CO-CHANNEL SPEECH SEPARATION USING STATE-SPACE RECONSTRUCTION AND SINUSOIDAL MODELLING

By

Yasser Mahgoub, B.Eng., M.Eng.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of

Doctor of Philosophy

Ottawa-Carleton Institute for Electrical and Computer Engineering
Faculty of Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
October 2009

Your file Votre référence
ISBN: 978-0-494-67893-0
Our file Notre référence
ISBN: 978-0-494-67893-0

Please note:

The pages iii and iv have been removed for privacy reasons.

Abstract

This thesis deals with the separation of mixed speech signals from a single acquisition channel; a problem that is commonly referred to as *co-channel speech separation.* The goal of the thesis is to present some contributions towards the design and implementation of a robust and enhanced co-channel speech separation system.

The phenomenon of co-channel speech commonly occurs due to the combination of speech signals from simultaneous and independent sources into one signal at the receiving microphone, or when two speech signals are transmitted simultaneously over a single channel. An efficient co-channel speech separation system is an important front-end component in many applications such as Automatic Speech Recognition (ASR), Speaker Identification (SID), and hearing aids.

The separation process of co-channel speech consists, mainly, of three stages: *Analysis*, *Separation*, and *Reconstruction.* The central separation stage represents the heart of the system in which the target speech is separated from the interfering speech. At the front, since the separation process works on one segment of co-channel speech at a time, a mean must be found in the analysis stage to accurately classify each segment into single or multi-speaker before separation. Precise estimation of each speaker's speech model parameters is another important task in the analysis stage. The speech signal of the desired speaker is finally synthesized from its estimated parameters in the reconstruction stage. In order to have a reliable overall speech separation system, improvements need to be achieved in all three stages.

This thesis introduces a classification algorithm that is capable of determining the voicing-state of co-channel speech. The algorithm uses some features of the

reconstructed state-space of the speech data as a measure to identify the three voicing-states of co-channel speech; Unvoiced/Unvoiced (U/U), Voiced/Unvoiced (V/U), and Voiced/Voiced (V/V). The proposed method requires neither a priori information nor speech training data. Nonetheless, simulation results show enhanced performance in identifying the three voicing-states at different target-to-interference ratio (TIR) values as well as at different levels of background noise compared to other existing techniques.

A time-domain method to precisely estimate the sinusoidal model parameters of co-channel speech is also presented. The method does not require the calculation of the discrete Fourier transform nor the multiplication by a window function which both degrade the estimate of the sinusoidal model parameters. The method incorporates a least-squares estimator and an adaptive technique to model and separate the co-channel speech into its individual speakers. The application of this method on speech data demonstrates the effectiveness of this method in separating co-channel speech signals with different TIRs.

Acknowledgements

All the praises and thanks be to **Allah**, the Lord of mankind.

I would like to express my sincere gratitude to my thesis supervisor, Professor **Richard Dansereau**, for his continuous support, guidance, encouragement and involvement throughout the entire research process. His help is greatly appreciated.

At Carleton University, I am indebted to all my professors at the Faculty of Engineering for providing me with the knowledge to develop my research. In particular, I would like to thank my program advisor, Professor **Rafik Goubran**, for his tremendous support and guidance during the first phase of my studies. Special thanks are also due to the fellow graduate students of the Department of Systems and Computer Engineering, for their participation in interesting and valuable discussions. I would especially like to thank Mohammad Radfar, Ahmed Khalil, Ahmed Rami Abu El-Quran, Ziad El-khatib, Lijing Ding and Zhong Lin.

Financial support provided by Carleton University in the form of a research assistantship and in the form of a graduate scholarship, and by the National Capital Institute of Telecommunications (NCIT) in the form of a research grant is gratefully acknowledged.

I am also very grateful to my mother, my wife, my brother and my children for their continuous prayers and support during the course of this work. They all have stood by me and offered a wonderful helping hand. Indeed, my thanks are extended to all whose visions sustained me throughout the research production, and to those whose willingness to help made it possible for me to complete this research.

Finally, I dedicate this dissertation to the memory of my father.

"And God has brought you forth from your mothers' wombs knowing nothing - but He has endowed you with hearing, and sight, and minds, so that you might have cause to be grateful."

(Quran, 16:78)

"Say: Verily my prayer, my sacrifice, my life and my death are all for Allah, the sustainer of the universe. He has no partner. And of this I have been commanded, and I am the first of the Muslims."

(Quran, 6:162-163)

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACF	Autocorrelation function
AM	Amplitude modulation
AMDF	Average magnitude difference function
AMIF	Average mutual information function
ANN	Artificial neural network
APPC	Adjacent pitch period comparison
ASA	Auditory scene analysis
ASR	Automatic speech recognition
AWGN	Additive white gaussian noise
BSD	Bark spectral distortion
BSS	Blind source separation
CASA	Computational auditory scene analysis
CCSS	Co-channel speech separation
CEP	Cepstrum
CNO	Constrained nonlinear optimization
DC	Direct current
DCT	Discrete cosine transform
DDF	Double difference function
DFT	Discrete Fourier transform
DG	Different gender
DMC	Difference-mean comparison
DTW	Dynamic time warping

EM	Estimate maximization
ESACF	Enhanced summary autocorrelation function
FFT	Fast Fourier transform
FHMM	Factorial hidden Markov model
FNN	False nearest neighbor
FOM	Figure of merit
HES	Harmonic enhancement and suppression
HILO	High-to-Low frequency energy ratio
HMM	Hidden Markov model
HMS	Harmonic magnitude suppression
HNR	Harmonic-to-noise energy ratio
HS	Harmonic selection
ICA	Independent component analysis
IEEE	Institute of electrical and electronics engineers
IFFT	Inverse fast fourier transform
IIR	Infinite impulse response
IP	Internet protocol
IS	Itakura-Saito
ITU	International telecommunication union
KWS	Keyword spotting
LLR	Log-likelihood ratio
LMSE	Least mean square error
LPC	Linear predictive coding
LPCC	Linear prediction cepstral coefficient
LPF	Low pass filter
LS	Least squares
MAP	Maximum a posteriori
MFCC	Mel-frequency cepstral coefficient
ML	Maximum likelihood

MLD	Maximum likelihood deconvolution
MLSA	Maximum likelihood speaker assignment
MMSE	Minimum mean-squared error
MOS	Mean opinion score
MPDA	Multi-pitch determination algorithm
MSE	Mean-squared error
ND	Nodal density
NMF	Nonnegative matrix factorization
NPN	No processing is needed
OLA	Overlap-add
PDA	Pitch determination algorithm
PESQ	Perceptual evaluation of speech quality
PSQM	Perceptual speech quality measure
SACF	Summary autocorrelation function
SAPVR	Spectral autocorrelation peak valley ratio
SDR	Signal-to-distortion ratio
SE	Speech enhancement
SFM	Spectral flatness measure
SG	Same gender
SHS	Subharmonic summation
SID	Speaker identification
SM	Sinusoidal modelling
SNMF	Sparse non-negative matrix factorization
SNR	Signal-to-noise ratio
SPL	Sound pressure level
SRR	Signal-to-residual ratio
SSR	Signal-to-signal ratio
ST	Same talker
STE	Short-time energy

STFT	Short-time fourier transform
SVD	Singular value decomposition
TIR	Target-to-interference ratio
TPM	Trajectory parallel measure
U/U	Unvoiced/unvoiced
U/V	Unvoiced/voiced
V/U	Voiced/unvoiced
V/V	Voiced/voiced
VAD	Voice activity detection
VoIP	Voice over internet protocol
WER	Word error rate
WRA	Word recognition accuracy
WSS	Weighted slope spectral
ZCR	Zero-crossing rate

CHAPTER 1

Introduction

1.1 Research Motivation

In many speech processing applications such as speech recognition, speaker identification, and speech enhancement, the input speech signal is often corrupted by the surrounding acoustic noise. This in turn deteriorates the perceived quality and intelligibility[1] of the speech and consequently degrades the overall performance of the speech processing algorithm. When the acoustic interference consists of competing speech signals from other talkers (known commonly as the cocktail-party effect), then further degradation results due to the similarity in nature between the desired and undesired signals. In such scenarios, therefore, a speech separation algorithm represents an essential front-end component to enhance speech quality and intelligibility for further processing. If we can separate the desired speech signal prior to its processing, this will help in enhancing the overall performance of the speech processing algorithm.

In some practical situations where only a single acquisition channel is available, single channel separation techniques must be used. This may be imposed by the system used (as telephone based applications) or by the availability of the desired

[1]Speech quality is a highly subjective measure which reflects the way the signal is perceived by listeners while speech intelligibility is an objective measure of the amount of information which can be extracted by the listeners from the given signal. See section 7.2 for more details.

signal (as prerecorded applications). They are especially interesting due to the simplicity in microphone installation but the major constraint of single channel methods is that there is no reference signal for the interference available. Therefore the power spectral density of the interfering speech has to be estimated based on the available co-channel speech signal only and this is what makes it a challenging task. This problem is commonly referred to as the co-channel speech separation problem.

The idea of a co-channel speech separation is to automatically process the mixed signal in order to recover each talker's original speech. Minimizing artifacts in the processed speech is a key concern, especially if the final goal is to use the recovered speech in machine-based applications such as automatic speech recognition and speaker identification systems. The goal therefore of co-channel speech separation algorithm is to:

- Improve the perceptual aspects of a degraded speech signal.
- Improve the performance of the final speech processing system.
- Increase the robustness of machine-based speech processing systems.

Although it is clearly evident that the human auditory system is very proficient at focusing on a particular speaker or speakers in a mixture [1], computer algorithms, on the other hand, designed to do the same task have demonstrated only a limited degree of success.

1.2 Goal of this Thesis

The goal of the thesis is to present some contributions towards the design and implementation of a robust and enhanced co-channel speech separation system. Figure 1.1 shows a schematic diagram for the framework of the proposed separation system. Referring to the figure, the sampled co-channel speech signal, $x(n)$, which is the sum of the desired speech, $s(n)$, and the interfering speech, $i(n)$, is first segmented into consecutive frames with certain amount of overlap. Each frame is processed

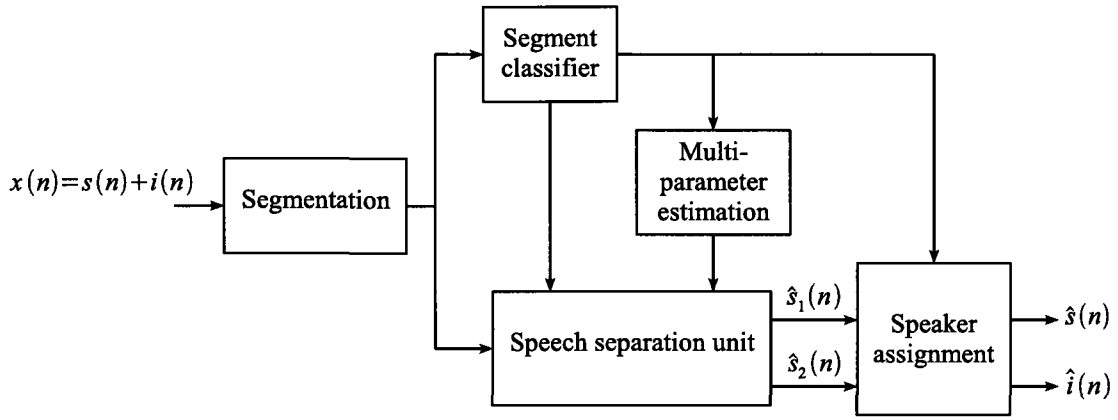


Figure 1.1: A schematic diagram for the framework of the proposed co-channel speech separation system.

separately and classified by the *segment classifier* into one of the possible three voicing states: *unvoiced*, *single-voiced* or *double-voiced*. If the current speech frame is classified as double-voiced, it is passed to the *multi-parameter estimation* unit to estimate sets of parameters that belong to each talker. Examples of these parameters include pitch frequencies and sinusoidal model parameters. These parameters are then fed to the main *speech separation* unit (the heart of the system) to recover the speech signal corresponding to each talker. Finally, a *speaker assignment* technique is used to assign each recovered frame to its proper speaker.

Thesis contributions are aimed to be in the following subsystems:

1. Co-channel speech segment classification into single and multi-talker.
2. Parameters' estimation of the co-channel speech segment such as pitch and model parameters.
3. The main separation algorithm.
4. Overall evaluation technique for the speech separation system.

1.3 Structure of this Thesis

The rest of this thesis is organized as follows:

Chapter 2 introduces the problem of co-channel speech separation and presents a quick review of the various approaches to solve this problem that are commonly reported in the literature.

Chapter 3 provides the necessary background theory and techniques related to the concepts of state-space reconstruction for nonlinear systems and its application to speech signal processing. The chapter also reviews the embedding theorem and methods for determining optimal embedding parameters. In addition, it presents recent work on the application of the state-space reconstruction to various speech-processing systems.

Chapter 4 presents a new classification method for the segment classifier shown in Figure 1.1. The method determines the voicing-state of co-channel speech based on nonlinear state-space reconstruction and utilizing the trajectory parallel measure (TPM). The proposed method requires neither a priori information nor speech training data. Nonetheless, simulation results show enhanced performance in identifying the three voicing-states using the proposed method compared to other existing techniques. Simulation results also show a reliable performance at different target-to-interference ratio (TIR) values as well as at different levels of background noise.

Chapter 5 introduces a time-domain method to precisely estimate the sinusoidal model parameters of co-channel speech as a contribution to the multi-parameter estimation block of Figure 1.1. The proposed method does not require the calculation of the discrete Fourier transform nor the multiplication by a window function. It incorporates a least-squares estimator and an adaptive technique to model and separate the co-channel speech into its individual speakers. The

application of this method on speech data demonstrates the effectiveness of this method in separating co-channel speech signals with different TIRs. Simulation results show the capability of this method in producing accurate and robust parameter estimation in low signal-to-noise (SNR) situations compared to other existing algorithms.

Chapter 6 presents a modified version of the TPM algorithm based on sample-by-sample manner as well as a simplified method for estimating the sinusoidal model parameters. These modifications are made to reduce the implementation cost and to improve the overall performance. The modified algorithms are then implemented and integrated together for the overall separation system to be tested.

Chapter 7 presents testing results of the overall separation system using databases of real speech. Both subjective and objective measures are used to evaluate the performance of the speech separation algorithm based on intelligibility and quality of the processed speech. The performance of the separation system is also evaluated under the conditions of a speech separation challenge in order to be compared with the performance of other systems.

Chapter 8 summarizes the results and draws conclusions arising from the research work. Significant contributions are highlighted and finally, future research directions are suggested.

1.4 Main Contributions

As mentioned before, the primary contribution of this thesis is the development of a set of new algorithms to implement an efficient co-channel speech separation system as shown in Figure 1.1. These contributions can be summarized as follows:

1. Development of a new voicing-state classification method to classify the co-channel speech segment into unvoiced, single-voiced, or double-voiced [2].
2. Testing the performance of the proposed voicing-state classifier at different levels of interfering speech and background noise [3].
3. Development of a time-domain method to precisely estimate the sinusoidal model parameters of co-channel speech [4].
4. Modifying the voicing-state classification method to reduce its computational complexity and make it reasonable for real-time implementation.
5. Modifying a simplified method for sinusoidal parameter estimation of co-channel speech.
6. Testing the performance of the overall system using a variety of subjective and objective measures to evaluate the intelligibility and quality of the processed speech under different values of TIR.

CHAPTER 2

Background and Literature Review

2.1 Introduction

Co-channel speech occurs when overlapping speech signals from multiple speakers are mixed together over a single channel. The process of extracting or separating the desired speech signal from the mixture is commonly referred to as co-channel speech separation (CCSS). Using a CCSS algorithm in the front-end of speech-processing systems can be of great benefit in enhancing the target speech signal in many applications. These applications include automatic speech recognition (ASR), speaker identification (SID), and speech enhancement (SE) techniques. Remarkably, the human auditory system can solve the CCSS problem with a greater accuracy than current machine-based separation systems. Historically, CCSS systems have been developed using algorithms that suppress the interfering signals, enhance the target signal, or estimate both signals simultaneously. So far, there is no one technique that works satisfactorily in all situations.

This chapter presents a brief background and literature survey on the problem of CCSS. It begins, in Section 2.2, with an explanation of the problem and its main applications. A literature survey on previous approaches to solving this problem follows in Section 2.3. Finally, Section 2.4 provides a review of multi-pitch determination algorithms.

2.2 The Problem of Co-Channel Speech Separation

Co-channel speech is defined as the composite signal of two or more speakers. This phenomenon commonly occurs due to the combination of speech signals from simultaneous and independent sources into one signal at the receiver (single microphone). Common situations where co-channel speech may occur can be summarized as follows:

1. When recording the speech of two people speaking simultaneously into a single microphone. An example is speech signals recorded onto an in-flight voice data recording box in an airplane's cockpit.

2. When a target speech is received simultaneously along with audible background voices from other speakers at the hearing aid of a hearing-impaired people.

3. When cross talk occurs while speech is transmitted over an imperfect communication channel.

The human auditory system shows a remarkable proficiency at focusing on a particular speaker or speakers in a mixture (commonly known as the "cocktail-party effect"). Amazingly, human listeners are capable of doing that even when using a single ear or when listening with two ears to a single-channel recording. On the other hand, computer algorithms designed to do the same task have demonstrated only a limited degree of success. The experiments of Miller [5] and Brokx and Nooteboom [6] showed that human listeners can achieve a significant degree of segregation for voices combined in a single channel. In fact, there are many factors involved for the human in segregating concurrent sounds that might not be available for the machine. Some of these factors, as indicated by Cherry [7], include:

1. Spatial information about the direction of the sound source.
2. Visual information such as lip movement and gestures.

3. Tonality information (based on average pitch, speed of speech, the speaker's gender, etc.).
4. Information on accents.
5. Information on transition probabilities (based on comprehension of speech, syntax, etc.).

Except for the last, the above factors can be eliminated in the scenario of co-channel speech when, for example, two messages from the same speaker are recorded simultaneously on a tape. Nevertheless, the human ear can still separate the speech in such an extreme scenario due to the vast memory of transition probabilities that allow humans to predict word sequences.

The goal of co-channel (also called monophonic) speech separation is to recover one or both of the speech signals from the composite signal. A two-speaker model of the co-channel speech separation problem is shown in Figure 2.1. The received *co-channel speech* signal, $x(n)$, is the sum of the *target speech* signal, $s(n)$, and the *interfering speech* signal, $i(n)$.

$$x(n) = s(n) + i(n). \tag{2.2.1}$$

The input target-to-interference ratio (TIR) is defined as the relative power ratio of $s(n)$ to $i(n)$.

$$TIR\ [\text{dB}] = 10\ \log_{10} \frac{\sum_n s^2(n)}{\sum_n i^2(n)}. \tag{2.2.2}$$

In the context of signal processing, CCSS can be viewed in four different ways:

1. Enhancement of the target speech.
2. Extraction of the target speech.
3. Suppression of the interfering speech.
4. Estimation of both speech signals.

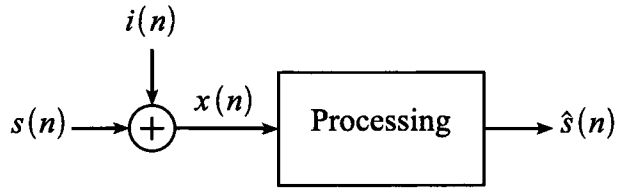


Figure 2.1: The co-channel speech separation problem.

Each of these objectives requires a different approach and would be used in a different application. For example, in automatic speech recognition applications, the TIR is usually high and it is preferable to enhance the target speech rather than to suppress the interference. If the final goal is to use the speech for speaker identification, it is probably sufficient to extract only speech segments in which target speech is minimally degraded by interfering speech (i.e., segments with high TIR). However, if the final speech is to be listened to by someone, the interfering speech would need to be suppressed (cancelled) as much as possible from the mixture with a minimal effect on the target speech. This approach can be applied, for example, in electronic aids for hearing-impaired people in low TIR environments. In this case, the intelligibility of the output speech plays a key factor in evaluating the separated signal. Finally, in other applications, such as analyzing recorded speech, it might be necessary to estimate both waveforms and assign them to the individual speakers.

The three main components of almost all CCSS techniques are shown in Figure 2.2 and described below. In this figure, $\hat{s}(n)$ and $\hat{i}(n)$ are estimates of $s(n)$ and $i(n)$ respectively.

Analysis

In this stage, the objective is to extract the various parameters of co-channel speech components, such as number of speakers, voicing states, and fundamental frequencies (pitches). In *frequency-domain* techniques, for example, the analysis stage consists of windowing and short-time Fourier transform (STFT) to estimate amplitudes,

Figure 2.2: Main components of the CCSS processor.

frequencies, and phases of the harmonic components of individual speech signals. The window size chosen needs to be long enough for individual harmonics to be resolvable, but short enough to minimize frequency broadening of harmonic peaks due to changing pitch.

Separation

This is the main and most important stage of the process. Here, the target speech is separated from the co-channel signal.

Reconstruction

The speech segment of the desired speaker is re-synthesized in this stage from its estimated parameters and the continuous speech stream is reconstructed. The inverse fast Fourier transform (IFFT) and overlap-add (OLA) are common algorithms used in this stage for frequency-domain techniques.

2.2.1 Challenges with CCSS algorithms

The common challenges that affect the performance of any CCSS algorithm can be summarized as follows:

1. Signals with similar power intensities.
2. Closely spaced pitch frequencies (e.g., male/male or female/female).
3. Intersection of pitch contours.
4. Background noise.
5. Complexity and processing delay (for real-time applications).

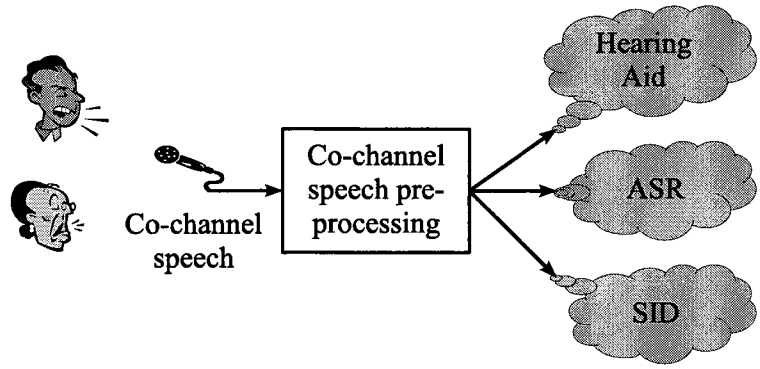


Figure 2.3: Applications of CCSS.

