by Fig.

Figure 6.7: The results of segmentation are shown at low SNR when noise standard deviation is 1. (a) noise-free signal; (b) noisy signal; (c) segmentation using time-scale domain.

6.11(b), which indicates that this method would be quite useful for segmentation and pitch analysis of the voice signals in noisy speaker identification/verification scheme [119]. It is noted that the same parameters are used as before.

6.6 Discussion

Packet loss is an undesirable and unavoidable problem in streaming audio. Consecutive packet loss always renders a number of gaps in the streaming data. Thus, the intelligibility of the received audio is impaired with increasing loss rates. To repair these damaged audio stream, lost detection and loss-recovery methods are developed at the either side of transmission or receiving. For the streaming audio over the Internet, gap detection is the one of the most common loss-detection techniques. Therefore, gap detection for missing streaming audio data would be the one of the good applications

Figure 6.8: The results on segmentation are shown for the choice of number of coefficients $(a_o, a_1, \dots, a_{M-1})$ and the corresponding basis functions $(\phi_o, \phi_1, \dots, \phi_{M-1})$ to be used for segmentation; (a) using only a_o , (b) using a_o and a_1 , (c) using a_o , a_1 and a_2 , (d) using a_o , a_1 , a_2 and a_3 , (e) a_o , a_1 , a_2 , a_3 and a_4 (The noise standard deviation is 0.5 and the segments are powers of 2).

for the methods presented in this chapter.

6.7 Conclusion

Two segmentation methods for narrowly spaced noisy multicomponent signals are presented. Our methods are not based on strong *a priori* assumptions which make our techniques more generally applicable with respect to other methods. In our segmentation schemes both the signal components (i.e. activity intervals) and gaps (i.e. quiet intervals) are segmented from the high noisy background. The proposed schemes are evaluated on a number of synthetic signals and real signals. The preliminary results of segmentation are quite promising. Moreover, a comparison has been shown between our wavelet transform based segmentation method with the PCA based time-domain

Figure 6.9: The results of segmentation are shown for the choice of number of coefficients $(a_o, a_1, \dots, a_{M-1})$ and the corresponding basis functions $(\phi_o, \phi_1, \dots, \phi_{M-1})$ to be used for segmentation; (a) using only a_o , (b) using a_o and a_1 , (c) using a_o , a_1 and a_2 , (d) using a_o , a_1 , a_2 and a_3 , (e) a_o , a_1 , a_2 , a_3 and a_4 (The noise standard deviation is 1 and the segments are powers of 2).

segmentation method with respect to noise level or SNR. However, further improvement of the algorithm is needed to use in some real cases where the noise could be correlated with the signal components. In addition to gap detection for lost detection of streaming audio, the proposed methods, for example, can be applied to speech signals in voice activity detection and in the detection and classification of the events of biomedical signals in clinical diagnosis.

Figure 6.10: Segmentation for a clean voiced speech; (a) clean voiced speech, (b) segmentation function describing the segments.

Figure 6.11: Segmentation for a noisy voiced speech; (a) noisy voiced speech, (b) segmentation function describing the segments.

Chapter 7

Conclusions And Recommendation

In this thesis, new methods for audio segmentation in the presence of noise were investigated. We have focused on the use of transforms to achieve robust, faster and accurate segmentation results. The major contributions of this thesis are summarized in this chapter.

7.1 Speech/Non-Speech Detection

To determine the different components of speech, namely active components and inactive components, a speech/non-speech detection technique is presented in Chapter 3. In this case, the active corresponds to speech or noisy speech segments and silence or background noise corresponds to the inactive components. The proposed method is mainly based on the reconstructed version of the original noisy signal using the binary Walsh basis functions and an analysis by synthesis scheme. In this algorithm, a method in [82] has been incorporated to select the minimum number of Walsh basis functions.

Since only a few Walsh basis functions are necessary, the computational complexity

of this method is relatively low. Furthermore, the use of Walsh transform makes the implementation of the algorithm to be simple. Unlike the previously reported VADs which have slower response time due to the processes like feature extraction, modelling of speech and non-speech and hangover scheme, the proposed method is faster in implementation with fewer parameters to be adjusted.