2.2.2 Main applications

As shown in Figure 2.3 below, the main applications of CCSS are:

1. Speech enhancement for hearing-impaired people.
2. Automatic speech recognition (ASR).
3. Speaker identification (SID).

Speech enhancement for hearing-impaired people

The main challenge in this application is that the target speech level is usually lower than the competing speech level (i.e., the input signal has a negative TIR). Intelligibility and quality are important characteristics of the extracted speech in this case and the output TIR would be a useful performance measure [8]. Figure 2.4 (presented in [5] and redrawn by [9]) illustrates listener performance in identifying monosyllabic nonsense words in the presence of up to eight competing speakers. The performance was measured in terms of word percentage correct identification versus masker intensity level. During the experiments, the intensity level of target speech was fixed at 95 dB SPL (i.e., TIR (dB) = 95 - Masker Intensity Level). As can be seen from the figure, the identification accuracy decreases as the number of competing voices or their intensity levels are increased. This is due to the fact that when fewer

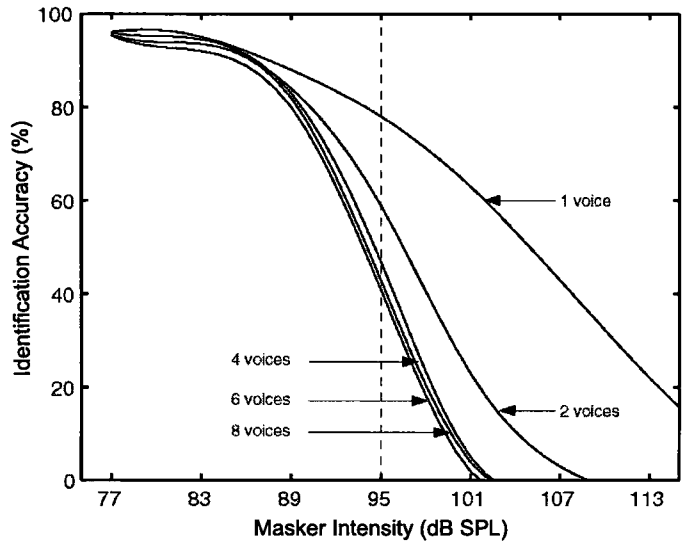


Figure 2.4: Word identification accuracy for human listeners as a function of the number of competing voices and masker intensity. Target speech level was held constant at 95 dB. (After [9], redrawn from [5].)

voices are present, there is a greater chance of the listener hearing segments of the target speech during gaps in the waveform. With the large number of competing voices, on the other hand, the masker level becomes approximately constant over time and the opportunity for hearing the target voice no longer arises.

Automatic speech recognition (ASR)

It is generally acknowledged that speech signals that are corrupted by other speakers represent a more difficult challenge for robust ASR than those corrupted by wideband noise sources. It is noted in [10] that even after applying the separation algorithm proposed by [11], the reduction of word error rate for ASR was unnoticeable. In [12], the performance of an ASR system was measured in terms of recognition accuracy of some relevant keywords at different TIRs ranging from -9 to 6 dB. Figure 2.5 is a redrawing of their results in comparing the ASR system performance with and without speech separation.

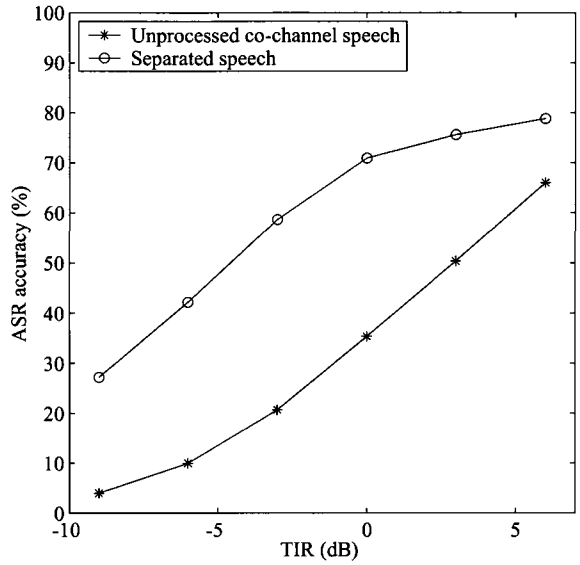


Figure 2.5: Recognition accuracy comparison of ASR system using unprocessed co-channel speech (asterisk-marked line) and separated target speech by the method presented in [12] (circle-marked line).

Speaker identification (SID)

Speaker identification (SID) is an important branch of speech processing. SID is the process of automatically identifying (recognizing) the identity of a speaker by using specific information included in the speaker's speech waveforms. The impact of co-channel speech on the performance of SID systems depends mainly on two factors: the TIR and the amount of overlap between target speech and corrupting speech [13]. Unlike wideband noise, speech interference does not spread the energy constantly over the entire utterance of the target speech. There will still be segments (gaps) in which the target speech is minimally corrupted by the interfering speech. These segments are usually denoted as *usable speech*. As reported in [14–16], the use of usable speech segments alone has shown significant improvement to the performance of the SID system compared to the results obtained using the original mixed signal. Figure 2.6 shows an example of the impact of co-channel speech on the performance of the SID system with and without usable speech extraction [15].

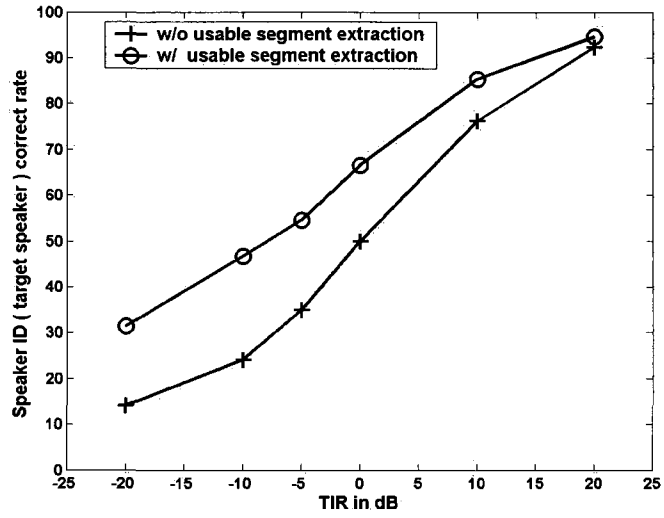


Figure 2.6: SID correct rate with and without usable speech extraction. SID is considered correct when the input speech is identified as the target speaker. (After [15].)

2.3 Previous Work

In the past few decades, extracting individual speech sources from mixtures of different signals has been attractive to many researchers. Despite the tremendous amount of work done to solve the CCSS problem during this period, however, no solution has been sufficiently developed to make its way out of the laboratory [17]. Work that resulted in significant improvement in this area can be traced back to the 1970s.

In general, approaches proposed to solve the CCSS problem can be categorized into four main branches, plus hybrid approaches that are based on these four:

1. General signal-processing approaches
2. Computational auditory scene analysis (CASA) approaches
3. Blind source separation (BSS) approaches
4. Model-based approaches

Typically, general signal-processing and CASA-based approaches seek discriminative

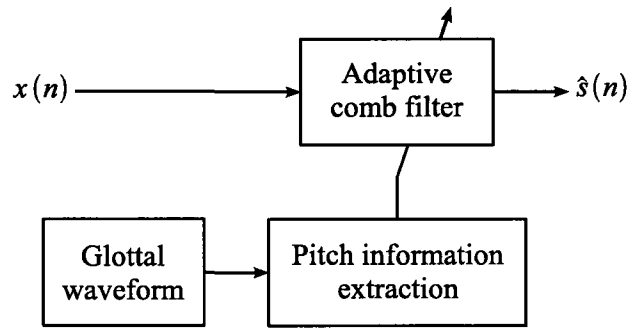


Figure 2.7: Block diagram of the adaptive comb filtering technique for speech enhancement.

features in the observation signal to separate the speech signals. In contrast, the BSS and model-based approaches rely on *a priori* knowledge of sources obtained during a training phase. Hybrid approaches seek the benefits of each individual approach and integrate two or more of them in one system to enhance the overall performance. In the following section, some of the common algorithms under each of the above categories are presented.

2.3.1 General signal-processing approaches

This group of approaches includes all non-auditory time- and frequency-domain algorithms. Signal-processing approaches that try to mimic the human auditory system for speech separation are discussed in the next section.

One of the earliest time-domain methods used for CCSS under this group was the adaptive comb filtering technique proposed in the 1970s by Shields [18] and Frazier [19, 20]. This method is based on the assumption that the speaker's pitch frequency is known *a priori*. The impulse response of a time-variant comb filter is modified according to the pitch frequency of the target speaker, and the filter is used to reinforce the desired signal. Figure 2.7 shows a block diagram of the adaptive comb filtering method as described by Lim [21]. For this technique to work properly, the desired

speech must be considerably stronger than the interfering speech. Unfortunately, evaluation tests run on this technique by Perlmutter *et al.* [22] and Lim [21] showed that the intelligibility of the desired speech after processing is always less than the original unprocessed co-channel speech. Furthermore, performance is degraded when the pitch frequencies vary rapidly.

Parsons (1976) [23] was one of the first researchers to propose a method of suppressing the effect of interfering speech in the frequency domain by means of harmonic selection (HS). In this method, the Fourier transform of the co-channel speech signal is dissected into components belonging to each speaker. The sound of the desired speaker is then reconstructed from its components as identified by the dissection. The process as a whole is outlined in Figure 2.8. The aim of the peak separation process is to identify all the frequency harmonics contained in the spectrum that belong to both speakers. The pitch extraction algorithm is then applied to estimate the fundamental frequency of each speaker. Subsequent to the pitch determination, this information is used to reconstruct the constituent voices. This method shows interesting results when the desired and interference signals have approximately the same power. One of the major drawbacks of this method is that pitch contours must be sufficiently separated. Parsons also indicated that errors in pitch estimation remarkably reduce the intelligibility of the restored speech. Experiments involving listeners with normal and impaired hearing executed by Stubbs and Summerfield [24, 25] using the HS algorithm on all-voiced sentences showed limited intelligibility improvements, especially for hearing-impaired listeners.

The algorithm of harmonic magnitude suppression (HMS) was suggested by Hanson and Wong (1984) [26] to handle a voice-interfered speech signal with a negative-dB TIR. The composite speech was first segmented and Hamming-windowed into 25.6 ms segments with a 50% overlap. For each segment, the interfering pitch period was estimated. Spectral magnitude sampling was used to estimate

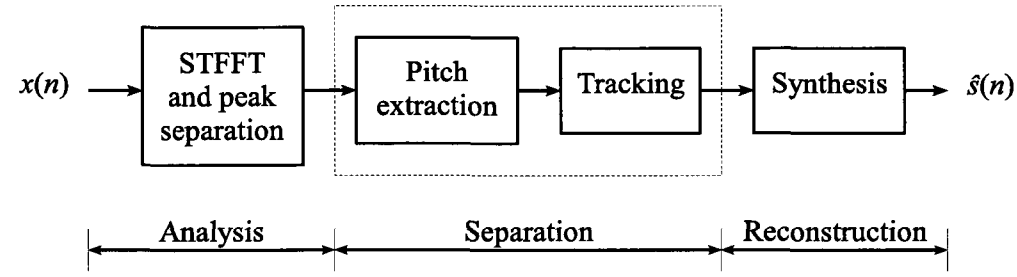


Figure 2.8: Block diagram of the harmonic selection technique [23].

the interfering pitch harmonics magnitudes. The signal magnitude spectrum was then estimated by a spectral magnitude subtraction method as shown in Figure 2.9. Extensive testing indicated intelligibility improvement for interfering speech cases with TIR ranging from -6 to -40 dB. However, the quality of the stronger speech was not retained. The technique still required an *a priori* estimate of the interfering speaker as well as a determination of the voicing state of each speaker.

Naylor and Boll (1987) [27] proposed an enhancement to the HMS system by estimating all model parameters directly from the co-channel speech. Furthermore, they disabled co-channel processing during unvoiced interference to avoid accidental suppression of the target speaker. However, their system was again subject to the assumption of a consistently strong interfering speaker, which is true only for large negative TIRs.

Meanwhile, Lee and Childers (1988) [28] modified Hanson and Wang's algorithm by using a two-stage scheme. In the first stage, they employed the HMS algorithm as a front-end to make initial spectral estimates of each speaker. The second stage used a spectral tailoring technique for minimizing the cross-entropy of the two speakers to obtain better spectral estimation. This work showed only slight improvement over the original method and required the pitch harmonics of the speakers to be well separated for best results.

Zissman and Weinstein (1989) [29] studied the relationship between intelligibility

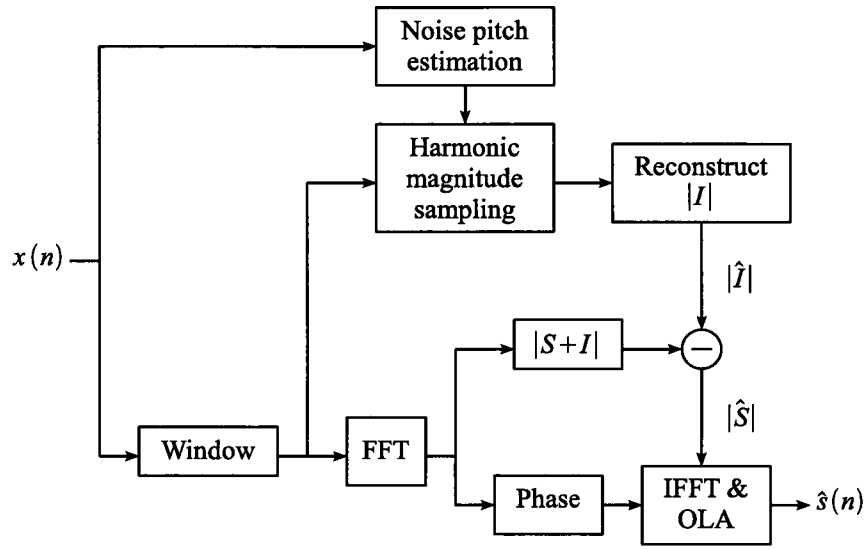


Figure 2.9: Block diagram of the harmonic magnitude suppression technique [26].

and the level of interference suppression during voicing regions of co-channel speech. They focused on signals with TIR ranges from -3 to -15 dB. Simulation results (evaluated by five human listeners) indicated that an interference rejection of 10 to 20 dB provides a significant improvement to target intelligibility. Final results, however, suggested that the effect of interference suppression on target intelligibility is speaker dependent.

Based on the sinusoidal modelling (SM) of speech, McAulay and Quatieri (1990) [30] and Silva and Almeida (1990) [31] used a least-squares estimation technique to calculate the sinusoidal parameters (i.e., amplitude, frequency, and phase) of the speech signal of each speaker. These parameters were assumed constant over the short analysis interval (one frame). Estimated parameters were then used by a synthesizer to reconstruct the speech waveform of the desired speaker. This technique was successful, but generally suffered from numerical stability problems in the presence of closely spaced component frequencies. As described in [30], the algorithm could be implemented using two different approaches: *peak picking* and *frequency sampling.*

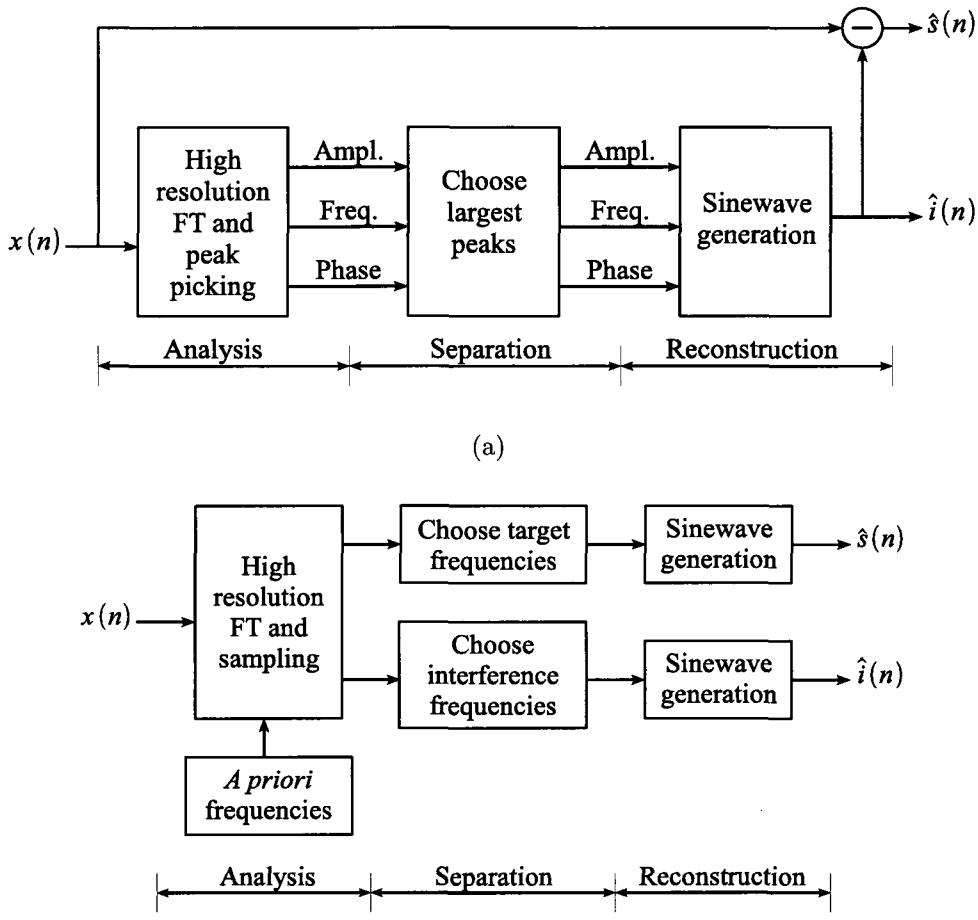


Figure 2.10: Block diagram of the sinusoidal modelling technique [30]: (a) peak-picking approach, and (b) frequency-sampling approach.

In the peak-picking approach, no *a priori* information about harmonic frequencies of the individual speakers was assumed. As shown in Figure 2.10(a), the largest peaks of the summed spectra were chosen and were used to reconstruct the larger of the two waveforms (interfering speech). The waveform estimate was then subtracted from the input waveform to produce an estimate of the lower signal (target speech). This approach showed poor separation due to spectral overlapping and required that the target signal has much lower energy than the interferer.

In the frequency-sampling approach [30,31], on the other hand, *a priori* knowledge

of the sine-wave frequencies of each of the two speakers was assumed. The frequency sets of each individual speaker were obtained either by peak picking or by estimating the pitch frequency of each speaker. Frequency sets were then used to sample the summed spectra to obtain amplitude and phase estimates for the sine-wave representation of each waveform. An estimate of the desired waveform could then be directly reconstructed or obtained by subtracting the reconstructed larger waveform from the summed waveforms. (See Figure 2.10(b).) In spite of the large *a priori* information required about pitch information for both the target and interfering speaker, enhancement over the peak-picking approach was small [30].

In 1991, Gu and Bokhovan [32] proposed a CCSS algorithm using frequency bin nonlinear adaptive filtering along with a multi-pitch estimation technique based on the hidden Markov models (HMM). The system showed some success in separating synthetic speech signals with fixed pitches. However, the performance noticeably deteriorated when real speech signals with natural pitches were used.

Naylor and Porter (1991) [33] investigated a speech separation algorithm that required no *a priori* information. The algorithm utilized a modified covariance spectrum estimator to estimate pitch frequencies of the two speakers. Next, a linear estimation technique was used to solve the complex spectrum of the co-channel signal into individual speaker components. Harmonic location error was worse at high frequencies, making this approach more sensitive to additive noise.

In 1994, Savic *et al.* [34] developed a co-channel speaker separation system based on maximum-likelihood deconvolution (MLD) that was applied to the input speech to estimate the excitation signal of each speaker. The restored speech of a particular speaker was produced by convolving the excitation signal with the vocal tract filter which was assumed to be known *a priori.*

The work by Morgan *et al.* (1995) [35] presented a CCSS system using the harmonic enhancement and suppression (HES) technique. This system used a

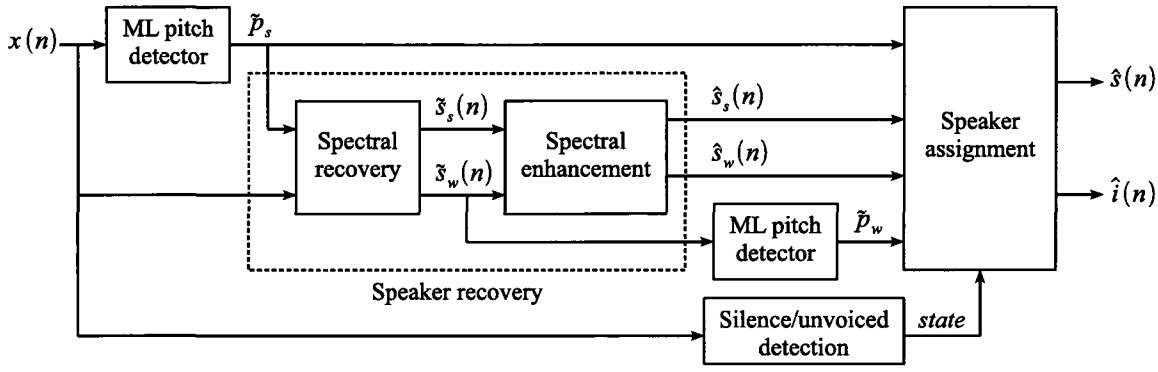


Figure 2.11: Block diagram of the HES co-channel speech separation system [35].

maximum likelihood (ML) pitch detector [36] that had been modified to avoid pitch-doubling errors. It provided an integer estimate of the pitch period of the stronger speech signal in a 40 ms analysis interval. Based on this initial integer estimate of the pitch period, a multi-resolution search was conducted to determine a fractional pitch period. The fractional pitch period was then used to construct two filter pairs in the frequency domain. These filters were applied to the co-channel signal spectrum and used to separate the stronger and weaker speakers, respectively. The recovered weaker signal, or residual signal, was further processed by suppressing energy at frequencies corresponding to the stronger speaker's formants. The recovered stronger and weaker signals were then assigned to the target or interfering speaker using a maximum likelihood speaker assignment (MLSA) algorithm and re-synthesized using the OLA technique. (See figures 2.11, 2.12, and 2.13 [11].) This technique is highly complex and, in addition, cannot be applied for signals with equal power intensities.

Benincasa (1997) [37] developed an algorithm for CCSS based on constrained nonlinear optimization (CNO). This method separated overlapping voiced speech signals by determining the best possible parameters (frequency, phase, and amplitude) for the harmonics of both speakers that provide the least mean square error (LMSE) between the original co-channel speech and the sum of the reconstructed

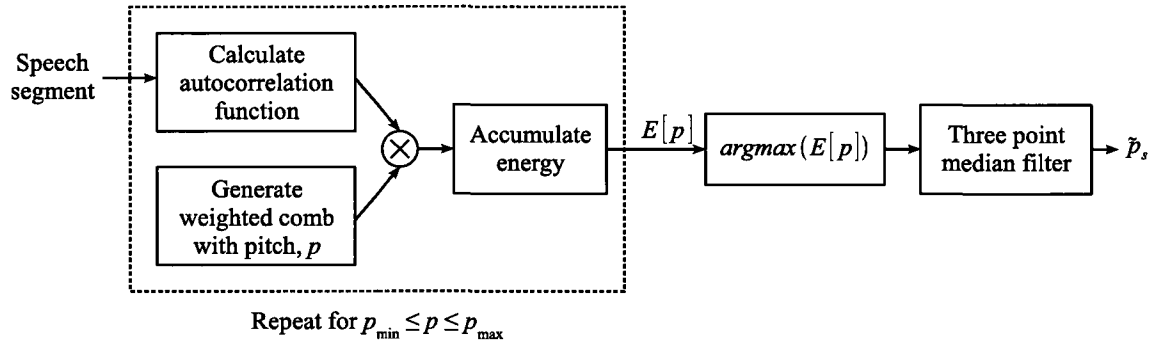


Figure 2.12: Block diagram of the ML single-speaker pitch detection algorithm [35].

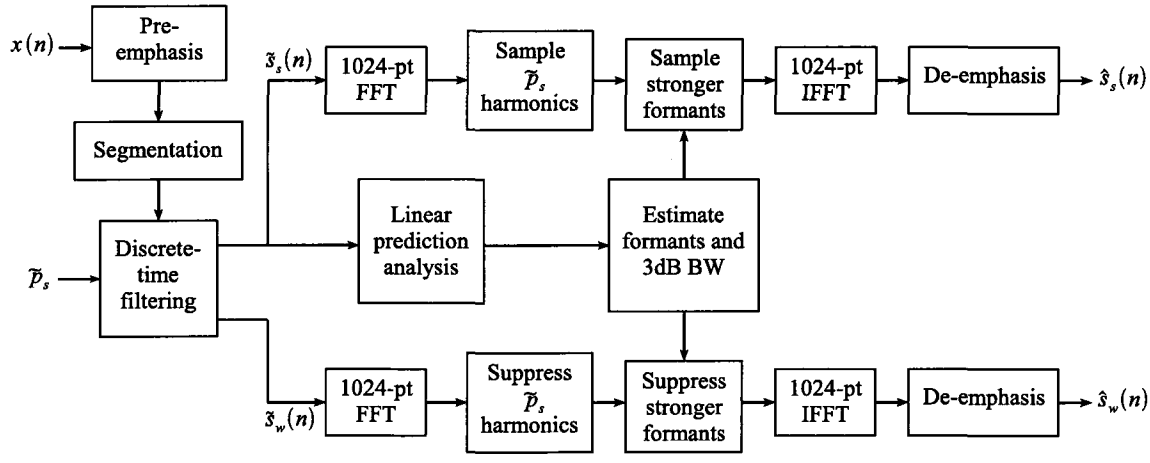


Figure 2.13: Block diagram of the HES speaker recovery system [35].

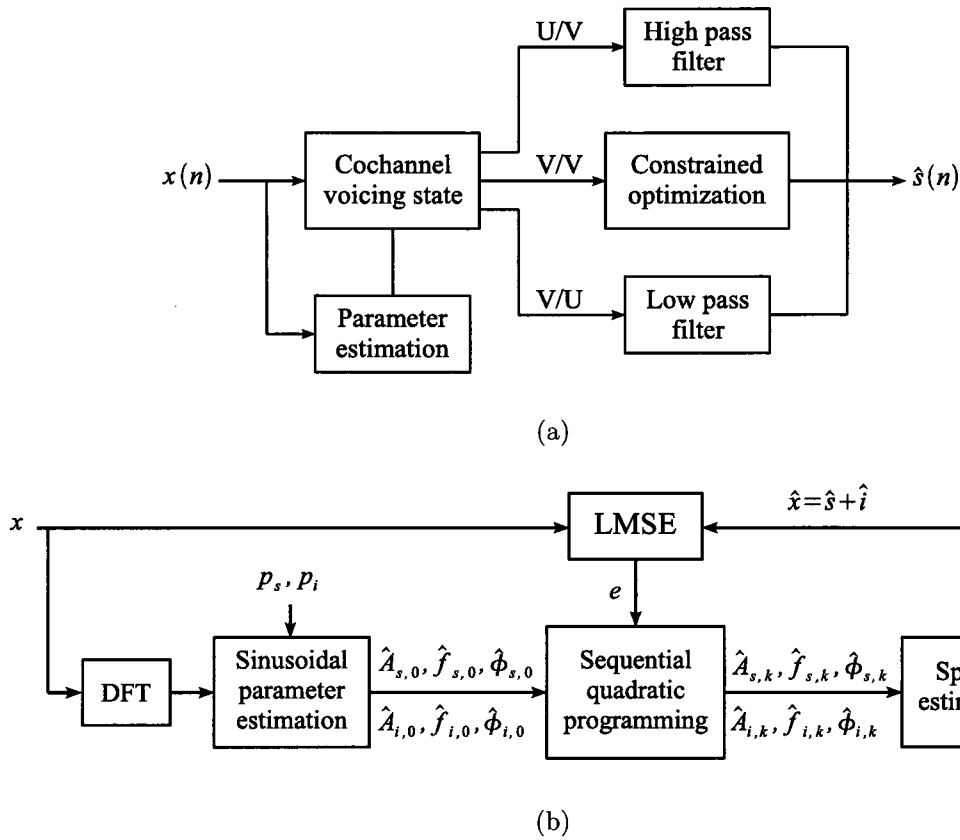


Figure 2.14: Block diagram of the constrained nonlinear least-squares speech separation system [37]: (a) system diagram, (b) constrained nonlinear least-squares optimization.

speeches. This is done in the frequency domain as shown in Figure 2.14 (a) and (b). Reconstruction of the desired and interfering speech is accomplished using an OLA technique. A drawback of this algorithm is that all voicing states and pitch frequencies of both speakers were assumed to be known *a priori* in order to estimate initial conditions of the harmonic parameters.

Between 1998 and 2005, extensive work by Yantorno and his colleagues addressed the effect of co-channel speech on SID systems [8]. Their research was focused mainly on extracting portions of co-channel speech in which the energy of the target speaker is much greater than the energy of other interferers. They referred to these portions as "usable speech," while the other portions were called "unusable speech." Speech

segments having a TIR magnitude of 20 dB and above were considered usable. Since TIR measure cannot be determined directly from co-channel speech in real applications, their approach was to find measures that correlate as much as possible with the TIR in detecting the usability of co-channel speech. Different measures were developed to improve usability detection and, consequently, speaker identification accuracy [38]. Some of these proposed measures are listed below:

- Spectral autocorrelation peak-to-valley ratio (SAPVR) [39, 40]
- Local kurtosis [41]
- Adjacent pitch period comparison (APPC) [42]
- Cyclostationarity and wavelet transform [43]
- Linear predictive analysis [44, 45]
- Difference-mean comparison (DMC) and nodal density (ND) [46]

To ensure that an efficient usable speech detection system was developed, several usable speech detection features containing complementary information were fused [38, 47]. As a result, the overall performance of the detection system was enhanced. The application of all techniques based on usable speech measures, as mentioned above, was limited to SID systems. Since the concept of usable speech is by definition application dependent, what is considered usable for speaker identification may not be usable for speech recognition and vice versa.

Based on the same concept of sinusoidal modelling of speech, Heming (2000) [48] expanded the subharmonic summation (SHS) technique [49] proposed for single-pitch detection to detect multiple pitches of overlapping voiced speech. Then a two-dimensional harmonic sieve arithmetic was exploited to estimate the sinusoidal model parameters. No reliable results using this technique were presented.

2.3.2 CASA approaches

According to Bregman [50], the human auditory system performs what is called *auditory scene analysis* (ASA) to separate the acoustic signal of mixed sounds

(cocktail party) into streams corresponding to individual sources. The ASA principle can be viewed as a two-stage process. In the first stage, the acoustic mixture is decomposed into elements that present significant acoustic events. Subsequently, a grouping process combines elements that are likely to belong to the same acoustic source, forming a perceptually meaningful structure called a *stream.* For a good review of the cocktail-party effect and ASA concept, see [51] and [17].

Inspired by the ASA principle, considerable work has been done in the last few years to build a computational auditory scene analysis (CASA) separation system. Fundamentally, CASA techniques were developed to separate mixtures of any sound sources (including speech) using signal-processing approaches in the same way that human listeners do [52].

In CASA methods, the input signal is initially decomposed into a matrix of time-frequency (T-F) cells via a bank of auditory bandpass filters (such as the gammatone filterbank [52]) in which bandwidths increase with increasing centre frequencies. This filterbank structure mimics cochlear filtering in the human auditory system. Sound source separation is then achieved by two main stages similar to ASA: *segmentation* (analysis) and *grouping* (synthesis). In the segmentation stage, the adjacent T-F cells that have common acoustic features are merged into two-dimensional segments. Those segments that are believed to belong to the same source are then grouped together in the grouping stage. The grouping process is mostly based on *cues* (sound properties) such as periodicity, amplitude modulation, common onset and offset, spatial location, and continuity. The structure of a typical CASA system is shown in Figure 2.15 [53, 54]. It consists of four stages: front-end analysis, feature extraction, segmentation, and grouping. The front-end processing decomposes the auditory scene into a time-frequency representation via bandpass filtering and time windowing. The second stage extracts auditory features corresponding to ASA cues, which will be used in subsequent segmentation and grouping. In segmentation and grouping, the system generates segments for both target and interference and

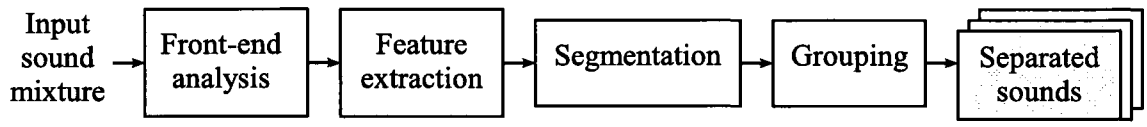


Figure 2.15: Schematic diagram of a typical CASA system.

groups the segments originating from the target into a target stream. A stream corresponds to a single sound source. Grouping itself is comprised of simultaneous and sequential organization. Simultaneous organization involves grouping of segments across frequency while sequential organization refers to grouping across time. Finally, the waveform of the segregated target can then be re-synthesized from the target stream.

Most CASA approaches rely on pitch estimation as the most powerful cue for speech separation [55–63]. For example, in 1986 Weintraub[1] [55] attempted to separate the voices of two speakers (male and female) by tracking their pitches using a dynamic programming algorithm. This algorithm was applied to auto-coincidence (a version of autocorrelation) functions of the output channels of a cochlear filterbank. Within a time frame, a Markov model then used these pitch tracks to determine the number of speakers as well as the voicing state of each speaker (i.e., silent, periodic, or non-periodic). Simulation results in [55] showed a significant performance degradation when the pitch tracks were estimated directly from the mixed speech signal compared to when *a priori* pitch tracks extracted from isolated speech signals before mixing were used.

Meddis and Hewitt (1992) [56] proposed a computational model to identify only two simultaneous vowels. First, a bank of 100 bandpass filters was adopted to simulate the cochlear response. The output of each filter was further transduced by a model of inner hair cells. Two-dimensional autocorrelation functions (ACFs) of the output

[1]It is widely believed that Weintraub was among the first researchers to develop a CCSS system using a computational auditory model, even before the term "CASA" was introduced.

channels were then utilized to extract pitch information. All ACFs were summed over frequency to obtain a summary autocorrelation function (SACF) where its peak was used to derive an estimate for the pitch of the stronger vowel. Individual ACFs showing peaks close to this value were combined and used to identify the first vowel using a template-matching procedure. The ACFs in the remaining channels were then combined and used to identify the second vowel.

In 1993, Cooke [57] proposed a CASA model that first generates local segments based on filter response frequencies and temporal continuity. These segments were then merged into groups based on common harmonicity and common amplitude modulation (AM). A pitch contour was then obtained for each group, and groups with similar pitch contours were put into the same stream.

Ellis (1996) developed a prediction-driven system that generated predictions using a world model and compared the predictions against the input [58]. The world model included three types of sound elements: noise cloud, transient click, and harmonic sound.

In 1999, Wang and Brown presented a CASA model to separate voiced speech from background interference based on oscillatory correlation. They used temporal continuity and harmonicity as major grouping cues in their system. First, segments were formed on the basis of similarity between adjacent filter responses (cross-channel correlation) and temporal continuity. Then grouping among segments was performed according to the dominant pitch extracted within each time frame. Finally, target and interference streams were segregated by means of oscillatory correlation. A stream was represented by a group of synchronized oscillators and different streams were represented by desynchronized oscillators [59]. Their simulation results using a common corpus (presented in [57]) of acoustic mixtures of voiced utterances and different types of interference, including white noise, "cocktail party" noise, and competing speech, indicated good results in comparison with previous CASA systems.

In most cases, the model was able to remove the interference and recover the low-frequency (below 1 kHz) portion of voiced target speech. The system, however, failed in handling the high-frequency (above 1 kHz) part of target speech and lost much of it.

The system described by Ottaviani and Rocchesso (2001) [60] consisted of two main stages: pitch analysis and signal re-synthesis. In the analysis stage, pitch tracks were estimated based on the enhanced summary autocorrelation function (ESACF) proposed by Tolonen and Karjalainen in [64]. The enhancing operation for the SACF gave prominence to the candidate peaks for pitch estimation and consequently reduced false detection. Based on pitch information computed in the first stage, the speech signal was further analyzed in the second stage using a highly zero-padded Fourier spectrum, which was selectively weighted to emphasize harmonics of the detected pitch. Finally, the separated speech signal was re-synthesized using inverse Fourier transform.

Hu and Wang (2004) [61] tried to solve the problem of high-frequency components in Wang and Brown CASA systems [59] by employing different methods to segregate resolved (low-frequency) and unresolved (high-frequency) harmonics of the voiced target speech. More specifically, they generated segments for resolved harmonics based on temporal continuity and cross-channel correlation. These segments were grouped according to common periodicity. Segments for unresolved harmonics, on the other hand, were generated based on common AM in addition to temporal continuity. These segments were further grouped based on AM rates, which were obtained from the temporal fluctuations of the corresponding response envelopes.

In 2006, Runqiang *et al.* proposed a speech separation system for speech recognition [62]. After preprocessing the input signal with the auditory peripheral filtering, dominant pitch tracks of the target speech were estimated using two features: normalized correlogram and frequency dynamic. Next, an initial grouping based on Hu and Wang's algorithm [61] was performed to produce initial streams. A regrouping

strategy was employed to refine these streams via amplitude modulation cues, which were finally organized by speaker recognition techniques into corresponding speakers. Finally, the output streams were reconstructed to compensate the missing data by a cluster-based feature reconstruction.

Although most monaural CASA systems have utilized pitch-based cues in segmentation and grouping, there have been other attempts to exploit other cues such as common onset and offset [52, 63, 65] and frequency modulation [66]. Onsets and offsets are important ASA cues for the reason that different sound sources in an environment seldom start and end at the same time. In addition, there is strong evidence for onset detection by auditory neurons. There are several advantages to applying onset and offset analysis to auditory segmentation. In the time domain, onsets and offsets form boundaries between sounds from different sources. Common onsets and offsets provide natural cues to integrate sounds from the same source across frequency. In addition, since onset and offset cues are common to all types of sounds, the proposed system can in principle deal with both voiced and unvoiced speech.

In a recent publication [66], Gu and Stern described a new approach to CCSS based on the detection of modulation frequencies and the grouping of sets of frequencies that appear to be co-modulated, regardless of the extent to which they are harmonically related to each other. Nevertheless, accurate estimation of modulation frequency can be quite difficult for natural signals.

A common problem in many CASA systems is the difficulty in dealing with the high-frequency part of target speech above 1 kHz [61]. This is due to the fact that in the high-frequency range, harmonics are generally unresolved since the corresponding auditory filters have wide passbands. Also, the grouping process is a difficult task such that in current approaches, one of the underlying signals is assumed to be fully voiced in order to ease the grouping [67, 68]. Another problem with CASA techniques is that they cannot replicate the entire process performed in the auditory system since

the process beyond the auditory nerve is not well known. For further information on CASA techniques refer to [52, 69].

2.3.3 Blind source separation (BSS) approaches

Blind source separation is a statistical technique which attempts to recover a set of original source signals from observed instantaneous mixtures of these sources. BSS relies on a main assumption that the mixing process must be linear. A standard approach to the BSS problem is independent component analysis (ICA) [70] and its extensions, in which we assume that the sources are all statistically independent of one another. In general, the ICA approach to BSS tries to invert the mixing process (demixing), to reconstruct the original components, by finding a linear transform of the mixtures such that the recovered signals are as independent as possible. ICA techniques for speech separation are not inspired by auditory analysis and require two conditions to solve this problem. First, the number of available mixtures must be equal to or greater than the number of original unknown sources (overdetermined BSS). Second, the sources must be perfectly aligned. Furthermore, BSS approaches usually apply machine-learning techniques to estimate the demixing matrix. Therefore, BSS algorithms are not commonly used in CCSS since these assumptions are difficult to satisfy in this case. In addition, CCSS algorithms cannot rely on the spatial location of the speakers to obtain a binaural input, as do algorithms that use multiple microphones. On the contrary, general signal-processing and CASA approaches require fewer conditions and, therefore, are more flexible in dealing with the CCSS problem. A comparison between the performance of CASA and BSS approaches for speech segregation can be found in [71].

To overcome the constraint of the number of observations in order to deal with the single-channel source separation problem, some researchers have recently proposed what are called underdetermined BSS techniques [72]. In these techniques, auxiliary information (such as *a priori* knowledge of statistical models of the sources) is

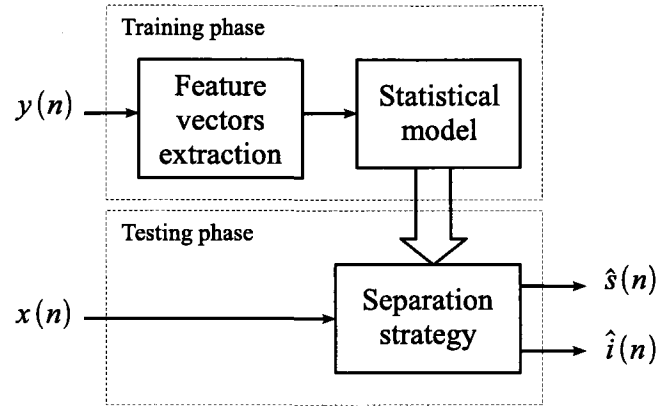


Figure 2.16: Schematic diagram of a single-channel BSS system [76].

commonly used to solve the problem. Figure 2.16 depicts a general schematic diagram for an underdetermined BSS techniques. The process consists of two phases: the training phase and the testing phase. In the training phase, the speech sources are projected onto a set of basis functions by using ICA [73], nonnegative matrix factorization (NMF) [74], or sparse coding [72]. The basis functions are chosen such that the projection coefficients are as sparse as possible. This means that only a few coefficients of the decomposition of the sources on the basis functions are significantly nonzero. In the testing phase, given the basis functions and coefficient distributions, techniques such as maximum a posteriori (MAP) estimation are used to estimate the sources. As indicated in [75], these techniques do not work well when the trained basis functions of the two sources overlap, which is the case in speech mixtures.

An example of ICA-based algorithms for underdetermined BSS given a single-channel recording is that presented by Jang *et al.* (2003) [73]. In this algorithm, a set of time-domain basis functions that encode the source signals were learned *a priori* from a training data set and then used to separate the unknown test sound sources. The original source signals were recovered by maximizing the log likelihood of the separated signals, calculated using the basis functions and the probability

density functions of their coefficients–the output of the ICA basis filters. Simulation results [73], using only six mixtures of different combinations of two music signals and two all-voiced speech signals, showed poor recovery performance of speech mixtures compared to mixtures containing music. In fact, this was due to the faster variation of speech signals and basis functions with time compared to music.

2.3.4 Model-based approaches

Recently, model-based speech separation and de-noising techniques has been an attractive topic for many researchers in the field of robust speech recognition. Model-based CCSS techniques are spiritually similar to model-based single-channel speech enhancement techniques. In this case, CCSS can be considered as a speech enhancement problem in which both the target and interference, which are non-stationary sources with similar probabilistic characteristics, must be estimated. Briefly, the following procedures are commonly applied in model-based CCSS techniques. First, patterns of the sources are obtained in the training phase. Then, those patterns whose combinations model the observation signal are chosen. Finally, the selected patterns are either directly used to estimate the sources [76–78] or used to build filters that when imposed on the observation signal result in an estimate of the sources [79–81]. Model-based, as well as underdetermined systems, rely heavily on the use of *a priori* information of sound sources. Therefore, they are fundamentally limited in their ability to handle novel interference.

2.3.5 Hybrid approaches

Although the above-mentioned approaches differ in solving the same problem, there are possibilities for combining any two of them to enhance the overall performance of the CCSS system. An example of combining CASA and BSS techniques is provided by Radfar *et al.* (2007) [76], who described a system in which underdetermined BSS was used as a grouping technique for the CASA approach. The authors reported good

performance for the separation of two speech signals in a single recording compared to the performance expected using CASA or BSS alone.