This method has been evaluated on speech signals in five types of noisy environments with a range of SNR varying from 0 dB to 20 dB. The results show that the proposed method is able to detect speech as well as non-speech frames with lower error rates than the standardized VADs of G.729, ETSI AMR1 and AMR2. Discriminability between speech and noise using ROC analysis has also shown its use in real time problems. Therefore, applications can be found in real time noise cancellation systems, noise reduction for speech enhancement, discontinuous transmission based on speech/non-speech detection, data compression and silence compression allowing the speech channel to be shared with other information in multimedia communication.

7.2 Segmentation Based on Genetic Algorithm

A segmentation scheme for speech signals in different noisy backgrounds is presented in Chapter 4. The main contribution of this thesis is to formulate the problem of segment boundary detection as an optimization problem. For that purpose, start and end points of speech segments are determined based on a genetic algorithm (GA)- a stochastic optimization algorithm. The possibility of being stuck in local optima is reduced by exploiting the ability of GA, which can utilize parallel exploration of the search space.

A new and effective evaluation function is designed using sample entropy, a regularity measure of time series, and heterogeneity measure. The optimal locations of

segment boundaries are determined through the evolution of genetic algorithm. The efficiency and simplicity of Walsh basis functions also makes the proposed algorithm to be effective in determination of segments present in the segmenting signals. The feasibility of proposed GA based segmentation technique for noisy speech segmentation has been shown by the performance measure using the speech signals in different noisy environments. It has been found that average deviation error in segment boundaries is less than 20% of the actual duration of the segment. For 96% of the segments determined by the presented method, the deviation of detected boundaries from manually determined ones is less than 30 ms.

7.3 Segmentation of Narrowly Spaced Noisy Audio Signals

The aim of this part to address the issue for the segmentation of audio signals where some signal components may have a relatively short duration while some components may have longer duration. Speech and biomedical signals can be such type of signals. The potential of time scaling approach based on Walsh transform was investigated for the detection of these narrowly spaced noisy audio segments. In this chapter, the problem is also investigated through a technique based on residual signal and convexity measure.

Different from the segmentation methods in the literature, the proposed methods do not need feature extraction. The experimental results as shown for noisy speech signals indicate that the ability of the proposed methods in various situations although the performance decrease slightly with the decrement of SNRs. Moreover, high precision rate and recall rate can be achieved while the overall onset error is found relatively

low. Since the algorithms are computationally inexpensive, it can be applied in many areas of real audio processing including online systems.

7.4 Segmentation of Noisy Multicomponent Audio Signals in Time-Scale Domain

The problem to segment of noisy audio signals based on minimal *a priori* information was considered. Details are describe in Chapter 6. Two efficient algorithms for segmenting audio signal in the presence of noise are presented. It is also assumed that signal components are closely spaced and time intervals between adjacent signal components are unknown.

A new segmentation method of multicomponent noisy signals using wavelet transform is proposed. Locations and durations of signal components (i.e. active intervals) and noisy narrow gaps (i.e. inactive intervals) of the input noisy signals are estimated. It has been shown that the Morlet wavelet transform is useful for segmenting a noisy signal, when the signal components are closely spaced. To estimate the locations and durations of the narrow noisy gaps as well as noisy segments, the windowed approximate entropy is calculated for the masked noisy signal. Since this method is not based on strong *a priori* assumptions, it makes this technique more generic. The preliminary results as obtained using a number of synthetic signals are found promising.

In the second segmentation method, the desired components are segmented from high noisy background using a linear binary Walsh transform. A subset of Walsh functions are chosen to cover principally the noise subspace so that the linear combination of the selected basis functions captures the discriminating features between signal and noise. The linear combination of the selected basis functions is able to identify the

adjacent signal components. The limitation of this method is that the segment length should be some integer power of 2 for the most accurate segmentation. The results, as shown for the synthetic signals are found quite promising. This proposed scheme can be applied on a number of real signals such as speech, music and PCG signals.

7.5 Recommendation

New segmentation techniques for audio signals in noisy conditions are presented in this thesis. Specifically, four aspects of audio segmentation are explored using the Walsh and wavelet transforms. Some of the possible extensions for the proposed works in this thesis are as follows.

One of the possible extensions for the works presented in this thesis is speech/music discrimination. In most of the speech/music discrimination techniques, a set of features are extracted to train the classifiers such as HMM or GMM from which the discriminated result to be determined. Usually, the extraction of features to characterize the signals is computationally expensive. Instead of this feature extraction, the reconstructed versions of input signal, which have been analyzed in this thesis can be investigated to analyze the dynamics of the signal. The discrimination between speech and music can then be performed by either using a HMM classifier or Independent Component Analysis (ICA) as in [120].

It would be possible to extend speech/non-speech detection algorithm presented in this thesis in video segmentation. Detection of speech and non-speech is the initial stage of video segmentation with the help of audio analysis. Further classification of non-speech frames into music, environmental sound and silence segments can also be accomplished by introducing the entropy metric (approximate entropy, windowed approximate entropy and sample entropy) which have been efficiently utilized in this

thesis.

Several proposals have addressed the problems of detection, estimation and classification of biological signals such as PCG signals, surface EMG signals and uterine EMG signals and acceleration signals for walking type analysis exploiting the wavelet transforms. Biological signals consist of short-lived, high-frequency components closely located in time as well as long duration components closely spaced in frequency [121]. Traditionally, electrocardiogram (ECG) signals consist of four main events of cardiac cycles: P wave, QRS complex, ST segment and T wave. The automatic detection of these characteristic components in ECG cardiogram is an important task in monitoring the heart abnormalities. The study of wavelet transform and genetic algorithm to segment ECG signals would be another future direction. In this case, the fitness function can be designed using entropy metric suite as the method proposed in Chapter 4.

Furthermore, segmentation and identification of first and second sounds of heart sound signals (PCG signals) is crucial to characterize the murmurs present in the cardiac cycles. It would be a good investigation to study the effects of different Walsh basis functions and wavelets to segment PCG signals. Detected segments can then be classified into different type of heart sounds using a classifier, such as neural network classifier.

Author's Publications

1. F. Sattar, M. Pwint and R. Doraiswami. On the Segmentation of Narrowly-Spaced Noisy Audio Signals. In *Proceedings of the 3rd IEEE International Conference on Multimedia and Expo, ICME'02*, August, 2002, Laussane, Switzerland.
2. M. Pwint and F. Sattar. A Segmentation Method to Closely-Spaced Noisy Audio Signals. In *Proceedings of the 7th IEEE International Symposium on Consumer Electronics, ISCE'03*, December, 2003, Sydney, Australia.
3. M. Pwint and F. Sattar. A Time-Scaling Approach to Detect the Closely-Spaced Segments of Noisy Speech Signals. In *Proceedings of the 10th IEEE International Conference on Electronics, Circuits and Systems, ICECS'03*, December, 2003, Sharjah, United Arab Emirates.
4. M. Pwint and F. Sattar. On the Detection of Closely-Spaced Noisy Audio Segments. In *Proceedings of the 4th International Conference on Information, Communications and Signal Processing, ICICS-PCM'03*, December, 2003, Singapore.
5. F. Sattar, R. Doraiswami and M. Pwint. A New Segmentation Technique for Noisy Multi-Component Signals Using Wavelet Transform. In *Proceedings of International Symposium on Circuits and Systems 2004, ISCAS'04*, May, 2004, Vancouver, Canada.
6. M. Pwint and F. Sattar. A Segmentation Method for Noisy Speech Using Genetic Algorithm. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, ICASSP'05*, March, 2005, Philadelphia, USA.
7. M. Pwint and F. Sattar. A New Speech/Non-Speech Classification Using Minimal Walsh Basis Functions. In *Proceedings of International Symposium on Circuits and Systems 2005, ISCAS'05*, May, 2005, Kobe, Japan.
8. M. Pwint and F. Sattar. Speech/Non-speech Detection Using Minimal Walsh Basis Functions. Submitted to *EURASIP Journal on Audio, Speech, and Music Processing*.
9. M. Pwint and F. Sattar. Genetic Algorithm Based Segmentation of Speech Signals in Noisy Environments. Submitted to *IEEE Transaction on Speech and Audio Processing*.
10. M. Pwint and F. Sattar. Segmentation of Narrowly Spaced Audio Signals Based on Time-Scaling. Submitted to *EURASIP Journal on Applied Signal Processing*.

11. F. Sattar, R. Doraiswami and M. Pwint. Segmentation of Noisy Multicomponent Audio Signals in Time-Scale Domain. Submitted to *IEEE Transaction on Multimedia*.