Roweis (2001) combined a model-based learning algorithm with a CASA system to enhance the separation of mixtures from two speakers [79]. He first trained speaker-dependent HMMs on the speech of isolated speakers. Next, these pre-trained models were combined into a factorial HMM (FHMM) consisting of two Markov chains that evolve independently. Given a co-channel speech mixture, separation was achieved by first inferring the underlying state sequence in the FHMM and computing the output predictions for each Markov chain. A binary mask for the CASA model was then determined by comparing the relative magnitudes of each model's predictions.

From the above review, it is readily seen that CCSS is a broad topic and has been addressed by different approaches. This thesis will focus mainly on the general signal-processing approaches to the CCSS problem.

2.4 Multi-Pitch Determination

Perceptually, one of the most powerful cues for CCSS is the *fundamental frequency* (pitch) of voiced speech. Specifically, human listeners are able to exploit the difference in pitch in order to separate the harmonics of the target voice from those of interfering voices. Consequently, much of the work on CCSS algorithms has focused on the problem of determining and tracking the multiple pitches present in the speech mixture (so-called multi-pitch analysis) as the first problem to be attacked in developing a speech separator. Pitch information is used in subsequent stages to separate the constituent signals. If the estimated pitch was not accurate enough, the resulting signal cannot provide good quality. Although a variety of automatic pitch determination algorithms (PDAs) are now known, they are generally used for single-speaker situations and work best at high signal-to-noise ratios (SNRs.) In a two-speaker scenario, each speaker is a source of extraneous and potentially misleading information about the other. Hence, the use of single-pitch determination algorithms

in such a case greatly increases the probability of error, especially when the voices are of comparable loudness and the pitches are nearly equal. Multi-pitch determination algorithms (MPDAs) are preferable to use in this case.

Generally, MPDAs can be classified into three categories: time-domain, frequency-domain, and time–frequency-domain algorithms. Time-domain PDAs directly examine the temporal structure of the speech waveform. Typically, peak and valley positions, zero-crossings, autocorrelations, or residues of comb-filtered signals are analyzed for detecting the pitch period. Frequency-domain PDAs distinguish the fundamental frequency by utilizing the harmonic structure in the short-term spectrum. Time–frequency-domain algorithms perform time-domain analysis on band-filtered signals obtained via a multichannel front-end. Practically, the main difficulty with almost all MPDAs is estimating the pitch of the weaker speaker.

2.4.1 Time-domain MPDAs

Time-domain MPDAs, generally rely on cancellation rather than autocorrelation. For example, the method proposed by de Cheveigné (1991) [82] is a two-dimensional extension of the classical average magnitude difference function (AMDF) algorithm [83] for single-pitch determination. Two cascaded comb filters of impulse responses $\delta(n) - \delta(n + d_A)$ and $\delta(n) - \delta(n + d_B)$ were applied to the input signal, where $\delta(n)$ is the delta function and d_A and d_B are the lag parameters, in samples, of the comb filters respectively. Accordingly, the two filters have nulls at frequencies $f_A = F_s/d_A$ and $f_B = F_s/d_B$ and all their multiples, where F_s is the sampling frequency. Pitch analysis was therefore performed by applying the filters to the co-channel signal on a frame-by-frame basis and searching the double difference function (DDF)

$$\mathrm{DDF}(d_A, d_B) = \sum_{frame} |x(n) - x(n + d_A) - x(n + d_B) + x(n + d_A + d_B)| \qquad (2.4.1)$$

for a global minimum over the two-dimensional lag parameter space. The coordinates of this minimum give the estimates of the two pitch periods. To avoid searching

other local minima of lags equal to period multiples, the algorithm restricted the search to lags corresponding to the feasible range of human pitch. De Cheveigne and Kawahara [84] evaluated their algorithm on mixtures consisting of perfectly periodic waveforms, with excellent results. However, their joint cancellation technique has certain limitations. It is sensitive to sampling rate, and the cancellation of one speaker may partially cancel the other speaker if their pitch periods are related by an integer multiple.

Chazan *et al.* (1993) [85] proposed another time-domain MPDA for detecting pitch periods for CCSS using two nonlinear comb filters connected in a cross-coupled scheme in conjunction with an estimate-maximization (EM) iterative method. The pitch of one speaker was obtained by suppressing the estimated pitch harmonics of the other speaker. Iterative parameter estimation worked as follows. First, the mixed signal was fed to the first comb filter and its parameters were computed so as to minimize the output error. After determining the parameters of this comb filter, its output was then subtracted from the mixed signal to obtain an estimate of the second speaker. Similarly, the estimated signal was fed to the second comb filter and the same process was repeated. This process iterated back and forth to find the best parameters for the two nonlinear comb filters. Final filter parameters were then extracted as the pitch candidates. The algorithm used relatively long frames (60 ms) to effectively estimate pitches of speakers with close spectral harmonics. In order to handle continuous pitch variations within the long analysis frame, a time-warped signal was used to model each speaker.

2.4.2 Frequency-domain MPDAs

Perhaps, the earliest MPDA for separating two concurrent speakers was the system described by Parsons (1976) [23]. In this approach, multi-pitch estimation was achieved using a modified version of Schroeder's histogram method [86] for single-pitch detection. In the two-speaker case, a histogram containing all integer submultiples

of all the peaks in the frequency spectrum was formed. Then, the histogram was searched for the highest entry that corresponds to the pitch of the stronger speaker. All harmonics of this pitch were identified and omitted from a second formed histogram. The highest entry of the new histogram provided an estimate for the pitch of the weaker speaker.

Kwon *et al.* (2000) [87] proposed a simplified system for determining multi-pitch of mixtures of two speech signals based on the sinusoidal model of speech [30]. Their algorithm consisted of three steps. First, all peaks of the frequency spectrum that contribute significant power compared to signal power were selected. Then, pitch candidates were generated from these peaks using a linear searching algorithm. Finally, the pitch was estimated by selecting the candidate that had passed the maximum number of peaks. Despite the relatively low computational complexity of this algorithm, it did not solve the problem of overlapping peaks.

Recently, techniques using statistical learning approaches for feature extraction were also applied to the problem of multi-pitch determination. Sha and Saul (2005) [88] described a real-time MPDA based on prior information about speech sources obtained via learning in a training phase. In this algorithm, fixed basis functions (each of which corresponds to a single pitch value) were trained offline using the NMF algorithm to model the particular timbres of voiced speech. During the testing phase, the mixed signals were then decomposed in terms of these basis functions to detect the presence of one or two voices and determine their pitches.

The algorithm proposed by Radfar *et al.* (2006) [89] exploited the same sinusoidal and harmonic modelling of speech signals by Quatieri [30] to detect and track the pitch contours. First, a set of pitch candidates was calculated using a sinusoidal spectrogram. Next, the best pitch contours that minimize the error between harmonic models of estimated signals and the spectral magnitude of the original mixed signal were picked. Overall results compared to Wu [68] and Chazan [85] techniques showed a slight reduction in pitch error rate in the range of -9 to 9 dB TIR.

2.4.3 Time–frequency-domain MPDAs

Time-frequency domain MPDAs are mainly inspired by the human auditory system. The temporal model of pitch perception proposed by Licklider [90] was the basis of most of these algorithms. For example, Weintraub (1986) [55] described a computational implementation of Lickliders theory using the so-called auto-coincidence functions of the output channels of a cochlear filterbank. Auto-coincidence functions were used in a dynamic programming algorithm to estimate and track the pitch period of each speaker. The large number of channels used (85) caused this approach to be computationally inefficient. However, Tolonen and Karjalainen (2000) [64] suggested a computationally efficient multi-pitch model using the same concept based on the SACF. (See Section 2.3.2.) Computational savings were made by splitting the input signal into two bands (below and above 1 kHz) rather than performing a multi-band frequency analysis. A generalized autocorrelation is then computed for the low-frequency band and for the envelope of the high-frequency band, and added to give a SACF. Further processing is then performed to enhance the representation of different pitches. Specifically, the SACF is half-wave rectified and then expanded in time by a factor of two, subtracted from the original SACF, and half-wave rectified again. This removes peaks that occur at sub-octave multiples and also removes the high-amplitude portion of the SACF close to zero delay.

One of the sophisticated algorithms for tracking the pitch of multiple speakers is proposed by Wu *et al.* (2003) [68]. Their approach consisted of four stages:

1. The input signal was filtered by a bank of gammatone filters to simulate cochlear filtering. In low-frequency channels (below 800 Hz), the correlogram was computed directly from the filter outputs, whereas in high-frequency channels the envelope in each channel was autocorrelated.
2. Clean correlogram channels (i.e., those likely to contain reliable information about the periodicity of a single speaker and relatively uncorrupted by noise) were identified.

3. A statistical approach was used to estimate the pitch periods present in each individual time frame.
4. Pitch periods were tracked across time using HMM.

This system proved to be robust to the presence of an interfering speaker and can track its pitch. However, Khurshid and Denham [89] found that it is less robust against background noise and that, while it accurately tracks the pitch of the dominant speaker in a mixture, its estimate of the non-dominant pitch can be poor.

To correct some of these problems, Khurshid and Denham (2004) [91] suggested an alternative approach based on the analysis of the output of a bank of damped harmonic oscillators that model the frequency analysis performed by the cochlea. Analysis of the fine time structure (consecutive zero crossings and amplitude peaks) of each oscillator output was performed to determine the driving frequency. An algorithm was then used to hypothesize the multiple pitches present in order to explain the observed frequency components. This was achieved by identifying salient spectral peaks and then assessing the support for every subharmonic of the peak that falls within the normal range of voice pitch. Such a frequency remapping leads to noise robustness and may be regarded as a simple model of simultaneous masking in the auditory nerve. Simple continuity constraints were used to track two pitches over time. Khurshid and Denham's comparison of the two systems showed that Wu *et al.*'s system more accurately tracked the dominant pitch, but that their own system more reliably tracked the non-dominant pitch and was more robust to noise.

2.5 Summary

This chapter reviewed the principles and major applications of CCSS systems and presented a brief survey of the research done in this area. The main goal of a CCSS algorithm is to automatically process the mixed signal in order to enhance the target speech, cancel the interfering speech, or recover each speaker's original speech. The use of CCSS techniques in the front-end of speech processing systems is a very useful tool in many applications. These include human-based applications such as electronic aids for people with hearing disabilities and machine-based applications such as ASR and SID systems. Several previous studies have developed algorithms for modelling and separating co-channel speech under different domains such as general signal-processing approaches, computational auditory scene analysis (CASA) approaches, blind source separation (BSS) approaches, and model-based approaches. In the general view, any CCSS algorithm consisted mainly of three stages: analysis, separation, and reconstruction. As shown in this chapter, some CCSS systems required some sort of *a priori* information on one or both of the mixed speech signals. The pitch contours of individual speakers were widely used as *a priori* information. This, in fact, limits their capability to work in different applications. Other algorithms tried to estimate pitch tracks directly from the co-channel signal using a suitable MPDA.

CHAPTER 3

Speech Processing Using State-Space Reconstruction

3.1 Introduction

One might think that analysis in the frequency domain should be the core of the work in this thesis. Indeed, as shown in the previous chapter, many CCSS approaches are based on frequency analysis of speech signals. However, this chapter demonstrates why the conscious renunciation of the Fourier transform is considered to be the strength of the methodology proposed to dealing with this problem. As opposed to conventional frequency-domain techniques, the main analysis approach exploited in this work is based on the state-space reconstruction of speech.

The concept of state-space representation has become quite popular as a result of the recent increase of research in the field of nonlinear dynamic systems, particularly chaotic systems[1] [92]. Using state-space representation, the system is described in terms of state variables using a set of differential equations and represented, geometrically, in an m-dimensional space where the axes are labelled with the state variables (degrees of freedom).

In practical situations where the state variables and the describing differential

[1]Chaotic systems are deterministic nonlinear systems that are very sensitive to initial conditions. Their output signals often appear to be random and irregular in time and frequency domains. However, chaotic signals are slightly predictable and give a structured and deterministic state-space.

equations of the system are often unknown (or perhaps unmeasurable), it is possible to reconstruct the state-space using only single observations from the system as a function of time. This is commonly referred to as state-space reconstruction and is performed by using the embedding theorem.

State-space reconstruction is a relatively new tool for analyzing speech and audio signals. Originally, this method was used to better understand the chaos of strange attractors and other non-periodic systems [92]. However, it also can be used to observe the regularity of periodic and quasi-periodic signals like speech. Visualizing audio and speech signals in state-space generates interesting shapes and provides a level of insight into the nature of sounds that is not available in the standard Fourier representation. In other words, the motivation behind using such a representation is to explicitly reveal properties that are implicit in the acoustic signal and might be ignored in conventional spectral analysis. Reconstruction of state-space is the fundamental basis of the algorithms proposed for the CCSS problem in the succeeding chapters of this thesis.

This chapter furnishes the reader with the necessary background theory and techniques related to the concepts of state-space reconstruction for nonlinear systems and its application to speech signal processing. First, the notion of state-space representation is introduced in Section 3.2, since nonlinear characterization of dynamic systems is based mainly on this concept. The embedding theorem and methods for determining optimal embedding parameters are also reviewed in this section. In Section 3.3, speech is demonstrated as a nonlinear process, giving the motivation to pursue speech processing using state-space reconstruction. At the end of this section, recent work on the application of the state-space reconstruction to various speech-processing systems is presented.

3.2 State-Space Analysis

3.2.1 The concept of state-space representation

State-space (also called phase-space) representation is a very powerful tool for analyzing complex dynamic systems based on the concept of *state*. A state of any dynamic system is the minimal set of variables (known as *state variables*) such that the knowledge of these variables at time $t = t_0$, along with the knowledge of the input for $t \geq t_0$, completely determines the behaviour of the system at any time $t \geq t_0$. Therefore, state-space is defined as the m-dimensional space whose coordinate axes constitute the state variables. Any state the system takes can be represented by a point (vector) in the state-space. The concept of state-space representation, by itself, is not new. It has existed for a long time in the field of classical dynamics as well as in other fields. However, the exploiting of this concept when only a single observation of the system is known was first presented by Packard [93] in 1980 and proved by Takens [94] a year later.

The representation of dynamic systems in the state-space form has many advantages:

1. It provides an insight into the behaviour of the system.
2. It is an effective tool to study nonlinear and time-varying systems.
3. It can be extended to handle systems with multiple inputs and outputs.

Variables (degrees of freedom) that contribute to the dynamics (changes in state) of a nonlinear system interact in a multiplicative rather than an additive manner. This is why, in general, the outcome of a nonlinear system is sometimes difficult to predict using time and frequency representations and usually displays chaotic behaviour. In order to understand the merit of state-space in representing such systems, we have chosen a deterministic system known to behave nonlinearly, the Lorenz oscillator. This system is one of the most striking examples of nonlinear chaotic behaviour. It was derived in 1963 by Edward Lorenz [95] from the simplified model of a convectional

(heat transfer) system in the atmosphere. The dynamics of the Lorenz model can be characterized with the following set of first-order differential equations:

$$\begin{aligned} \dot{x}(t) &= \sigma(y(t) - x(t)) \\ \dot{y}(t) &= \rho x(t) - y(t) - x(t)z(t) \\ \dot{z}(t) &= x(t)y(t) - \beta z(t), \end{aligned} \tag{3.2.1}$$

where $x(t)$, $y(t)$, and $z(t)$ correspond to the three state variables (two temperature measures and a velocity measure), the over-dot notation corresponds to the derivative (rate of change) of the variable in question, and σ, ρ, and β are constant parameters. As can be noticed from the above set of differential equations, influences of the state variables on the current state of the system are not independent and additive, but are instead mutually dependent and multiplicative. Changes in $x(t)$, for example, are dependent not only on the value of $x(t)$, but also on the values of $y(t)$ and $z(t)$. This interactive nature of state variables along which the system may change determines the complexity of nonlinear systems and is also the key to quantifying systems with unknown state variables. For the Lorenz model at hand, it suggests a nonlinear, three-dimensional, and deterministic system.

Figure 3.1(a) shows 30 seconds of the $x(t)$ variable of the Lorenz system in the time domain using the parameter set $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. The corresponding Fourier transform of this time signal is shown in Figure 3.1(b). Although the time-domain representation seems to have some sort of irregular periodic behaviour, the frequency-domain plot gives no insight at all into the nature of the system. We tend to assume random behaviour despite the deterministic equations. However, if we look at the state-space representation of this system in Figure 3.1(c), we will notice a very regular and predictable behaviour. The figure shows a chaotic attractor that keeps "drawing the same pattern" in the state-space, no matter what initial conditions were chosen for the differential equations. This type of attractor is commonly known as a *strange attractor.*

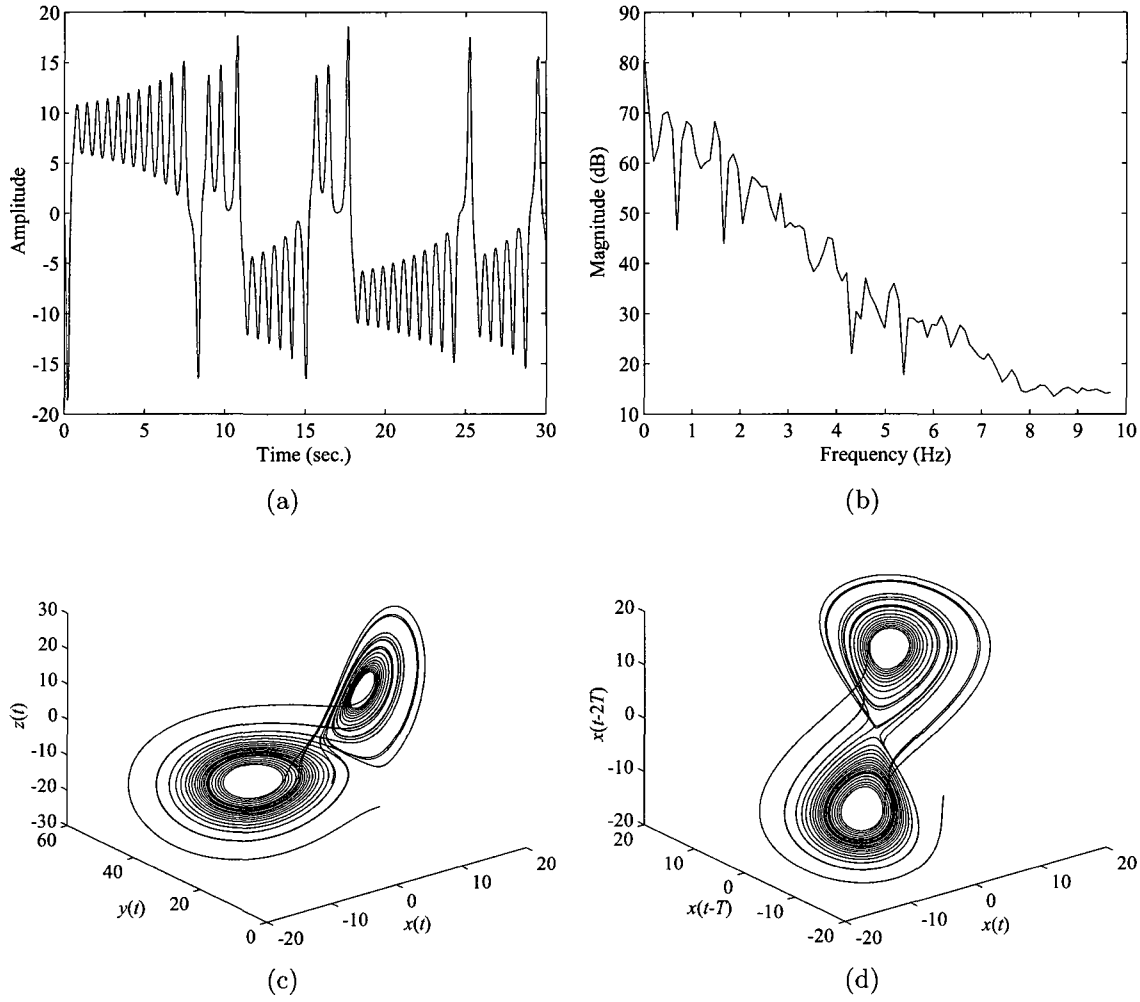


Figure 3.1: Lorenz attractor: (a) time-domain representation of the signal $x(t)$, (b) the corresponding Fourier transform of $x(t)$, (c) a three-dimensional plot of the system's state-space, and (d) the reconstructed state-space using the method of delay. (See Section 3.2.2).

From the above example, we can conclude that in certain situations the state-space emphasizes characteristics of the underlying system that can barely be seen in the original time or frequency spaces. Within the state-space, several features such as recurrence and parallelness between trajectories can be extracted and utilized in further processing. For example, small variations of angles between vectors indicate smooth changes of the state variables, while large variations reveal random behaviour. This is an important feature that can be used, for example, in classifying speech signals into voiced and unvoiced segments.

For experimental and naturally occurring systems, including speech, the state variables as well as the set of describing differential equations of the system are often unknown (or perhaps unmeasurable). Only one or more output signals are observed as a function of time. An important question arises: How can state variables be found from such unidimensional observations? The answer lies in the embedding theorem [94].

State-space reconstruction methods have been developed as a means to *reconstruct* the state-space and develop new predictive models. In these methods, the observed time signal is used to form a space that is topologically equivalent to the original state-space of the system. (See Figure 3.1(d) for the Lorenz model example.) This is possible because the interactive nature of the state variables of a nonlinear system (as described above) prescribes that the change of a single state variable will be influenced by the change of all other state variables. Therefore, access to one of these variables can be utilized to evaluate the dynamics of the entire system by unfolding the unidimensional signal into the appropriate number of dimensions to reveal the underlying dynamics. Further discussion about state-space reconstruction and the embedding theorem is presented in the following section.

3.2.2 State-space reconstruction and the embedding theorem

In many real applications, often only unidimensional measurements are available for the system under study. For example, in speech and audio systems, we usually have access to a single observable output from the system (e.g., the acoustic pressure wave measured by a microphone) that has been sampled to produce a time series. In many cases, this scalar time series contains the only information available from that system. An important challenge that must be met is the calculation of the system's real multi-dimensional state-space trajectory from such a unidimensional observation.