Bibliography

- [1] D. H. Milone and A. J. Rubio. Prosodic and accentual information for automatic speech recognition. *IEEE Trans. Speech and Audio Processing*, 11(4):321–333, 2003.
- [2] J. Ajmera, G. Lathoud, and I. McCowan. Clustering and segmenting speakers and their locations in meetings. *ICASSP'04*, 1:I605–608, 2004.
- [3] S. S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. *DAPRA Broadcast News Transcription and Understanding Workshop*, 1998.
- [4] J. C. Segura C. Benítez J. Ramírez, A. de la Torre, and A. Rubio. A new voice activity detector using subband order-statistics filters for robust speech recognition. *ISCAS'04*, 1:I849–852, 2004.
- [5] M. Jelinek and F. Labonté. Robust signal/noise discrimination for wideband speech and audio coding. *Proc. Speech Coding Workshop*, pages 151–153, 2000.
- [6] U. Mittal and N. Phamdo. Signal/noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Trans. Speech and Audio Processing*, 8(2):159–167, 2000.
- [7] F. Beritelli, S. Casale, and A. Cavallaro. A robust voice activity detector for wireless communications using soft computing. *IEEE J. Select. Areas Comm.*, 16(9):1818–1829, 1994.
- [8] GSM 06.94 ETSI. Digital cellular telecommunications system (phase 2+); voice activity detectors (VAD) for adaptive multi-rate (AMR) speech traffic channels; european telecommunications standards institute. 1999.
- [9] M. Marzinik and B. Kollmeier. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Trans. Speech and Audio Processing*, 10(2):109–118, 2002.

-
- [10] H. Sheikhzadeh, R. L. Brennan, and H. Sameti. Real-time implementation of HMM based MMSE algorithm for speech enhancement in hearing aid applications. *Proc. ICASSP'95*, 1:808–811, 1995.
 - [11] A. Rezayee and S. Gazor. An adaptive KLT approach for speech enhancement. *IEEE Trans. Speech and Audio Processing*, 9(2):87–95, 2001.
 - [12] J. Wei, L. Du, Z. Yan, and H. Zeng. A new algorithm for voice activity detection. *Proc. Int. Symposium. on Circuits and Systems*, 2:588–591, 2003.
 - [13] K. Srinivasan and A. Gersho. Voice activity detection for cellular networks. *Proc. Speech Coding Workshop*, pages 85–86, 1993.
 - [14] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd. The voice activity detector for the PAN-European digital cellular mobile telephone service. *ICASSP'89*, 1:369–372, 1989.
 - [15] S. G. Tanyer and H. Özer. Voice activity detection in nonstationary noise. *IEEE Trans. Speech and Audio Processing*, 8(4):478–482, 2000.
 - [16] Y. Wu and Y. Li. Robust speech/non-speech detection in adverse conditions using the fuzzy polarity correlation method. *Proc. Systems, Man and Cybernetics*, 4:2935–2939, 2000.
 - [17] ITU-T Recommendation G.729 Annex B. A silence compression scheme for G.729 optimized for terminals conforming to recommendation v.70. 1996.
 - [18] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *Signal Processing Letter*, 6(1):1–3, 1999.
 - [19] Y. D. Cho and A. Kondo. Analysis and improvement of a statistical model-based voice activity detector. *Signal Processing Letter*, 8(10):276–278, 2001.
 - [20] S. Gazor and W. Zhang. A soft voice activity detector based on a Laplacian-Gaussian model. *IEEE Trans. Speech and Audio Processing*, 11(5):498–505, 2003.
 - [21] F. Kubala, H. Jin, S. Mastoukas, L. Nguyen, R. Schwartz, and J. Makhoul. The 1996 bbn byblos hub-4 transcription system. *Speech Recognition Workshop*, pages 90–93, 1997.
 - [22] P. C. Woodland, M. J. F. Gales, D. Pye, and S. J. Young. The development of the 1996 HTK broadcast news transcription system. *Speech Recognition Workshop*, pages 73–78, 1997.