Takens' theorem [94] suggests how the unknown state-space of a dynamical system can be reconstructed from a time series of only one observed variable. This theorem is commonly known as the *delay embedding theorem.* It states: If a certain condition is met, it is possible to *reconstruct* a state-space trajectory from time-delayed replicas of the observed time series with one-to-one mapping between the reconstruction and the actual attractor of the underlying system. To further explain this theorem, let us denote the single observable N-point time series obtained from the system by $x(n)$, where $n = 0, 1, \ldots, N-1$. Starting from this unidimensional time series, we can build a new dynamical system of arbitrary dimension m (i.e., in the Euclidean space $\mathbb{R}^m$) that behaves similarly to the original system. The m-dimensional vectors of the new system can be formed by unfolding the observed samples and embedding them into the m-dimensional space using d-sample delayed replicas of $x(n)$. That is,

$$\mathbf{x}_m(n) = \Big[\; x(n),\;\; x(n+d),\;\; \cdots,\;\; x(n+(m-1)d) \;\Big], \tag{3.2.2}$$

where d is defined as the reconstruction or embedding delay, m is referred to as the embedding dimension, and $n = 0, 1, \ldots, N-1$. Here, we assume, for simplicity of formulation, that we have access to the extra samples $x(N), x(N+1), \cdots, x(N-1+(m-1)d)$. According to Takens' theorem, for almost any time series $x(n)$ and almost any embedding delay d, the attractor of the new constructed m-dimensional system

will be equivalent topologically to the attractor of the original system if

$$m \geq 2m_0 + 1, \tag{3.2.3}$$

where m_0 is the dimension of the original manifold. Later results by Sauer *et al.* [96], reduced the lower limit of (3.2.3) to

$$m \geq 2m_{box}, \tag{3.2.4}$$

where $m_{box} \leq m_0$ is called the *box-counting* or *fractal* dimension of the original manifold and is calculated as

$$m_{box} = \lim_{\epsilon \to 0} \frac{\log B(\epsilon)}{\log(1/\epsilon)}, \tag{3.2.5}$$

where $B(\epsilon)$ is the number of hyper-boxes of side length ϵ required to cover the original attractor in the Euclidean space $\mathbb{R}^{m_0}$. This lower limit for choosing the embedding dimension, m, assures the elimination of any self-crossing of the trajectories in the reconstructed state-space of chaotic systems [96].

By using the state-space reconstruction described above, the dynamics of a nonlinear system of multiple degrees of freedom (e.g., Lorenz system) are retrieved and features that provide important knowledge about the system's behaviour can be extracted. For example, the so-called invariant geometrical properties, such as the Lyapunov exponent (a quantity that characterizes the rate of separation of close trajectories) and correlation dimension (a measure of the dimensionality of the state-space), can be calculated from the reconstructed state-space and yet will still correspond to the actual values of the underlying system's attractor.

In practice, the state-space reconstruction is usually normalized using

$$\acute{\mathbf{x}}_m(n) = \frac{\mathbf{x}_m(n) - \overline{\mathbf{x}}_m}{\sigma_{\mathbf{x}}}, \tag{3.2.6}$$

where the mean vector, $\overline{\mathbf{x}}_m$, and the standard deviation scalar, $\sigma_{\mathbf{x}}$, of the reconstructed attractor are defined, respectively, as

$$\overline{\mathbf{x}}_m \triangleq \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{x}_m(n) \tag{3.2.7}$$

and

$$\sigma_{\mathbf{x}} \triangleq \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} \|\mathbf{x}_m(n) - \bar{\mathbf{x}}_m\|^2}. \tag{3.2.8}$$

Going back to the Lorenz attractor of Figure 3.1, we can see that the reconstructed state-space in Figure 3.1(d) (using $m = 3$ and $d = 15$ at a sampling frequency of 200 Hz) gives almost precisely the same portrait of the true state-space in Figure 3.1(c). Not only is the overall topological shape preserved, but also local details of the similarity of orbits are easily recognized in the two plots. The purpose of the Lorenz example is to show that the state-space of a system can be reconstructed even with access to only one of the many possible dimensions of change.

3.2.3 Choice of embedding parameters

The choice of proper values of the time delay, d, and the embedding dimension, m, need to be made carefully to achieve a reliable state-space reconstruction. According to Takens' theorem, as mentioned in the previous section, the embedding dimension must be greater than approximately twice the dimension of the original manifold [94]. On the other hand, there is no theory that places bounds on the choice of the embedding delay. However, the embedding delay must be chosen carefully for the quality of the reconstruction. Unfortunately, in most practical applications, the dimension of the original state-space structure is unknown. Therefore, some other means of determining an appropriate embedding dimension and time delay must be found. The literature has suggested many methods for determining the best values of these two parameters. We will briefly discuss some of these methods in this section.

Choosing the embedding dimension

The embedding dimension, m, determines how many coordinates will be used in reconstructing the state-space. In choosing m, the goal is not to determine precisely the dimension of the original manifold. In fact, the goal is to be sure that the embedding dimension is sufficient to allow the dynamics of the underlying system

to be disclosed with minimal distortion. Generally, the choice of the embedding dimension according to (3.2.3) or (3.2.4) assures that the final attractor does not lie on an ambiguous subset in the embedding space. Due to the validity of Takens' theorem, m can be arbitrarily chosen. However, if m is too large, the locality of the dynamics gets lost and the computational cost increases. Furthermore, if the signal is contaminated with noise, the higher m will be occupied by this extra noise instead of the meaningful dynamics of the system. On the other hand, if m is too small, the manifold will be folded in on itself. In this case, points close in the state-space may be close because of projection rather than because of the dynamics of the system (i.e., false neighbors). Therefore, it is necessary to find a systematic way of choosing an acceptable minimal embedding dimension that keeps the dynamics of the system unfolded.

One widely used method of estimating an acceptable minimal embedding dimension is the method of false nearest neighbors (FNN) developed by Kennel *et al.* [97]. The central idea of this method is based on the following concept: With the gradual increase of the embedding dimension, any property of the system that is dependent on the distance between two points in the state-space will stop changing when a sufficient embedding dimension is reached. This means that if the number of coordinates is too small, then points close in the reconstructed state-space may be close due to projection rather than to the dynamics of the system. In order to determine the sufficient embedding dimension using the FNN method, a search is first made for the nearest neighbor vector for each point of the reconstructed attractor, unfolded into an m-dimensional state-space. When the embedding dimension is increased to $m+1$, it is possible to discover the percentage of neighbors that were actually "false" neighbors and did not remain close because the m embedding dimension was too small. When the percentage of false neighbors drops to an acceptable value, it is possible to state that the attractor was completely unfolded and the minimum embedding dimension was reached.

Let us denote $\mathbf{x}_m(n)$ as a point in the reconstructed state-space of dimension m and embedding delay d and define $\mathbf{x}_m(p(n))$ as its nearest neighbor in terms of Euclidean distance. The squared distance between these two points is given by

$$\begin{aligned} D_m(n)^2 &= \left\|\mathbf{x}_m(n) - \mathbf{x}_m(p(n))\right\|^2 \\ &= \sum_{i=0}^{m-1} [x(n+id) - x(p(n)+id)]^2. \end{aligned} \tag{3.2.9}$$

Now, if we increase the embedding dimension by one and measure the squared distance between the same points in the new space, we will get

$$\begin{aligned} D_{m+1}(n)^2 &= \left\|\mathbf{x}_{m+1}(n) - \mathbf{x}_{m+1}(p(n))\right\|^2 \\ &= \sum_{i=0}^{m} [x(n+id) - x(p(n)+id)]^2. \end{aligned} \tag{3.2.10}$$

The difference in the squared distance between the two dimensions is then

$$D_{m+1}(n)^2 - D_m(n)^2 = [x(n+md) - x(p(n)+md)]^2. \tag{3.2.11}$$

This indicates how far the two neighbouring points have moved from each other due to dimension increase. Normalizing the square root of (3.2.11) with respect to the original distance given by (3.2.9) at the lower dimension results in a ratio of how far the displacement is compared to the original position of the two points. This ratio can be compared to a predetermined threshold, R_{th}, and used to identify the percentage of false nearest neighbors as follows [97]:

$$\text{FNN}(m)\ \% = \frac{100}{N} \sum_{n=0}^{N-1} H\left(\frac{\left|x(n+md) - x(p(n)+md)\right|}{D_m(n)} - R_{th} \right), \tag{3.2.12}$$

where $H(\cdot)$ is the Heaviside function defined as

$$H(y) = \begin{cases} 0, & y < 0 \\ 1, & y \geq 0. \end{cases} \tag{3.2.13}$$

In practical situations where the number of data points is not very large, another criterion can be applied to handle this issue of limited data size. In this case the

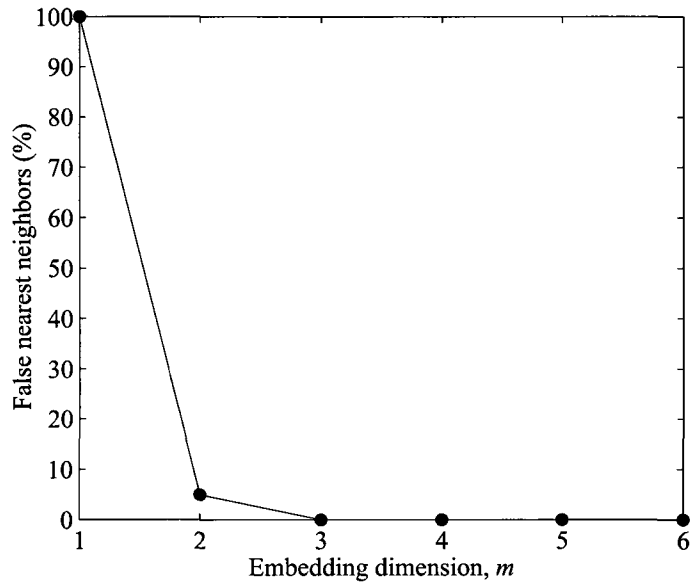


Figure 3.2: Percentage false nearest neighbors as a function of embedding dimension m for the Lorenz attractor (3.2.1).

following test can be used to identify FNN [97]:

$$\text{FNN}(m)\ \% = \frac{100}{N}\sum_{n=0}^{N-1} H\left(\frac{D_{m+1}(n)^2}{\sigma_{\mathbf{x}}^2} - A_{th}\right), \tag{3.2.14}$$

where $\sigma_{\mathbf{x}}^2$ is the squared value of (3.2.8) and A_{th} is another threshold. Other improvements to the FNN method especially when dealing with noise-corrupted data can be found in [98–100].

Equations (3.2.12) and (3.2.14) are used to check the neighbors in successively higher embedding dimensions until the percentage drops to a negligible number of FNN. The corresponding dimension to this value is then selected as the lowest embedding dimension that gives reconstruction without self-crossing.

An example of calculating the FNN is presented in Figure 3.2 for the Lorenz attractor (3.2.1). It is evident from the plot that an embedding dimension of three is quite sufficient to reconstruct the state-space of this attractor.

The FNN method can also be used to measure the relative contamination of noise

in the signal. However, the method is very sensitive to noise giving larger values of m as pointed out in [101]. In fact, the effect of noise is greater for larger values of d. This is a serious drawback of the method because in real applications we are led to choose a larger m than we really need.

Choosing the embedding delay

To perform state-space reconstruction using Takens' method, the delay parameter, d, which specifies the time lag between samples must be determined carefully. An optimum choice of the embedding delay will result in the state-space reconstruction being fully opened out, which will improve any subsequent analysis. If the d chosen is too small, the vectors $\mathbf{x}_m(n)$ and $\mathbf{x}_m(n+1)$ will be very similar, and consequently a correlated trajectory (probably stretched along the diagonal) is produced. On the other hand, if the d value is too large, the reconstructed trajectory becomes too disperse. In this case, much of the information is lost and the structure of the dynamics is difficult to ascertain.

Two criteria are commonly used to estimate the embedding delay based on the analysis of the autocorrelation function (ACF) or the average mutual information function (AMIF) [101]. The ACF provides a measure of the similarity between samples of the time signal. Typically, the value of d is set as the delay τ at which the ACF, $R(\tau)$, first passes through zero.

$$R(\tau) = \frac{\sum_{n=1}^{N} x(n)x(n+\tau)}{\sum_{n=1}^{N} x(n)^2}. \tag{3.2.15}$$

Choosing d as the delay corresponding to the first zero of $R(\tau)$ indicates that, on average over the observations, the coordinates $x(n+id)$ and $x(n+(i+1)d)$ are independent.

In the second method, as suggested by Fraser and Swinney [102], the embedding delay is chosen by looking at the AMIF, $I(\tau)$, of two points, $x(n)$ and $x(n+\tau)$,

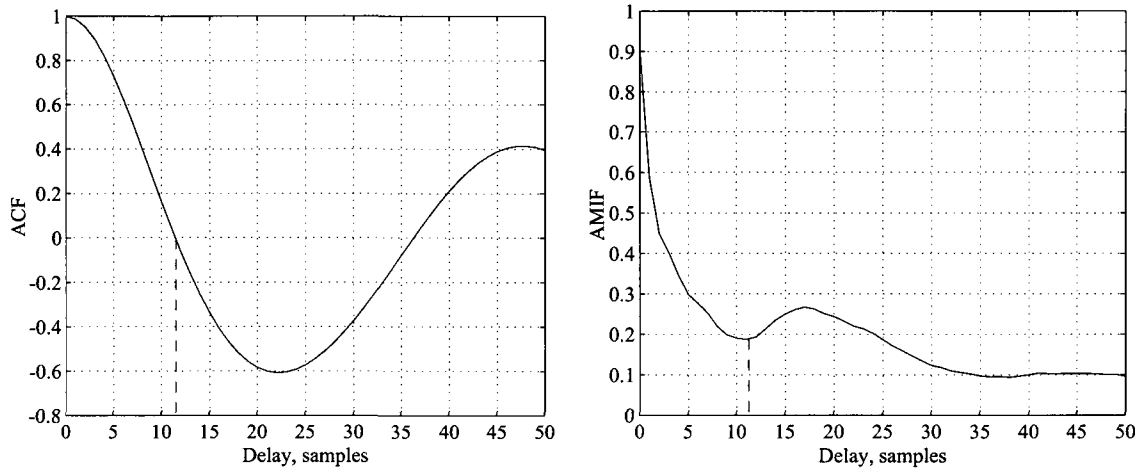
and taking the optimal delay, as the delay corresponds to the first minimum of this function.

$$I(\tau) = \sum_{n=1}^{N} P(x(n), x(n+\tau)) \log_2 \left[\frac{P(x(n), x(n+\tau))}{P(x(n))P(x(n+\tau))} \right]. \tag{3.2.16}$$

Calculating $I(\tau)$ from the time series is quite straightforward. To find $P(x(n))$, a histogram is formed by counting the number of times a given value of $x(n)$ appears divided by the total number of data. Likewise, $P(x(n+\tau))$ is computed from all values of $x(n+\tau)$. If the time series is long and stationary, then $P(x(n+\tau))$ will be almost the same as $P(x(n))$. The joint distribution $P(x(n), x(n+\tau))$ is obtained by plotting $x(n+\tau)$ against $x(n)$ and dividing this two-dimensional space into a grid of square bins. Then, a 3D histogram is formed by counting the number of times a square in this grid is occupied, divided by the total number of data. Finally, the AMIF at time delay τ is calculated by averaging these probabilities over all the bins. The process is repeated for the suggested range of delays. The first local minimum of $I(\tau)$ indicates minimal mutual information and hence maximal independence.

Figure 3.3 shows an example of estimating the embedding delay parameter for the Lorenz attractor using the ACF and the AMIF. Both methods seem to give approximately the same value of optimum time delay that can be used in the state-space reconstruction.

While the ACF is simpler and can be computed using fewer number of data compared to the AMIF, it in fact measures the *linear* dependence of coordinates, which gives a less accurate estimation of d when dealing with nonlinear data. AMIF, on the other hand, is more precise in the case of chaotic data. One may also be confronted with some problems when attempting to estimate embedding delay using either method. For example, the ACF may get approximately zero only after an extremely long time, or the AMIF may not have a clear first minimum. In these cases, a maximum limit for searching the best estimated delay should be set. Other methods for estimating embedding delay can be found in [92].



(a) (b)

Figure 3.3: Estimating embedding delay for the Lorenz system (3.2.1) by using (a) the first zero-crossing of the ACF and (b) the first minimum of the AMIF.

3.3 Nonlinear Analysis of Speech

Linear frequency-domain representation techniques are successfully applied to speech signals in almost all applications. Most of these techniques rely on a clear harmonic spectrum, which certainly contains useful information about speech. However, time or frequency representations of speech signals tell only half the truth. The important speech characteristics are defined by global non-stationary and nonlinear phenomena. These phenomena can never be described adequately by linear methods.

3.3.1 Speech as a nonlinear process

For decades, the classical linear model (i.e., one that obeys the superposition principle, such as the source-filter model) has been applied to speech processing with limited success. However, recent studies [103–107] have shown much evidence of significant nonlinear characteristics present in the production of speech signals that cannot be fully described by linear models. This means that the classical linear model is satisfactory only as an approximation of the overall nonlinear process of speech production. Accordingly, with nonlinear models we should obtain a more accurate

description of speech and, possibly, better performance of practical speech-processing applications. The main evidence of nonlinearity in the speech production system, as well as in the final speech signal, can be summarized in the following observations [103, 107].

Evidence of nonlinearity in the speech production mechanism

1. Interdependence and nonlinear coupling between the excitation signal and the vocal-tract filter during certain voiced sounds.
2. Laminar air flow through the vocal tract.
3. Nonlinear behaviour in the vibration of vocal folds to generate the glottal waves.
4. Turbulent air flow through a constriction in the vocal tract during unvoiced sounds.

Evidence of nonlinearity in the speech signal

1. Lack of ideal modelling of the speech signal by a Gaussian density function [108]. This suggests that a linear prediction scheme is not optimal in the mean squared sense [107]. Experiments [109] showed that the prediction error of a nonlinear predictor has smaller energy compared to linear predictor.
2. Large deviations in periodicity during voiced sound, which are indicated by the sudden appearance of subharmonics of the pitch frequency.

3.3.2 Applications of nonlinear speech processing

Nonlinear processing approaches for speech signals are potentially useful compared to linear approaches in many applications, including:

- Speech generation systems.
- Speech acquisition systems.
- Speech transmission.
- Human perception of speech.

While some problems are more tractable with nonlinear techniques than with linear techniques, limitations do exist. As reported in [110], some of these limitations for nonlinear speech processing are:

- The lack of a unifying theory between the different nonlinear processing tools (such as neural networks, homomorphic, polynomial, morphological, and ordered statistics filters).
- The greater computational burden compared to linear techniques.
- The difficulty of analyzing a nonlinear system because several useful tools are not valid, especially in frequency-domain analysis. The best results to date have been for nonlinear filter analysis.
- The lack at times of a closed formulation to derive the nonlinear models. Therefore, an iterative procedure must be used and local minima problems exist.

Recently, there has been a remarkable increase in publications that have studied the possibilities of applying nonlinear signal-processing techniques to speech. The main contributions were in the following fields.

Speech coding

The essential goal of speech coding is to represent digitized speech by the fewest possible number of bits while attempting to preserve the intelligibility and quality of speech needed for a certain application. Although it is possible to obtain sufficient accuracy in speech coding using conventional linear techniques (linear predictive coding (LPC), for example), nonlinear speech-processing techniques offer a better alternative. As discussed above, nonlinear models reflect the physical reality of speech more accurately than linear models. Therefore, it is reasonable to expect that speech coding based on nonlinear models will provide better performance in terms

of optimizing both quality and bit rate than coding based on linear models [109]. For example, it was reported in [111] that nonlinear prediction helped to achieve an improvement of about 2-3 dB in prediction gain over conventional linear prediction.

Recently, new nonlinear prediction-based speech coders using artificial neural networks (ANN) were proposed [112]. These techniques are shown to be more capable of taking account of nonlinearity and the non-Gaussian nature of speech.

Speech synthesis

Speech synthesis (or text-to-speech) is the artificial production of human speech by computers. In contemporary telephone services, speech synthesis systems play an important role in many man-machine interactive applications. Probably the most important qualities of any synthesized speech are intelligibility and naturalness. Intelligibility defines the degree of ease with which the output sound is understood, while naturalness describes how close this sound is to human speech.

It has been demonstrated that the implementation of speech synthesizers based on nonlinear modelling of speech can effectively produce a high-quality output signal in terms of naturalness and richness. For example, Banbrook *et al.* [113] developed a speech synthesis technique to produce high-quality sustained vowels by estimating the Lyapunov exponents from the reconstructed state space. The proposed synthesizer demonstrated some advantages over conventional techniques, such as the reduction of "buzziness" in the synthesized sound and the possibility for inclusion of simple coarticulation between phonemes.

Mann and McLaughlin [114] proposed another method to produce stable voiced speech using recurrent radial basis function neural networks that modelled the state-space dynamics. Despite the reported naturalness enhancement of the algorithm in an informal listening test compared to conventional techniques, the spectral plots of the synthesized speech showed a difficulty in modelling higher frequencies using this method.

Speech and speaker recognition

In automatic speech recognition (ASR), the target is to identify what is being said regardless of who is speaking. Speaker recognition, on the other hand, is the process of automatically identifying who is speaking regardless of what is being said.

Conventional speech and speaker recognition techniques are solely based on linear modelling theory (source-filter model) where the central processing space is the frequency domain. Typically, a vector of distinct features is first extracted from the frequency spectrum at a fixed rate and then used to search for the most likely word or speaker candidate. Examples of feature sets used by speech and speaker recognizers using linear method are the Mel-frequency cepstral coefficients (MFCC) and the linear prediction cepstral coefficients (LPCC). Recognizers based on nonlinear processing techniques, on the other hand, extract features from the state space that is reconstructed directly from the time-domain signal. Since the dynamics of the speech signal are more represented in the nonlinear model, this implies that features extracted from the state space can potentially contain additional and different information than spectral representation.

A recent study [115] showed that human listeners can still identify speakers even by listening to the residual signal obtained from a high-dimensional linear prediction analysis. On the other hand, this was a difficult task when the residual signal of a nonlinear analysis was used. This reveals that a remaining useful amount of information was ignored by the linear model, which was not able to cope with nonlinearities present in speech. Other papers concluded that it is possible to improve the identification rates with a combination of linear and nonlinear models. For example, simulation results conducted by Lindgren *et al.* [116, 117] and Petry and Barone [118] showed that the combination of MFCC or LPCC with state-space features (such as Lyapunov exponents and correlation dimension) leads to an increase in recognition accuracy over the case when only one method is used alone.