-
- [23] B. Ramabhadran, J. Huang, U. Chaudhari, G. Iyengar, and H. J. Nock. Impact of audio segmentation and segment clustering on automated transcription accuracy of large spoken archives. *Eurospeech'03*, pages 2589–2592, 2003.
 - [24] L. Wilcox D. Kimber. Acoustic segmentation for audio browsers. *Interface Conference*, 1996.
 - [25] A. Martin, D. Charlet, and L. Mauuary. Robust speech/non-speech detection using LDA applied to MFCC. *Proc. ICASSP'01*, 1:237–240, 2001.
 - [26] L. S. Huang and C. H. Yang. Novel approach to robust speech endpoint detection in car environments. *ICASSP'00*, 3:1751–1754, 2000.
 - [27] F. Beritelli. A robust endpoint detector on differential parameters and fuzzy pattern recognition. *IEEE ICSP'98*, pages 601–604, 1998.
 - [28] S. B. Samad A. Hussain and L. B. Fah. Endpoint detection of speech signal using neural network. *TENCON'00*, 1:104–108, 2000.
 - [29] J. C. Junqua, B. Mak, and B. Reaves. A robust algorithm for word boundary detection in the presence of noise. *IEEE Trans. Speech and Audio Processing*, 2:406–412, 1994.
 - [30] G. D. Wu and C. T. Lin. Word boundary detection with mel-scale frequency bank in noisy environment. *IEEE Trans. Speech and Audio Processing*, 8:541–554, 2000.
 - [31] M. J. Carey and E. S. Parrisand H. Lloyd-Thomas. A comparison of features for speech, music discrimination. *ICASSP'99*, 1:149–152, 1999.
 - [32] H. Sundaram and S.-F. Chang. Audio scene segmentation using multiple features, models and time scales. *ICASSP'00*, 6:2441–2444, 2000.
 - [33] P. Delacourt and C. Wellekens. Audio indexing: use of second-order statistics for speaker-based segmentation. *Multimedia Computing and Systems'99*, 2:959–963, 1999.
 - [34] S. S. Chen, E. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanvesky, and P. Olsen. Automatic transcription of broadcast news. *Speech Communication*, 37:69–87, 2002.
 - [35] U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll. New approaches to audio-visual segmentation of TV news for automatic topic retrieval. *ICASSP'01*, 3:1397–400, 2001.

-
- [36] Y. Wang, Z. Liu, and J.-C. Huang. Multimedia content analysis using both audio and visual cues. *IEEE Signal Processing Magazine*, 17(6):12–36, 2000.
 - [37] T. Muramoto and M. Sugiyama. Visual and audio segmentation for video streams. *ICME'2000*, 3:1547–1550, 2000.
 - [38] H. Jiang, T. Lin, and H.-J. Zhang. Video segmentation with the assistance of audio content analysis. *ICME'2000*, 3:1507–1510, 2000.
 - [39] <http://mathworld.wolfram.com>.
 - [40] H. Zhang and D. Rutkowski. Orthogonal sequency division modulation - a novel method for future broadband radio services. *IEEE Vehicular Technology*, 2:810–814, 1995.
 - [41] M. Adjouadi and F. Candocia. A stereo matching paradigm based on the Walsh transformation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(12):1212–1218, 1994.
 - [42] C. Wang, W.-J. Zhang, and X.-Z. Fang. Adaptive reduction of blocking artifacts in DCT domain for highly compressed images. *IEEE Trans. Consumer Electronics*, 50(2):674–654, 2004.
 - [43] N. Ahmed, H.-H. Schreiber, and P.-V. Lopresti. On notation and definition of terms related to a class of complete orthogonal functions. *IEEE Trans. Electromagn. Compat.*, 15:75–80, 1973.
 - [44] M. C. Yoon and D. H. Chin. Cutting force monitoring in the endmilling operation for chatter detection. *Engineering Manufacture*, 219(6):455–465, 2005.
 - [45] D. E. Goldberg. *Genetic Algorithm in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.
 - [46] A. J. F. Van Rooji, L. C. Jain, and R. P. Johnson. *Neural Networks Training Using Genetic Algorithms*. New York: World Scientific, 1996.
 - [47] J. R. Koza. *Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems*. Report No. STAN-CS-90-1314, Stanford University, 1990.
 - [48] M. Ozdemir. *Evolutionary computing for feature selection and predictive data mining*. Ph. D Thesis, Rensselaer Polytechnic Institute, Troy, New York, 2002.
 - [49] A. Chipperfield, P. Fleming, H. Pohlheim, and C. Fonseca. Genetic algorithm toolbox. *Department of Automatic Control and Systems Engineering, University of Sheffield*, 1995.