3.3.3 Speech analysis using reconstructed state space

Recently, several techniques used in the analysis of dynamic nonlinear systems have been applied to speech signals in order to investigate some of the short-term nonlinear characteristics of speech. State-space reconstruction is among the most widely used techniques in this domain [113, 116–127]. When the speech signal is modelled as the output of a nonlinear system, the state space can be reconstructed using the method of delay. As discussed in Section 3.2.2, this can be obtained by combining several delayed versions of the observed signal into a trajectory in a multi-dimensional space. Since speech signals are non-stationary, the embedding procedure is applied to short, consecutive segments called frames. Using state-space reconstruction enables us to identify useful features such as recurrence and parallelness of trajectories that can be utilized later for further processing. For example, small variations of angles between vectors indicate smooth changes of the state variables while large variations reveal random behaviour. This is an important feature that can be used, for example, in classifying speech signal into voiced and unvoiced segments.

It is generally acknowledged that voiced speech can be sufficiently embedded into a low-dimensional state space, whereas unvoiced speech has a noise-like nature and consequently needs a higher dimension to embed [106]. In practice, the short-term nature of the segmented speech makes determination of the true embedding dimension unnecessary [124]. For instance, a dimension of 3 or 4 is usually sufficient to reconstruct well-structured trajectories of voiced speech. The number of dimensions can be further increased, but beyond 4 or 5 no noticeable improvement can be observed for most practical applications.

On the other hand, the careful choice of embedding delay is important to improve the analysis by opening up or unfolding the attractor. The attractor should contain no areas where trajectories intersect in state space. The choice of an optimal delay parameter, d, depends on the sampling rate and signal properties. Embedding delay should be large enough for a reconstructed trajectory to be maximally "open" in state

space on average. On the other hand, it is desirable to keep d relatively small for better time resolution. In single-speaker applications, it is possible to use a constant value of d over all speech frames. A constant value of $d = 12$ at a sampling rate of 16 kHz, for example, was found to provide a good discrimination between structured (voiced) and unstructured (unvoiced) speech [124]. However, in multi-speaker situations where co-channel speech is present, it is preferable to recalculate d at each frame.

Time-delay embedding is the method of choice for many nonlinear speech-processing applications. It is possible, however, to use other embedding techniques, as long as they preserve topological properties of the original state space of a system. One particular alternative embedding technique is the singular value decomposition (SVD) introduced in [128]. This technique has some advantages over time-delay embedding due to its smoothing capabilities, leading to improved results on some types of signals (e.g., voiced fricatives and noisy speech). However, in most cases smoothing can also be achieved by simply performing moderate low-pass filtering of the signal before embedding it. Overall, SVD embedding can be a useful alternative to time-delay embedding, but its computational cost makes it less practical for real-time implementation.

Figure 3.4 shows examples of applying time-delay embedding to voiced and unvoiced speech. Two segments (one voiced and one unvoiced) taken from the speech waveform in 3.4(a) were embedded in the state space using the method of delay. In Figure 3.4(c) for the voiced-speech segment, the trajectory approximately repeats itself after every pitch period such that at each point in the state space we see a synchronized bundle of similar yet distinct trajectories due to the periodic nature of voiced speech. Figure 3.4(b) for the unvoiced-speech segment, on the other hand, shows the unstructured behaviour of unvoiced speech. As can be seen, vectors in the attractor are scattered randomly in the state space due to the high dimensionality of the unvoiced speech.

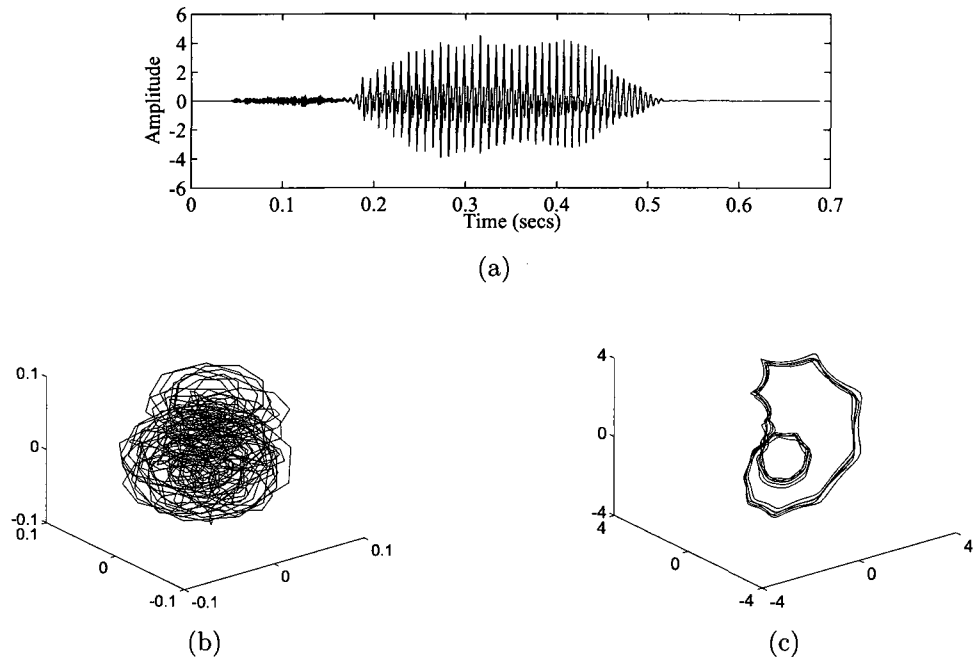


Figure 3.4: State-space embedding of speech: (a) speech waveform of a male speaker uttering the word "she," (b) embedded unvoiced 50 ms segment at 0.1 s, and (c) embedded voiced 50 ms segment at 0.3 s.

3.4 Simulation Results

To show and prove the existence of nonlinear characteristics in speech production, we conducted a simple experiment. In this experiment, several short segments of increasing loudness and sustained vowels (voiced speech) uttered by different male and female speakers were recorded. During the recording time, the speaker increased the level (volume) of the voice without moving neither the tongue nor the jaw, in order to avoid effects due to variations of the vocal tract. In addition, each speaker was asked not to change the tone (pitch) of the produced sound while raising his or her voice. This was to ensure that the level change of the output sound was caused only by increasing the air pressure in the lungs rather than by modifying the vibration pattern of the vocal folds. The recorded segments were visually inspected to make sure that the fundamental frequency of the voiced sound did not change with time.

Finally, the speech segments were analyzed and the changes in the power spectrum along each segment were investigated. Figure 3.5(a) shows a speech segment of the sustained vowel /aa/ uttered by a male speaker for one second. The pitch contour of this segment is plotted in Figure 3.5(b). It is quite evident that the calculated pitch value did not change much with time despite the change in the sound loudness.

To measure nonlinear characteristics in the speech samples, the energy ratio of the high to low frequency bands above and below 1 kHz was used. This was denoted as HILO ratio and was plotted against time, as shown in Figure 3.5(c), for the speech sample of Figure 3.5(a). Noticeable change in the HILO energy ratio along time (see Figure 3.5(c)) may reveal speech nonlinearity, as it indicates nonuniform gain change between low and high frequency bands. Since the vocal tract transfer function is assumed to be approximately fixed (time-invariant) over the testing duration, the only suggested interpretation of this gain change over time is nonlinearity in the sound production mechanism. This includes vocal folds and vocal tract interaction.

Figure 3.6 adds more insight into the picture by depicting two shorter frames at the beginning (lower-level sound) and at the end (higher-level sound) of the test segment of Figure 3.5(a). The power spectral densities of these two frames are presented in Figure 3.6(c) and Figure 3.6(d), respectively. It is clear that while the fundamental frequency did not change with time, frequencies in the higher band (above 1 kHz) have more gain than frequencies in the lower band. This is reflected in the time-domain signal as sharper transitions and fluctuations.

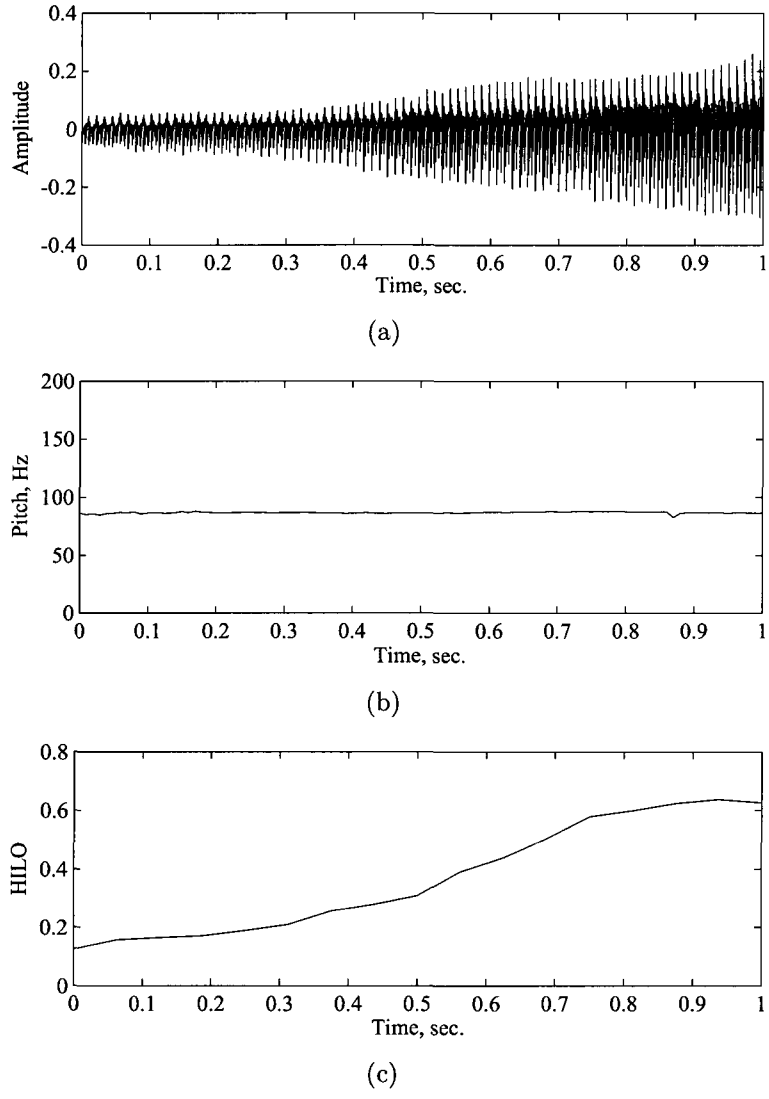


Figure 3.5: Example of nonlinearity in speech production: (a) an increasing voiced-speech segment of the vowel /aa/ spoken by a male speaker, (b) the measured pitch contour of the speech segment (approximately constant), and (c) variation of the HILO energy ratio with time.

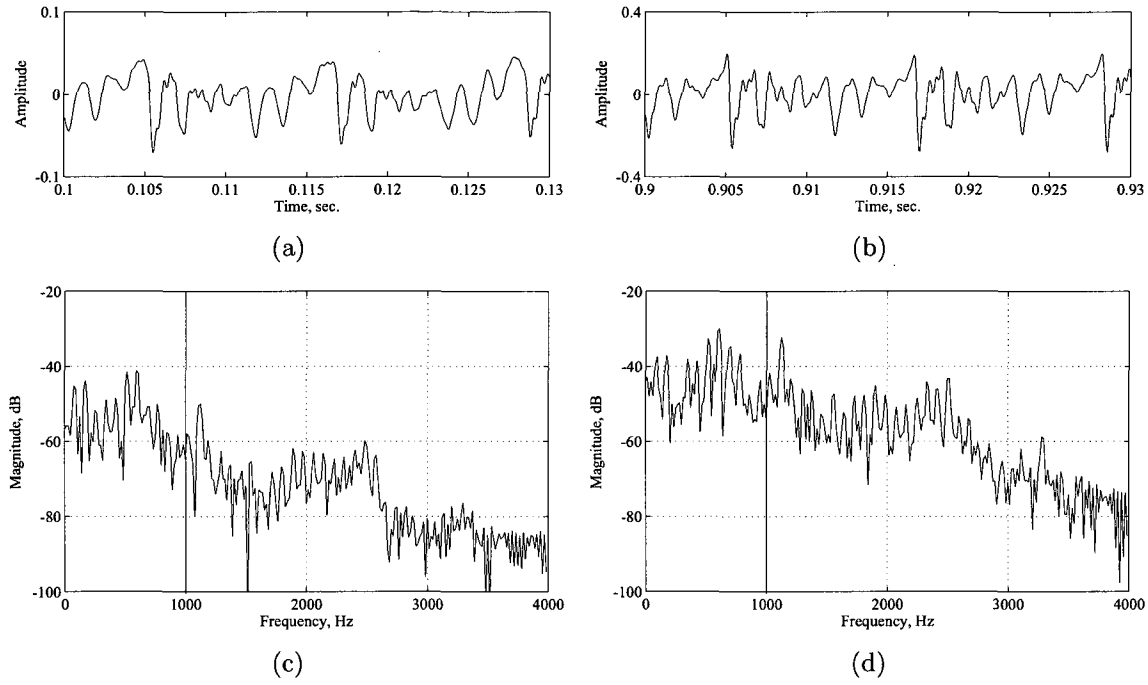


Figure 3.6: Snap shots of two different frames of the speech segment in Figure 3.5(a) along with their corresponding magnitude spectra: (a) starting lower-level frame, (b) ending higher-level frame, (c) magnitude spectrum of the starting frame, and (d) magnitude spectrum of the ending frame.

3.5 Summary

This chapter reviewed the concept of state-space reconstruction and the embedding theorem in modelling the nonlinear dynamics of speech. To reconstruct the state-space, the one-dimensional time-domain speech signal was embedded into an appropriate higher dimensional space. This reconstructed state-space had approximately the same dynamical properties as the original speech-generating system and was thus an effective model for speech synthesis. To improve the quality of the reconstructed state-space, some methods for determining the best values for the embedding dimension and time delay were described. Evidences of nonlinearity in both the speech production mechanism and the speech signal were mentioned and demonstrated using a simple experiment. Following this, a new algorithm for voicing-state classification of co-channel speech based on nonlinear state-space reconstruction is presented in the next chapter.

CHAPTER 4

Voicing-State Classification of Co-Channel Speech

4.1 Introduction

Voicing-state classification of single-speaker speech is a signal-processing method by which the speech waveform is segmented into voiced sound, unvoiced sound, or silence. For the co-channel case, voicing-state classification is extended to other states due to speech overlap. This process is required in a CCSS system as a means of selecting an appropriate separation-processing technique [37].

The use of a voicing-state classifier for co-channel speech can be of great help in enhancing the target speech signal in many applications. Voicing-state classifiers can also provide an important tool in any objective quality measurements of speech signals corrupted by other quasi-periodic signals such as music and speech.

In this chapter, a new approach to voicing-state classification of co-channel speech based on nonlinear state-space reconstruction is presented. The method's performance under varying levels of target-to-interference ratio (TIR) and under varying levels of signal-to-noise ratio (SNR) is investigated and compared with other existing techniques. See also [2, 3].

The organization of this chapter is as follows. In the next section we review some known methods for classifying single-speaker and co-channel speech signals into

different voicing states. The new method for voicing-state classification of co-channel speech is described in Section 4.3. In Section 4.4, simulation results for testing the performance of the new algorithm and comparing it with some existing techniques are presented.

4.2 Voicing-State Classification

4.2.1 Single-speaker classifiers

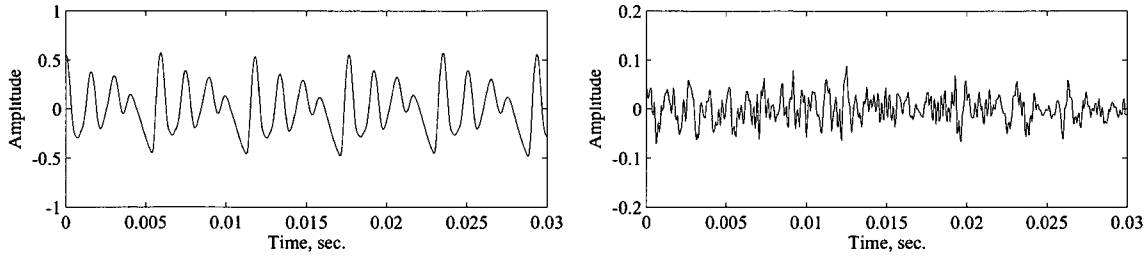
There are several ways to classify different events in speech produced by a single speaker. According to the mode of excitation, a speech segment can, in general, be classified into one of the following three states [108, 129]:

Voiced sound (V), in which the vocal folds vibrate with a slowly varying fundamental frequency. Therefore, the resulting waveform has a quasi-periodic nature and is characterized by relatively high energy. The fundamental frequency (pitch) of the produced voice is determined by the rate at which the vocal folds vibrate. Typically, women and children tend to have a higher pitch, while men tend to have a lower pitch. Vowel sounds such as /a/ in "father," /i/ in "eve," /e/ in "hate," and /u/ in "boot" fall in the voiced sound category. An example of the waveform of a voiced speech segment is shown in Figure 4.1(a).

Unvoiced sound (U), in which the vocal folds do not vibrate and the sound is generated by air turbulence in the vocal tract. As shown in Figure 4.1(b), the resulting speech waveform is random in nature and has relatively low energy. Examples of unvoiced sounds are the phonemes /s/ as in "see," /S/ as in "she," and /f/ as in "for."

Silence (S), where no sound is produced.

In fact, the classification of speech waveform into well-defined regions of voiced, unvoiced, or silence is neither exact nor easy. This is mainly due to state overlapping



(a) (b)

Figure 4.1: Two 30 ms speech segments of (a) voiced speech and (b) unvoiced speech.

during voicing transition as well as mixed excitation modes such as in voiced fricative sounds like /v/ in "vote," /TH/ in "then," and /z/ in "zoo." Other modes of excitation such as plosive sounds (produced by abrupt closing and releasing of the air flow) also affect the classification process.

In the last three decades, voicing-state classification of the speech of a single speaker has been addressed exhaustively by many researchers using different approaches [130–137]. In almost all methods, the classification problem was formulated as a three-state pattern recognition problem. The speech signal was segmented into small fixed-length frames, and the classification was performed by extracting uncorrelated feature vectors from these frames. The features were then used in a decision criterion to determine the voicing state of the current frame. Features were derived either in the time domain or the frequency domain. In the next few pages, we will briefly present some features that are commonly used in the single-speaker voicing-state classification process.

Common time-domain features

- **Short-term energy (STE)** [130, 132]

  Short-term energy (STE) is defined as

$$\text{STE [dB]} = 10 \log_{10} \left(\varepsilon + \frac{1}{N} \sum_{n=0}^{N-1} x^2(n) \right), \tag{4.2.1}$$

where $x(n); n = 0, 1, \ldots, N-1$ are the speech samples and ε is a small constant to avoid calculating log(0). Typically, voiced speech has higher energy than unvoiced speech or silence.

- **Zero-crossing rate (ZCR)** [130, 132]

  Zero-crossing rate (ZCR) is the number of sign-changes per second in the input signal. This feature reflects the dominant spectral concentration in the speech waveform. A low ZCR indicates the dominance of low-frequency energy as in the case of voiced speech, while a high ZCR indicates the dominance of high-frequency energy as in the case of unvoiced speech. For a speech frame of N samples, ZCR can be calculated as follows.

$$\text{ZCR} = \frac{F_s}{N} \sum_{n=1}^{N} \mathbb{I}\{x(n)x(n-1) < 0\}, \tag{4.2.2}$$

  where F_s is the sampling frequency and the indicator function $\mathbb{I}\{\cdot\}$ is 1 if its argument is true and 0 otherwise. A major drawback of the ZCR is its high susceptibility to DC offset.

- **First normalized autocorrelation coefficient** [130, 132]

  This parameter measures the amount of correlation between adjacent speech samples.

$$\phi_1 = \frac{\sum_{n=1}^{N} x(n)x(n-1)}{\sqrt{\left(\sum_{n=1}^{N} x^2(n)\right)\left(\sum_{n=0}^{N-1} x^2(n)\right)}}. \tag{4.2.3}$$

  Due to the concentration of energy at low-frequency, voiced speech tends to have highly correlated adjacent samples and hence a correlation coefficient close to 1. Unvoiced speech, on the other hand, tends to have a correlation coefficient close to zero due to its randomness nature.

- **First coefficient of the LPC predictor** [130–132]

  In the linear predictive coding (LPC) analysis of speech, the predicted value

$\tilde{x}(n)$ of the nth speech sample $x(n)$ is given by

$$\tilde{x}(n) = \sum_{p=1}^{P} \alpha_p x(n-p), \tag{4.2.4}$$

where a_p's are called the predictor coefficients and are determined by minimizing the mean-squared prediction error between $x(n)$ and $\tilde{x}(n)$.

$$\overline{e^2(n)} = \frac{1}{N} \sum_{n=0}^{N-1} (x(n) - \tilde{x}(n))^2. \tag{4.2.5}$$

The first predictor coefficient, α_1, can vary from a value of about -5 for voiced speech to a value of about 1 for unvoiced speech [130].

- **High-to-low frequency energy ratio (HILO)** [132]

  This is the ratio of the signal energy in the high-frequency band above f_H Hz to the energy in the low-frequency band below f_L Hz. To calculate this ratio, the speech signal is first filtered with a low-pass filter of a cut-off frequency of f_L Hz to produce the output signal $x_L(n)$. Then, the speech signal is filtered with a high-pass filter of a cut-off frequency of f_H Hz to produce the output signal $x_H(n)$. The value of HILO is then calculated for a frame of N samples as

$$\text{HILO [dB]} = 10 \log_{10} \left(\varepsilon + \frac{\sum_{n=0}^{N-1} x_H^2(n)}{\sum_{n=0}^{N-1} x_L^2(n)} \right), \tag{4.2.6}$$

  where ε is a small constant to avoid calculating log(0). Since voiced speech is more concentrated in the low-frequency band (below 2 kHz), it is expected to give lower HILO value compared to unvoiced speech.

Common frequency-domain features

- **Spectral flatness measure (SFM)** [138]

  This measure is used to determine the noise-like or tone-like nature of the signal.

The SFM is defined as the ratio of the geometric mean to the arithmetic mean of the signal's magnitude spectrum, i.e.,

$$\text{SFM [dB]} = 10 \log_{10} \left(\frac{\sqrt[K]{\prod_{k=0}^{K-1} |X(k)|}}{\frac{1}{K} \sum_{k=0}^{K-1} |X(k)|} \right), \tag{4.2.7}$$

where $|X(k)|$ is the magnitude spectrum (discrete Fourier transform (DFT)) of $x(n)$ at bin number k and K is the total number of bins. A high SFM indicates that the speech has a relatively flat spectrum in all bands similar to white noise. A low SFM indicates that the spectral power of speech is concentrated in a relatively small number of bands. Therefore, unvoiced speech has a high SFM close to 0 dB, while voiced speech has a low SFM close to -60 dB.

- **Harmonic-to-noise energy ratio** (HNR) [134]

  This ratio measures the strength of harmonic component in a speech signal. HNR is defined as

$$\text{HNR} = \frac{\sum_{k=0}^{K-1} |X_H(k)|^2}{\sum_{k=0}^{K-1} |X_N(k)|^2}, \tag{4.2.8}$$

  where $X_H(k)$ and $X_N(k)$ are the harmonic and noise components resulting from the harmonic-plus-noise decomposition of $X(k)$.

- **Mel-frequency cepstral coefficients (MFCC)** [135]

  The cepstral representation (spectrum of a spectrum) of the speech signal provides a good representation of the local spectral properties of the signal for the given frame analysis. For each speech frame, the MFCCs is calculated using the following steps:

  1. Compute the DFT of the windowed signal.
  2. Map the powers of the spectrum obtained above onto the mel scale.

3. Take the logs of the powers at each of the mel frequencies.

4. Compute the discrete cosine transform (DCT).

Some early approaches [139] tried to use frame-based periodicity (pitch) detection as the only feature for single-speaker voicing-state classification. However, such algorithms have not been highly accurate. This is mainly because pitch detection algorithms require long speech frames (at least twice the pitch period), while voiced speech is only approximately periodic over short time intervals [130]. Thus, separating the voicing-classification task from the pitch detection algorithm allows performing voicing classification on shorter speech segments.

4.2.2 Two-speaker classifiers

As shown in the previous section, natural speech signals can be classified roughly into one of three possible states of excitation: voiced (V), unvoiced (U), or silence (S). When two speech signals interfere, there are nine possible combinations of voicing states (classes) that can occur. These combinations are illustrated in the matrix of Table 4.1 for the MIT-CBG database [140] with several strategies of separation processing [11]. The MIT-CBG database is a "read speech" database that contains recorded speech of three male speakers each is uttering 210 sentences (630 in total). From the table it can be noticed that:

- The only portions of co-channel speech that require processing for separation (denoted P) are those when one speaker is in the voiced state and the other speaker is in the unvoiced state (i.e., V/U or U/V), and when both speakers are in the voiced state (i.e., V/V).

- No processing is needed (denoted NPN) if at least one of the speakers is silent.

- When both speech components are unvoiced (U/U), this segment cannot be processed well with current techniques (denoted X) due to the high ambiguity.

These separation strategies suggest that the possible voicing states to be classified for the purpose of co-channel speech separation can be reduced to the following five classes:

1. Silence: when both speakers are silent.
2. Unvoiced/Unvoiced (U/U): when either both speakers are producing unvoiced sounds or one speaker is silent, while the other speaker is in the unvoiced state.
3. Voiced/Unvoiced (V/U): when the desired speaker is producing a voiced sound, while the interfering speaker is either producing an unvoiced sound or is silent.
4. Unvoiced/Voiced (U/V): when the desired speaker is either producing an unvoiced sound or is silent, while the interfering speaker is producing a voiced sound.
5. Voiced/Voiced (V/V): when both speakers are producing voiced sounds.

These classes could be further reduced to three classes. If we consider the silence state to be a subset of the unvoiced state, classes 1 and 2 above are assumed to be one class. This is justifiable since in all these classes either no processing is needed or no processing can be done with standard techniques. In addition, if there is no need for the classification algorithm to discriminate between speakers in the V/U and U/V classes, classes 3 and 4 would be combined as one class. This is accepted when the speaker assignment algorithm is done at a different stage.

The upper right-hand corner of each box in Table 4.1 gives the observed percentage coincidence for each combination of states for some experiments conducted with the MIT-CBG database [29]. These percentages, of course, vary widely depending on the nature of conversation and the interaction between the speakers. However, the table gives an important indication that in the presence of crosstalk, a good portion of the overlapped speech (about half of the time for this speech corpus) falls into the V/V category.

Speaker 2 \ Speaker 1	Silence	Unvoiced	Voiced
Silence	3% NPN	4% NPN	10% NPN
Unvoiced	4% NPN	6% X	14% P
Voiced	10% NPN	14% P	35% P

Table 4.1: Processing strategies available for the combinations of vocal excitation for the MIT-CBG database [11].

Most single-speaker classification features mentioned in the previous section are not sufficient candidates alone to identify all voicing-state classes of co-channel speech. This is because, in co-channel speech, the classifier is required not only to discriminate between voiced and unvoiced speech but also to distinguish mixed excitation of the two speakers as in the V/U and V/V classes.

Previous work on voicing-state classification of co-channel speech has shown some success using either *a priori* information about the individual speakers [55] or training data sets [141]. However, *a priori* information is not always available in many practical situations. Also, methods that use training data sets are speaker and environment dependent. Each time the recording conditions or the background noise level changes, a new set of training data is required. Furthermore, in most proposed co-channel speaker separation systems, researchers process only those frames in which the desired speaker is voiced.

In [142], Zissman and Weinstein borrowed a speaker identification technique to label co-channel speech intervals as target-only, interferer-only, or two-speaker (target plus interferer). Two classifiers were applied in this technique: a vector-quantizing classifier and a modified Gaussian classifier. A feature vector containing 20 MFCCs

was used in both training and detection phases at different TIRs to identify which of the three possible classes were present. Experiment results using both classifiers showed relatively high performance with supervised and unsupervised learning.[1]

The HES speech separation approach proposed by Morgan *et al.* [11] assumed that at any co-channel speech interval, one speaker is much stronger than the other. Hence, a traditional single-speaker method for voicing-state classification was used to classify the current frame as either voiced or silence/unvoiced. Classification was performed by calculating five-feature vectors for each frame and analyzing them using a multivariate Gaussian classifier. The five features used were: the HILO ratio, the ZCR, the first coefficient of the LPC predictor, the log of the prediction error, and the ML magnitude passed through a comb filter at the pitch period of the stronger speaker. The task of the classifier was to use these features to generate a binary voicing decision corresponding to the voicing state of the current frame. A major drawback of this system was that it identified only the presence or absence of voiced speech and did not identify the voicing state of each speaker. The performance also degraded when both speakers had comparable power intensities.

Benincasa [141] trained a Bayesian classifier using supervised learning to automatically detect the five voicing classes mentioned earlier for co-channel speech. The feature vector used consisted of 37 features, including the STE, the fundamental frequency, the first normalized autocorrelation coefficient, the ZCR, the HILO ratio, 16 MFCCs, and 16 modified covariance coefficients.

In [14, 39, 40], attempts to locate only the usable speech segments (single-speaker voiced frames) using the adjacent pitch period comparison (APPC) and the spectral autocorrelation peak-to-valley ratio (SAPVR) measures were presented. No great attention was given to the other voicing-state classes of co-channel speech in these methods.

[1]Supervised learning refers to a machine-learning technique for learning a function from training data where the training data consist of pairs of input objects (speech frames in our case) and desired outputs or labels (voicing states). Unsupervised learning, on the other hand, refers to the machine-learning technique where the training data consist only of input objects.

4.3 New Voicing-State Classification Method

As shown in the previous section, most of the conventional techniques used to classify the voicing state of single-speaker and co-channel speech rely on the pattern recognition approach and treat the speech system as a linear system [37, 130, 132]. However, nonlinear approaches are known to give better access to the full dynamics of the speech production system than linear techniques. The application of nonlinear dynamical methods to speech characterization and analysis has produced numerous new and promising approaches over the last two decades. For example, the results in [124] and [121] have shown superior performance in solving the general problems of pitch determination and speech enhancement by using nonlinear methods compared to the linear techniques.

In this section, a new approach to voicing-state classification of single-speaker and co-channel speech based on nonlinear state-space reconstruction is presented. The method's performance under varying levels of TIR and under varying levels of SNR is investigated and compared with other existing techniques. The following three voicing-state classes are considered for co-channel speech:

1. Unvoiced/Unvoiced (U/U): where both speakers are either in the unvoiced state or the silence state.

2. Voiced/Unvoiced (V/U): where only one speaker is in the voiced state.

3. Voiced/Voiced (V/V): where both speakers are in the voiced state.

The silence state is assumed to be a subset of the unvoiced class. Also, no need to differentiate between speakers in the V/U class is assumed.

4.3.1 State-Space Reconstruction

State-space (also called phase-space) reconstruction is the first step in our nonlinear time series analysis. It basically views a single-dimensional data series, $x(n)$, in an m-dimensional Euclidean space $\mathbb{R}^m$. Using this method, the trajectories connecting data points (vectors) in the state space are expected to form an attractor that preserves the topological properties of the original unknown attractor. A common way to reconstruct the state space is the method of delays introduced by Takens [94]. In this method, m-dimensional vectors, $\mathbf{x}_n$, in the state space are formed from the time-delayed samples of the original signal, $x(n)$, as follows.

$$\mathbf{x}_m(n) = \Big[\; x(n),\;\; x(n+d),\;\; \ldots,\;\; x(n+(m-1)d) \;\Big], \tag{4.3.1}$$

where d is the embedding delay and m is the embedding dimension (number of coordinates). Delay reconstruction requires a proper choice of parameters d and m. The value of d can be calculated as the time (in samples) of the first zero crossing of the autocorrelation function. This allows opening up the attractor in the state-space reconstruction. The embedding dimension, m, has to be large enough (e.g., $m = 5$ in the present application) to avoid false neighbour trajectories. For more details on methods of estimating these two parameters, please refer to Section (3.2.3).

4.3.2 Method description

Voiced speech is known to have a quasi-periodic nature. Thus, it can be fully represented using a low-dimensional state space. Furthermore, the trajectories of voiced speech with temporal distances that are multiples of the pitch period tend to be close and parallel to each other. Co-channel and unvoiced speech, on the other hand, do not have this quasi-periodic nature and therefore require a higher embedding dimension. Figure 4.2 shows three speech frames with different voicing states V/U, V/V and U/U, and their corresponding state-space reconstructions, respectively, for $m = 3$.

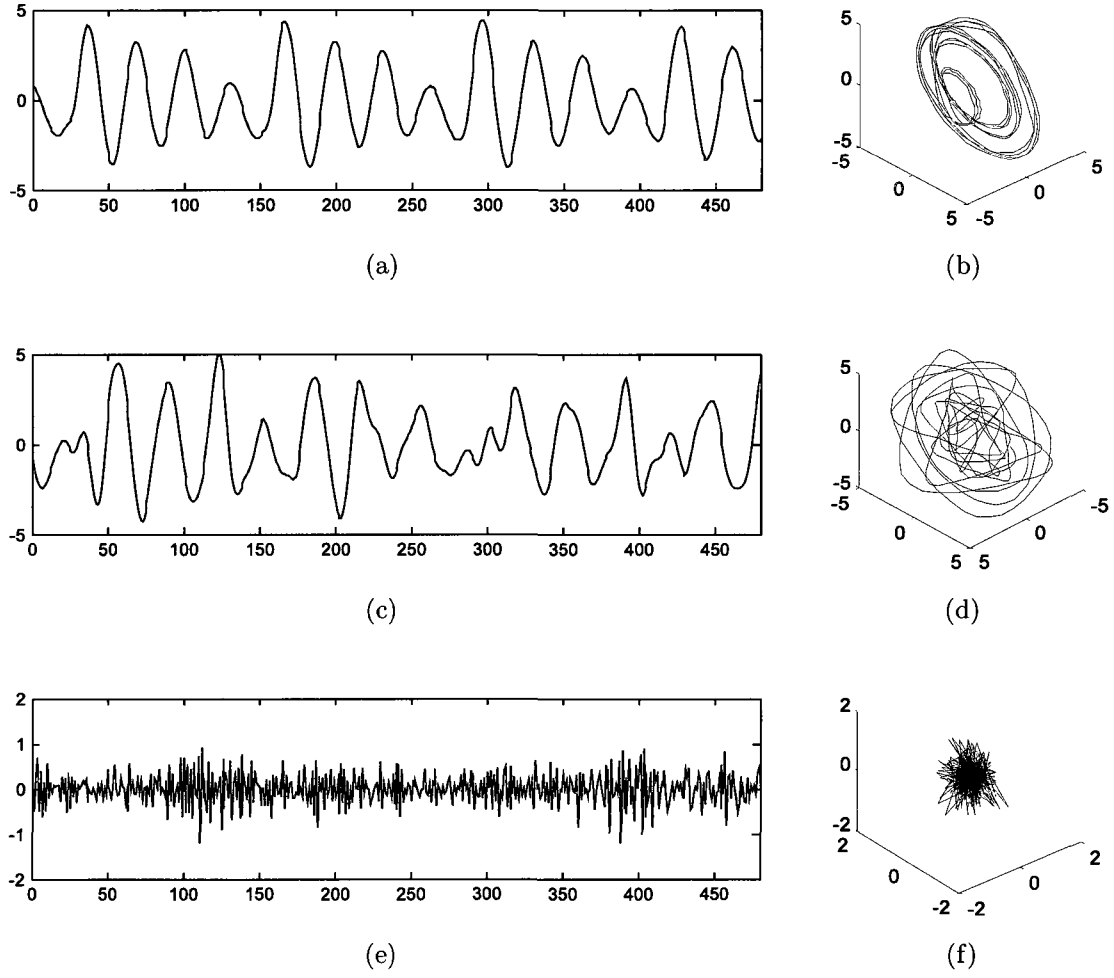


Figure 4.2: Three frames of (a) single-voiced, (b) double-voiced and (c) unvoiced speech signals and their corresponding state-space reconstructions (d), (e) and (f), respectively, for $m = 3$.

The trajectory parallel measure (TPM) [123, 143] of the state-space attractor can be used to measure the degree of similarity in the neighbour trajectories and hence detect the presence of voiced speech. If $\mathbf{t}(n)$ is the query trajectory at time index n and $\mathbf{t}(n_i)$ is the i^{th} nearest trajectory to $\mathbf{t}(n)$, then

$$\begin{aligned} \mathbf{t}(n) &= \mathbf{x}_m(n+1) - \mathbf{x}_m(n) \\ \mathbf{t}(n_i) &= \mathbf{x}_m(n_i+1) - \mathbf{x}_m(n_i), \end{aligned} \tag{4.3.2}$$

where $\mathbf{x}_m(n)$ is the query point and $\mathbf{x}_m(n_i)$ is the nearest neighbour point on the i^{th} trajectory. The angle between $\mathbf{t}(n)$ and $\mathbf{t}(n_i)$ is given by

$$\psi(n, n_i) = \frac{\mathbf{t}(n) \cdot \mathbf{t}(n_i)}{|\mathbf{t}(n)|\ |\mathbf{t}(n_i)|}. \tag{4.3.3}$$

The TPM of the whole attractor can then be calculated as [143]

$$\text{TPM} = \frac{1}{N} \sum_{n=0}^{N-1} H\left(\alpha - \frac{1}{L_{NN}} \sum_{i=1}^{L_{NN}} \psi(n, n_i) \right), \tag{4.3.4}$$

where N is the total number of samples in the frame, α is a constant threshold, L_{NN} is the total number of nearest neighbor trajectories, and $H(\cdot)$ is the Heaviside function, which is defined as

$$H(y) = \begin{cases} 0, & y < 0 \\ 1, & y \geq 0. \end{cases} \tag{4.3.5}$$

For example, a TPM value of 0.5 indicates that half of the trajectories in the attractor are nearly parallel to each other.

The proposed algorithm for classifying co-channel speech frames into one of the three voicing-state classes is presented in the Figure 4.3 flowchart. In this algorithm, the speech signal is first segmented into overlapping frames with a duration of at least twice the maximum human pitch period (about 12.5 ms). The voicing-state decision is obtained by a sequence of two-way decisions as follows.

1. The current speech frame is checked for the presence of voiced speech and classified as either voiced (V/U or V/V) or unvoiced (U/U) using the following two measures:

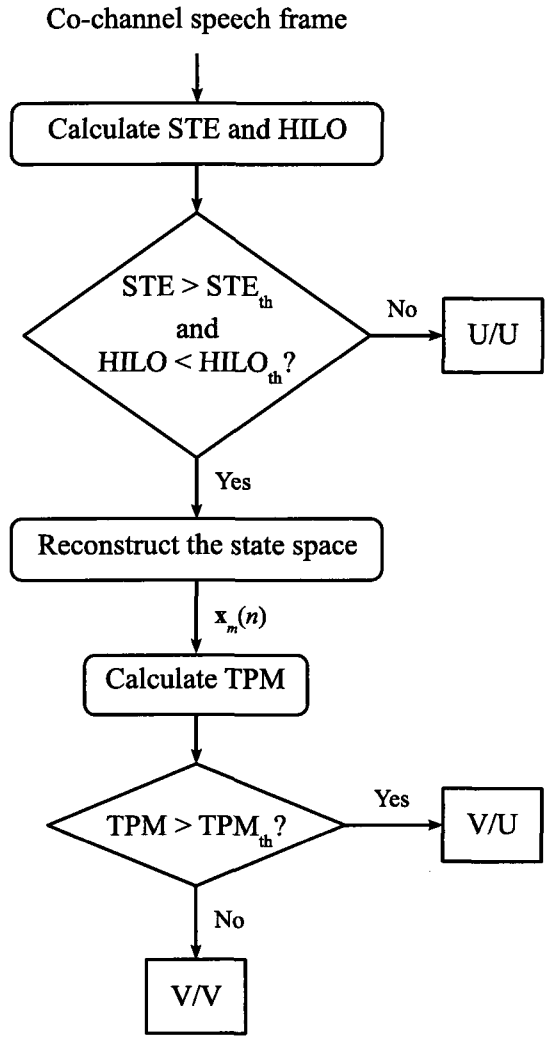


Figure 4.3: Flow chart of the proposed algorithm.

STE: the frame short-time energy in dB calculated using equation (4.2.1).

HILO: the ratio of the energy in the signal above f_H Hz to that below f_L Hz. This value is calculated using equation (4.2.6).

These two features were specifically chosen since both V/U and V/V classes are more pronounced in low frequency and have higher energy compared to the U/U class. This in turn improves the discrimination process compared to other features.

An STE threshold, E_{th}, is used and is calculated as

$$E_{th} = 0.1 * \frac{1}{\acute{M}} \sum_{k=1}^{M} E_k, \tag{4.3.6}$$

where E_k is the STE of the k^{th} frame as given by (4.2.1), M is the total number of frames, and $\acute{M}$ is the total number of non-silent frames. The reason for averaging frame energies over $\acute{M}$ instead of M in (4.3.6) is to eliminate the effect of silent frames on the energy threshold calculation and use only frames with active speech. All speech frames are classified as unvoiced except when both the STE is higher than E_{th} and the HILO ratio is lower than a chosen threshold, HILO_{th}.

2. For those frames classified as voiced in step 1, the state space is reconstructed by using the method of delay embedding (4.3.1) with a suitable embedding dimension m and time delay d.

3. For a vector $\mathbf{x}_m(n)$ in the state space, the nearest neighbour points on the nearest neighbour trajectories are located. A nearest neighbour, $\mathbf{x}_m(n_i)$, is found by searching for the point that minimizes the distance

$$\min_{n_i} \|\mathbf{x}_m(n) - \mathbf{x}_m(n_i)\| . \tag{4.3.7}$$

A further constraint is imposed such that nearest neighbours should have a temporal separation greater than the minimum pitch period (about 2.5 ms).

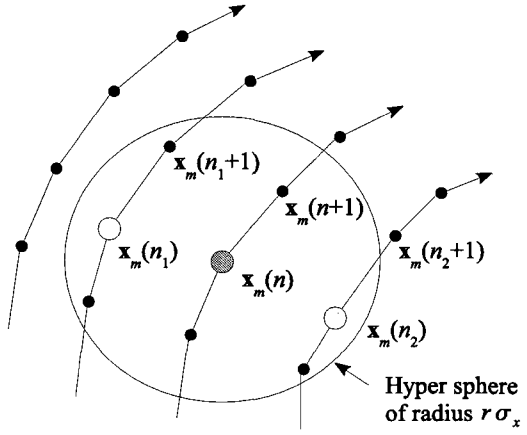


Figure 4.4: Searching for nearest neighbours (white circles) to the query point (gray circle).

The search for the nearest neighbours is also limited to a maximum spatial distance of $r\sigma_x$ in the state space, as shown in Figure 4.4 [143], where r is a constant and σ_x is the standard deviation of the speech samples. The reason for taking the radius of the hyper sphere as a function of σ_x is to make it adaptable to changes in the signal level.

4. The overall TPM of the given frame is calculated using (4.3.4) and compared to a threshold TPM_{th}. If the TPM is greater than TPM_{th}, then this frame is classified as V/U. Otherwise it is considered as a V/V frame.

4.4 Simulation Results

The performance of the proposed algorithm for voicing-state classification was evaluated by conducting two sets of computer simulations using Matlab. The first set of simulations was carried out to verify the possibility of using the TPM feature with single-speaker speech in the voicing-classification process. In the second set of simulations, the performance of the proposed method described in Section 4.3.2 to classify the voicing states of co-channel speech was evaluated. All simulations were accomplished using the TIMIT database [144]. The TIMIT database is a speech corpus that contains 6,300 files of approximately two to four seconds each of digitally recorded speech at a sampling rate of 16 kHz. Ten sentences are spoken by each of 630 male and female speakers from eight major dialect regions of the United States. All simulations in this thesis were performed using speech files from the first dialect region (New England) of the TIMIT corpus.

4.4.1 Single-speaker case

The first set of simulations was conducted to evaluate the performance of the TPM in classifying voicing states of the speech produced by a single speaker. It is important to study the performance of the TPM with a single speaker since in many cases, even with co-channel speech, good portion of the signal contains a speech that is produced by a single speaker.