-
- [50] M. S. White and S. J. Flockton. *Adaptive recursive filtering using evolutionary algorithms*. Evolutionary Algorithms in Engineering Applications, Springer Verlag, 1997.
 - [51] F. Pan, R.-C. Han, and D.-M. Feng. An identification method of time-varying delay based on genetic algorithm. *Machine Learning and Cybernetics*, 2:781–783, 2003.
 - [52] K. S. Tang, K. F. Man, S. Kwong, and P. J. Fleming. GA approach to multiple objective optimization for active noise control. *Algorithms and Architectures for Real-Time Control*, pages 13–19, 1995.
 - [53] B. Bhanu, S. Lee, and J. Ming. Adaptive image segmentation using a genetic algorithm. *IEEE Trans. Systems, Man, and Cybernetics*, 25(12):1543–1567, 1995.
 - [54] M. Yu and L. Udpa N. Eua-anant, A. Saudagar. Genetic algorithm approach to image segmentation using morphological operations. *ICIP'98*, 3:775–779, 1998.
 - [55] M. Haseyama, M. Kumagai, and H. Kitajima. A genetic algorithm based image segmentation for image analysis. *ICASSP'99*, 6:3445–3448, 1999.
 - [56] C. Rosenberger and K. Chehdi. Genetic fusion: application to multi-components image segmentation. *ICASSP'00*, 4:2223–2226, 2000.
 - [57] S. Y. Ho and K. Z. Lee. An efficient evolutionary image segmentation algorithm. *Evolutionary Computation'01*, 2:1327–1334, 2001.
 - [58] C. T. Lin, H. W. Nein, and J. Y. Hwu. GA-based noisy speech recognition using two-dimensional cepstrum. *IEEE Trans. Speech and Audio Processing*, 8(6):664–675, 2000.
 - [59] E. A. Durant and G. H. Wakefield. Efficient model fitting using a genetic algorithm: pole-zero approximations of HRTFs. *IEEE Trans. Speech and Audio Processing*, 10(1):18–27, 2002.
 - [60] E. A. Durant, G. H. Wakefield, D. J. Van Tasell, and M. E. Rickert. Efficient perceptual tuning of hearing aids with genetic algorithms. *IEEE Trans. Speech and Audio Processing*, 12(2):144–155, 2004.
 - [61] E. Y. Kim, S. H. Park, K. Jung, and H. J. Kim. Genetic algorithm-based segmentation of video sequences. *Electronics Letters*, 36(11):946–947, 2000.
 - [62] P. Chiu, A. Girgensohn, W. Polak, E. Rieffel, and L. Wilcox. A genetic algorithm for video segmentation and summarization. *ICME'2000*, 3:1329–1332, 2000.

-
- [63] S. M Pincus. Approximate entropy as measure of system complexity. *Proc. Nat. Acad. Sci*, 8:2297–2301, 1991.
 - [64] D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman. Sample entropy analysis of neonatal heart rate variability. *Am J Physiol*, 283(3):R.789–797, 2002.
 - [65] S. M Pincus. Quantification of evolution from order to randomness in practical time series analysis. *Methods Enzymol*, 240:68–89, 1994.
 - [66] K. K. L. Ho, G. B. Moody, C.-K. Peng, J. E. Mietus, M. G. Larson, D. Levy, and A. L. Goldberger. Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics. *Circulation*, 96:842–848, 1997.
 - [67] S. M. Pikkujämsä, T. H. Mäkikallio, K. E. J. Airaksinen, and H. V. Huikuri. Determinants and interindividual variation of R-R interval dynamics in healthy middle-aged subjects. *Am J Physiol Heart Circ Physiol*, 280:H1400–H1406, 2001.
 - [68] T. A. Kuusela, T. T. Jartti, K. U. O. Tahvanainen, and T. J. Kaila. Nonlinear methods for biosignal analysis in assessing terbutaline-induced heart rate and blood pressure changes. *Am J Physiol Heart Circ Physiol*, 282:H773–781, 2002.
 - [69] P. M. Ngan. Motion detection using approximate entropy. *DICTA'97*, pages 379–384, 1997.
 - [70] F. Sattar, R. Doraiswami, and M. Pwint. A new segmentation technique for noisy multi-component signals using wavelet transform. *ISCAS'04*, 3:III229–232, 2004.
 - [71] S. M. Pincus. Approximate entropy (ApEn) as a complexity measure. *Chaos*, 5:110–117, 1995.
 - [72] J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol*, 278(6):H2039–2049, 2000.
 - [73] N. Bailey. Sculptor: A real time phase vocoder. *LINUX Journal*, 1998.
 - [74] A. Imai, N. Seiyama, T. Mishima, T. Takagi, and Eiichi Miyasaka. *Application of Speech Rate Conversion Technology to Video Editing*. NHK Laboratories Note No. 475, 2001.
 - [75] M. R. Portnoff. Implementation of the digital phase vocoder using the fast Fourier transform. *IEEE Trans. Speech and Audio Processing*, 24(3):243–248, 1976.

-
- [76] J. Laroche and M. Dolson. Improve phase vocoder time-scale modification of audio. *IEEE Trans. Speech and Audio Processing*, 7(3):323–332, 1999.
 - [77] F. Keiler D. Arfib and U. Zöler. *DAFX - Digital Audio Effects*. John Wiley Publisher, 2002.
 - [78] E. Favreau. Phase vocoder applications in GRM tools environment. *DAFX'01*, 2001.
 - [79] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi. Applying support vector machines to voice activity detection. *Proc. Int. Conf. on Signal Processing*, 2:1124–1127, 2002.
 - [80] B. L. McKinley and G. H. Whipple. Model based speech pause detection. *ICASSP'97*, 2:1179–1182, 1997.
 - [81] R. R. Martin and M. M. Anguh. Quadtrees, transforms and image coding. *Computer Graphics Forum*, 10:91–96, 1991.
 - [82] A. Quddus and M. Gabbouj. Wavelet-based corner detection technique using optimal scale. *Pattern Recognition Letter*, 23:215–220, 2002.
 - [83] M. Adjouadi, F. Candocia, and J. Riley. Exploiting Walsh-based attributes to stereo vision. *IEEE Trans. Signal Processing*, 44(2):409–420, 1996.
 - [84] Z. Hu, J. Schalkwyk, E. Bernard, and R. Cole. Speech recognition using syllable-like units. *ICSLP*, 2:1117–1120, 1996.
 - [85] G. Ruske and T. Schotola. An approach to speech recognition using syllabic decision units. *ICASSP'78*, 3:722 – 725, 1978.
 - [86] M. Sharma and R. Mammone. Automatic speech segmentation using neural tree networks. *Neural Networks for Signal Processing Workshop*, pages 282–290, 1995.
 - [87] V. K. Prasad, T. Nagarjan, and H. M. Murthy. Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communication*, 42:429–446, 2004.
 - [88] M. Ostendorf and S. Roukos. A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Trans. Speech and Audio Processing*, 37(12):1857–1869, 1989.
 - [89] A. Basu and T. Svendsen. A time-frequency segmental neural network for phoneme recognition. *ICASSP'93*, 1:509–512, 1993.

-
- [90] G. Aversano, A. Esposito, and M. Marinaro. A new text-independent method for phoneme segmentation. *Midwest Symposium on Circuits and Systems (MWS-CAS'01)*, 2:516–519, 2001.
 - [91] B. Raj and R. Singh. Classifier-based non-linear projection for adaptive endpointing of continuous speech. *Computer Speech and Language*, 17(1):5–26, 2003.
 - [92] Y. Zhang, X. Zhu, Y. Hao, and Y. Luo. A robust and fast endpoint detection algorithm for isolated word recognition. *IEEE Intelligent Processing Systems*, pages 1819–1822, 1997.
 - [93] T. Zhang and C. Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans. Speech and Audio Processing*, 9:441–457, 2001.
 - [94] J. Ajmera, I. A. Mccowan, and H. Bourlard. Robust HMM-based speech/music segmentation. *ICASSP'02*, 1:I297–300, 2002.
 - [95] J. Piquier, J. L. Rouas, and R. André-Obrecht. A fusion study in speech/music classification. *ICME'03*, 1:I409–412, 2003.
 - [96] S. MacDougall, A. K. Nandi, and R. Chapman. Multiresolution and hybrid Bayesian algorithms for automatic detection of change points. *IEE Proc. Vision, Image and Signal Processing*, 145(4):280–286, 1998.
 - [97] X. Jin and C. H. Davis. A genetic image segmentation algorithm with a fuzzy-based evaluation function. *IEEE Proc. Fuzzy Systems*, 2:R.938–943, 2003.
 - [98] G. Sparacino, F. Bardi, and C. Cobelli. Approximate entropy studies of hormone pulsatility from plasma concentration time series: influence of the kinetics assessed by simulation. *Ann Biomed Eng*, 28(6):665–676, 2000.
 - [99] J. Bruhn, H. Ropcke, B. Rehberg, T. Bouillon, and A. Hoeft. Electroencephalogram approximate entropy correctly classifies the occurrence of burst suppression pattern as increasing anesthetic drug effect. *Anesthesiology*, 93(4):981–985, 2000.
 - [100] D. A. Coley. *An Introduction to Genetic Algorithms for Scientists and Engineers*. World Scientific, 2001.
 - [101] K. S. Tang, K. F. Man, S. Kwong, and Q. He. Genetic algorithms and their applications. *Signal Processing Magazine, IEEE*, 13(6):22–37, 1996.
 - [102] P. A. A. Esquef, V. Välimäki, K. Roth, and I. Kauppinen. Interpolation of long gaps in audio signals using the warped Burg's method. *Int. Conference on Digital Audio Effects (DAFx-03)*, 2003.

-
- [103] I. Kauppinen and J. Kauppinen. Reconstruction method for missing or damaged long portions in audio signal. *Journal of the Audio Engineering Society*, 50(7):594–602, 2002.
 - [104] L. Lu, L. Wenyin, and H.-J. Zhang. Audio textures: Theory and applications. *IEEE Trans. Speech and Audio Processing*, 12(2):156–167, 2004.
 - [105] C. Perkins, O. Hodson, and V. Hardman. A survey of packet loss recovery techniques for streaming audio. *IEEE Network*, 12(5):40–48, 1998.
 - [106] L. Lu, H.J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Trans. Speech and Audio Processing*, 10(7):504–516, 2002.
 - [107] W. H. Shin, B. S. Lee, Y. K. Lee, and J. S. Lee. Speech/non-speech classification using multiple features for robust endpoint detection. *ICASSP'00*, 3:1399–1402, 2000.
 - [108] <http://www.ee.columbia.edu/~dpwe/sounds/noise>.
 - [109] B. Ninnes and S. Henriksen. Time and frequency scale modification of speech signals. *ICASSP'00*, 3:1295–1298, 2000.
 - [110] T. Quatieri, R. B. Dunn, and T.E. Hanna. A subband approach to time-scale expansion of complex acoustic signals. *IEEE Trans. Speech and Audio Processing*, 3(6):515–519, 1995.
 - [111] R. Webster. *Convexity*. Oxford University Press, 1994.
 - [112] A. Merlo, D. Farina, and R. Merletti. A fast and reliable technique for muscle activity detection from surface EMG signals. *IEEE Trans. Biomedical Engineering*, 50:316–323, 2003.
 - [113] O. Rioul and M. Vetterli. Wavelet and signal processing. *IEEE Signal Processing Magazine*, pages 14–38, 1991.
 - [114] C. Torrence and G. M. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–76, 1998.
 - [115] S. M. Pincus and W. M. Huang. Approximate entropy: statistical properties and applications. *Commun. Stat. Theory Meth.*, 21:3061–3070, 1992.
 - [116] M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini. Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. *IEEE Trans. Biomedical Engineering*, 50(3):365–374, 2003.

- [117] J. Astola S. Aghaian and K. Egiazarian. *Binary Polynomial Transforms and Nonlinear Digital Filters*. Marcel Dekker Inc, 1995.
- [118] S. Haykin. *Adaptive Filter Theory*. Second ed, Prentics-Hall, 1991.
- [119] A. Drygajlo and M. El-Maliki. Speaker verification in noisy environment with combined spectral subtraction and missing feature theory. *ICASSP'98*, 1:121–124, 1998.
- [120] P. Vanroose. Blind source separation of speech and background music for improved speech recognition. *24th Symposium on Information Theory*, pages 103–108, 2003.
- [121] R. Fischer and M. Akay. Fractal analysis of heart rate variability. *Time Frequency and Wavelets in Biomedical Signal Processing*, M. Akay, Ed. Piscataway, NJ: IEEE, pages 719–728, 1998.