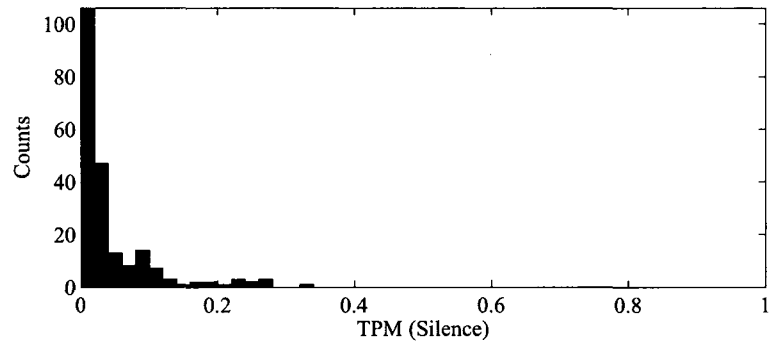
In the first experiment, the goal was to calculate the probability densities of the TPM feature for the three different voicing states of single-speaker speech: silence (S), voiced (V), and unvoiced (U). Thirty speech files from the TIMIT database for 15 male and 15 female speakers were used. Each file was manually segmented into regions of silence, unvoiced, and voiced speech. Afterwards, the speech segments were subdivided into non-overlapped 30 ms frames that were labelled according to their voicing states. For each frame, the state space was reconstructed and the TPM measure was calculated according to (4.3.4) using the values: $m = 3$, $\alpha = 0.85$, and

$r = 0.6$. It is important to note that the speech signal was low-pass filtered before embedding in the state space in order to smooth the flow of trajectories and improve the final results. Finally, the calculated TPM values for all frames, along with the voicing-state labels, were used to create the three histogram plots of Figure 4.5. Each histogram represents an approximate distribution (number of counts) of the TPM values for each voicing class.

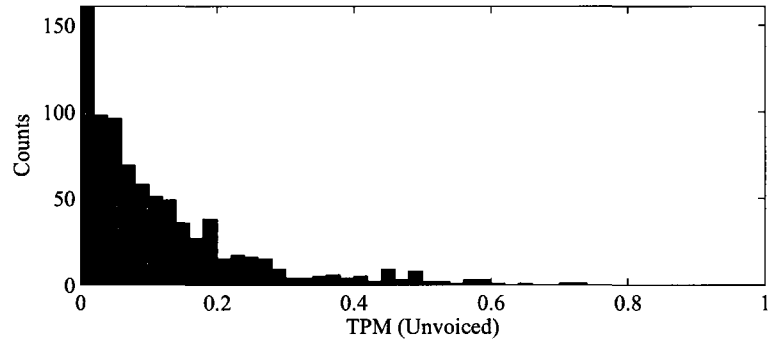
As seen from figures 4.5(a) and 4.5(b), distributions of the TPM for the S and U states are concentrated mainly around low values. Figure 4.5(c), on the other hand, clearly indicates that the distribution of the TPM for the V state is more concentrated around high values. As mentioned earlier, high TPM tends to distinguish voiced sounds from unvoiced sounds and silence. However, the TPM cannot be used alone to discriminate between the S and U states due to the great amount of distribution overlap for these two classes. Histograms also show that a fixed threshold for the TPM in the range between 0.5 and 0.7 can be used to discriminate voiced speech from non-voiced speech (i.e., unvoiced and silent).

In the second experiment, a new set of speech files from the TIMIT database was used. The goal of this experiment was to test the capability of the TPM measure in classifying the voicing states of a single speaker under different conditions. Since the first experiment demonstrated that TPM cannot be used alone to discriminate between the S and U states, the unvoiced data in the present experiment includes both unvoiced and silence regions.

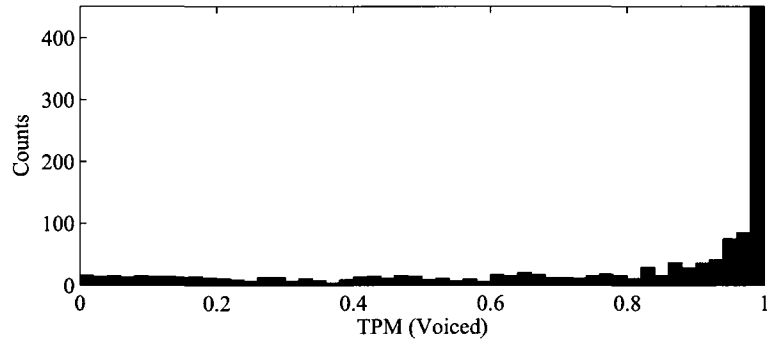
A new 30 speech files (15 males and 15 females) were selected randomly from the TIMIT database and segmented manually into regions of voiced (V) and unvoiced (U and S) speech. Afterward, these segments were subdivided into non-overlapped 30 ms frames that were labelled according to their voicing states (i.e., U or V). This set of labels was referred to as the reference set. The TPM measure was calculated for each frame, similar to the first experiment. A TPM threshold value, TPM_{th}, was set to 0.5 and another set of labels (called the estimated set) was created by comparing



(a)



(b)



(c)

Figure 4.5: Histograms of the distribution of TPM values calculated from the speech frames of a single speaker: (a) silence frames, (b) unvoiced frames, and (c) voiced frames.

Identified voicing-state		Reference voicing-state					
		Clean		SNR = 25 dB		SNR = 10 dB	
		U	V	U	V	U	V
Male	U	98.43%	1.57%	99.01%	0.99%	99.90%	0.10%
	V	24.82%	75.18%	27.37%	72.63%	48.22%	51.78%
Female	U	98.25%	1.75%	98.84%	1.16%	99.43%	0.57%
	V	6.33%	93.67%	9.95%	90.05%	34.02%	65.98%

Table 4.2: Performance of the proposed classification algorithm for single-speaker speech.

the frame's TPM value with TPM_{th}. If the TPM value was greater than TPM_{th}, the current frame was considered voiced and was labelled as (V). Otherwise, the frame was labelled as (U). Finally, the reference and estimated sets of labels were compared and the confusion matrix of Table 4.2 was calculated.

To examine the robustness of the TPM measure under the effect of background noise, a white noise signal was added to the speech data prior to calculating the TPM and the estimated set of labels. The reference set of labels was still calculated by using the clean speech signal. Noise was added at different SNR levels. The resulting confusion matrixes for the cases SNR = 25 dB and SNR = 10 dB are shown in Table 4.2.

It can be concluded from the table that, as the background white noise increased, the percentage of correctly identified voiced segments decreased rapidly. On the other hand, the percentage of correctly identified unvoiced segments increased slightly. This is logical, since the more noise we have in the signal, the more the TPM is drifted to low values. Furthermore, the percentage of correctly identified voiced segments was always higher for female speakers than for male speakers. This was due to the fact that female speakers tend to have higher pitch and hence a higher number of pitch periods per frame. This consequently increased the possibility of finding nearest neighbours and increased the TPM value of voiced female speech.

Speaker 2 \ Speaker 1	Silence	Unvoiced	Voiced
Silence	1% NPN	3% NPN	6% NPN
Unvoiced	3% NPN	9% X	18% P
Voiced	6% NPN	18% P	36% P

Table 4.3: Processing strategies available for the combinations of vocal excitation for the selected speech files from the TIMIT database.

4.4.2 Co-channel case

The second set of simulations was conducted to evaluate the performance of the TPM feature in classifying voicing states of co-channel speech.

In the first experiment, a matrix of observed percentage coincidence for each combination of voicing states similar to the one in Table 4.1 was generated for the TIMT database. Sixty sentences from the TIMIT database uttered by 30 male and 30 female speakers were randomly paired and linearly added at 0 dB TIR to form 30 co-channel speech files. Prior to addition, each file was segmented into non-overlapped 30 ms frames and manually labelled as either silent, unvoiced, or voiced. A frame was labelled voiced if 30% of it falls within a voiced phonetic marking on the TIMIT database. The percentage of occurrence of each combination was calculated and the results were listed in Table 4.3. By a quick comparison between Table 4.1 and Table 4.3, it can be concluded that the TIMIT database has fewer silence periods, more unvoiced periods, and similar voiced periods compared to the MIT-CBG database.

To determine the statistics of the TPM values for co-channel speech, another experiment was conducted. In this experiment, the TPM value of each frame of the 30 co-channel speech files used in the previous experiment was calculated. A set of

reference voicing states was manually created from the speech files prior to mixing as follows:

1. The frame was labelled as U/U if both speakers were either in the unvoiced state or silent.

2. The frame was labelled as V/U if only one speaker was in the voiced state.

3. The frame was labelled as V/V if both speakers were in the voiced state.

To avoid ambiguity in making reference labels, two conditions were applied when manually inspecting the individual and combined voicing states:

1. Frames that were identified as V/V and have TIR greater or less than 15 dB were considered V/U frames.

2. Transition frames were removed from the statistics.

The measured TPM values for each voicing class were then used to create the three histogram plots of Figure 4.6. Each histogram represents approximately the distribution (number of counts) of the TPM values for each class.

As seen from figures 4.6(a) and 4.6(c), distributions of the TPM for the U/U and V/V classes are concentrated mainly around low values. Figure 4.6(b), on the other hand, clearly indicates that the distribution of the TPM for the V/U (and U/V) class is more concentrated around high values. As mentioned earlier, high TPM tends to distinguish voiced sounds from unvoiced sounds and co-channel speech. However, the TPM cannot be used alone to discriminate between the U/U and V/V classes due to the great amount of distribution overlap for these two classes.

The distributions suggest that discrimination between the V/U and V/V classes can be done using a fixed threshold even without too much training. This is due to the small amount of overlap between the two distributions. The noticeable increase of occurrence in the V/V class close to unity is most likely due to frames with fundamental frequencies that are multiples of each other.

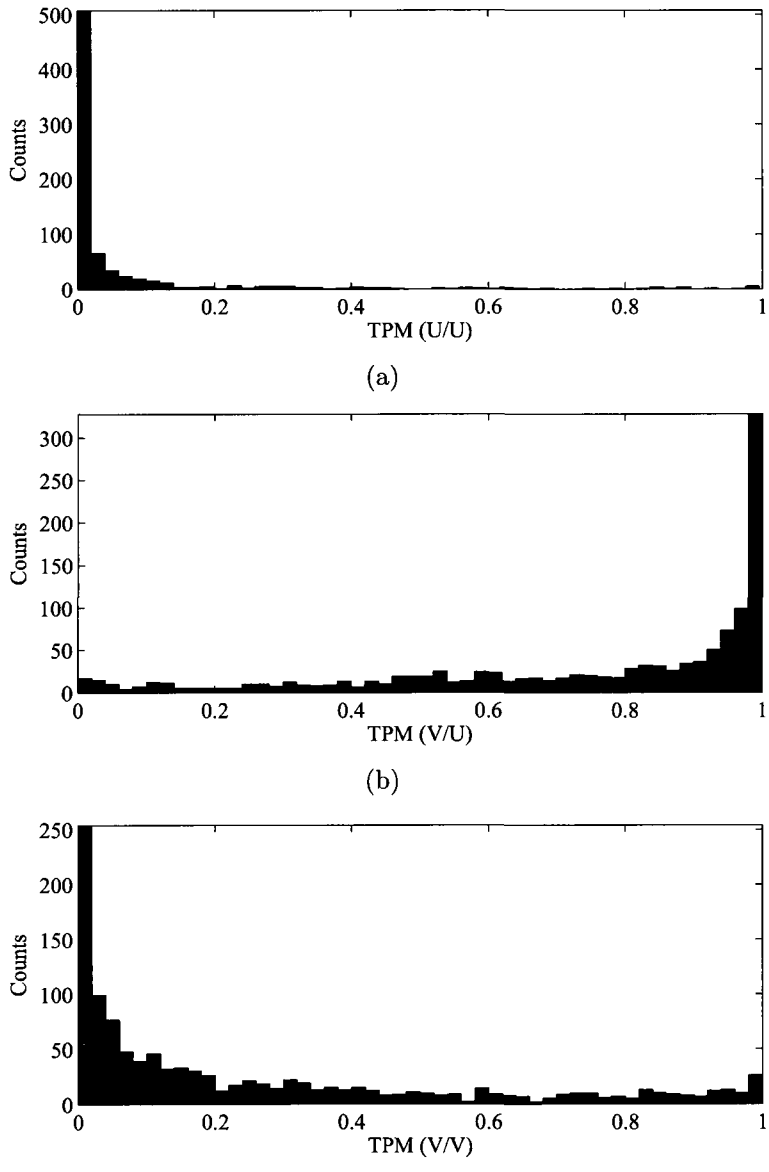


Figure 4.6: Histograms of the distribution of TPM values calculated from co-channel speech frames: (a) U/U frames, (b) V/U frames, and (c) V/V frames.

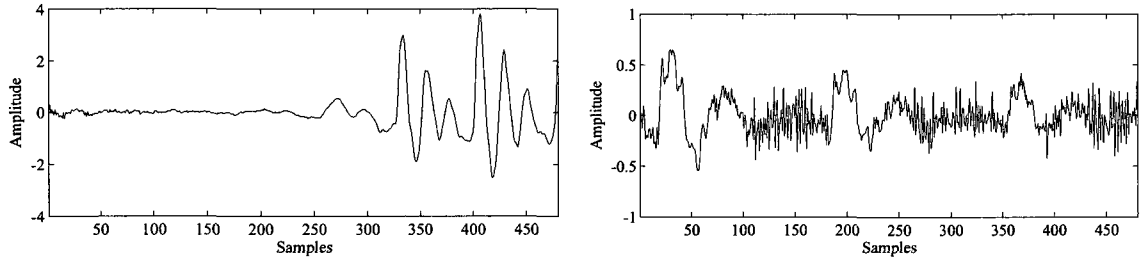
The performance of the proposed algorithm in classifying voicing states of co-channel speech was evaluated by conducting a set of computer simulations. Sixty co-channel speech files consisting of the speech recordings of male/female, male/male and female/female mixtures were randomly created from the TIMIT database. The speech data was sampled at 16 kHz and segmented into 30 ms frames with 50% overlap. Speech data mixtures were tested at TIR values of -5 dB, 0 dB, and 5 dB. The values of f_L = 2 kHz, f_H = 4 kHz, and HILO_{th} = -3 dB were used for the HILO measure calculation. Embedding dimensions of m = 2, 3, and 5 for the state-space reconstruction were compared. The TPM thresholds were set as follows: $\text{TPM}_{th} = 0.5$, $\alpha = 0.85$, and $r = 0.6$.

To start, the voicing states of the speech of individual speakers were identified to have a baseline for comparison. This produced two sets of two-level labels in which a value of 1 was given to every voiced frame and a value of 0 was given to every unvoiced frame. The two sets of labels were then added together to create a third set of three-level labels. This set was referred to as the reference set. A value of 0 corresponded to U/U class, a value of 1 corresponded to V/U class, and a value of 2 referred to V/V class. The reference set was further examined by visual inspection to correct any errors. Voiced-speech frames with TIR less than 15 dB were considered V/V. Finally, a fourth set of three-level labels was obtained by applying the proposed approach on the mixed speech. The last two sets ware compared to determine if any error in classification has occurred.

Table 4.4 summarizes the results obtained by applying the proposed algorithm to the three different mixtures at TIR values of 0 dB, 5 dB, and -5 dB (note that a TIR of 5 dB used in the male/female mixtures means that the male speech signal was 5 dB stronger than the female speech signal). No background noise was added to the speech mixtures during this set of experiments. It can be observed from the above table that the algorithm's capability in determining the U/U state is better than V/U and V/V. However, the overall performance still exceeds 85% at TIR = 0 dB. The accuracy of

Identified voicing-state		**Reference voicing-state**								
		TIR = 0 dB			TIR = 5 dB			TIR = -5 dB		
		U/U	V/U	V/V	U/U	V/U	V/V	U/U	V/U	V/V
Male/ Female	U/U	99%	2.8%	2.4%	99.3%	14.2%	15.5%	99.2%	14.5%	18.2%
	V/U	0.2%	75.5%	16.8%	0.5%	68.1%	17.2%	0.5%	77%	19.8%
	V/V	0.8%	21.7%	80.8%	0.2%	17.7%	67.3%	0.3%	8.5%	62%
Male/ Male	U/U	98.4%	5.4%	4.6%	98.3%	13.6%	10.6%	Same as TIR = 5 dB case		
	V/U	0.2%	68.8%	10.5%	1.0%	57.1%	11.2%			
	V/V	1.4%	25.8%	84.9%	0.7%	29.3%	78.2%			
Female/ Female	U/U	96.2%	3.6%	2.9%	98.7%	12.1%	15.4%	Same as TIR = 5 dB case		
	V/U	1.5%	87.7%	22.9%	0.7%	81.2%	26.7%			
	V/V	2.3%	8.7%	74.2%	0.6%	6.7%	57.9%			

Table 4.4: Performance of the proposed classification algorithm for co-channel speech.



(a) (b)

Figure 4.7: Some sources of error in the proposed method: (a) transition frame of the onset of a voiced sound produced by a female speaker and (b) mixed excitation frame of the phoneme /z/ produced by a male speaker.

identifying V/U frames is higher than the V/V frames for female/female mixtures while the opposite is true for male/male mixtures. The performance of determining the two classes is approximately the same for female/male mixtures. This is due to female voiced speech having a higher pitch frequency (and consequently a greater number of pitch periods per frame) than male voiced speech. This results in more neighbour trajectories in the state space. The overall performance is degraded at TIR = 5 dB and -5 dB compared to TIR = 0 dB case, as shown in Table 4.4.

Three major sources of error are observed from the simulations:

1. Transition frames (onsets and offsets of voiced speech).
2. Frames with mixed excitation.
3. Frames when the pitch frequency of one speaker is approximately an integer multiple of the other speaker's pitch frequency.

While it was sometimes easy to locate transition frames during single-speaker speech, it was difficult to do so with co-channel speech. To reduce the effect of the transition frames, a three-tap median filter was used on the three-level label sets.

Another set of experiments was performed to study the effect of background noise on the performance of the algorithm. Speech mixtures used in the previous experiments were as clean of background noise as possible. For this next set of experiments, Gaussian noise was added to the speech at decreasing levels of SNR to

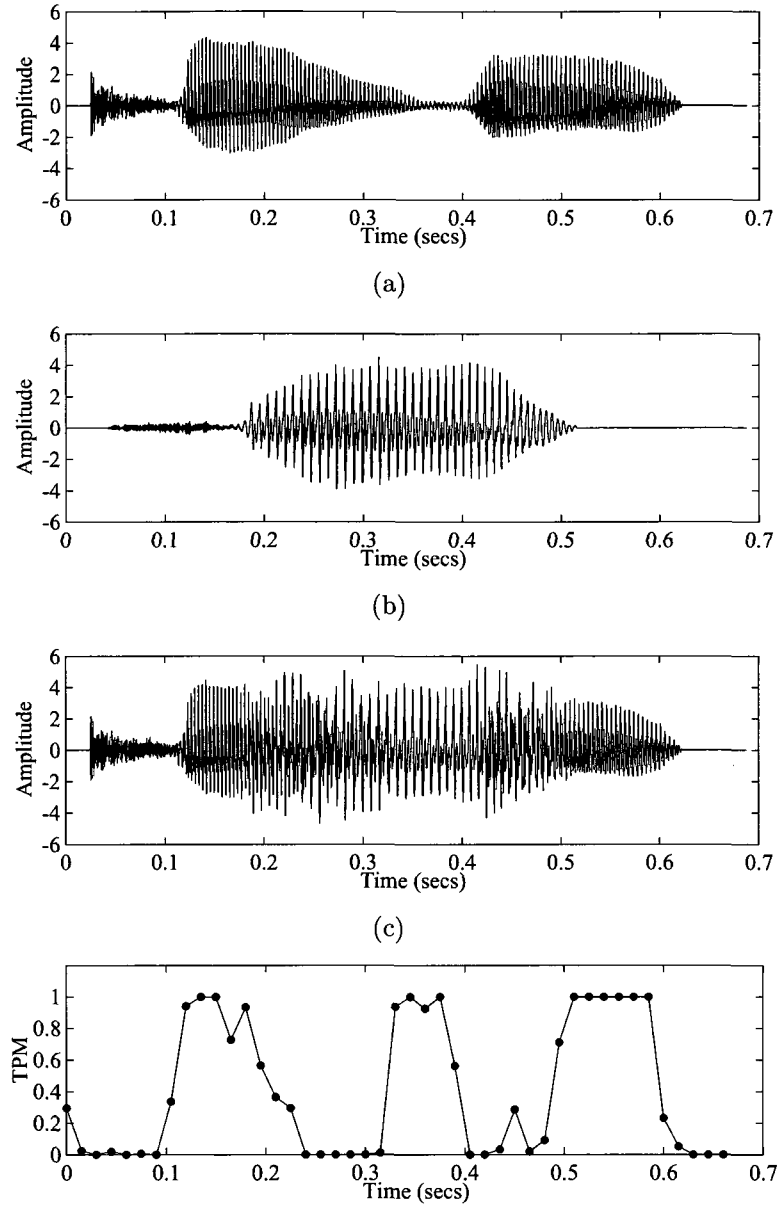
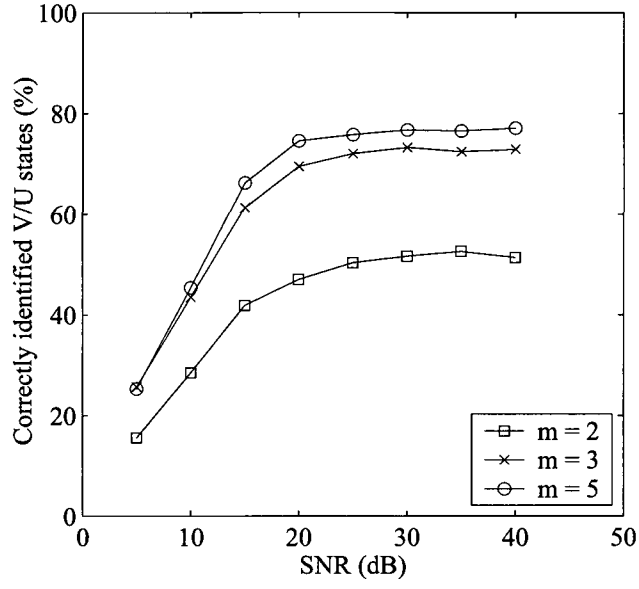


Figure 4.8: Segmental TPM applied to co-channel speech: (a) first waveform of a female speaker uttering the phrase "toll rate," (b) second waveform of a male speaker uttering the word "she," (c) the mixed co-channel speech, and (d) the TPM plot for the waveform in (c).

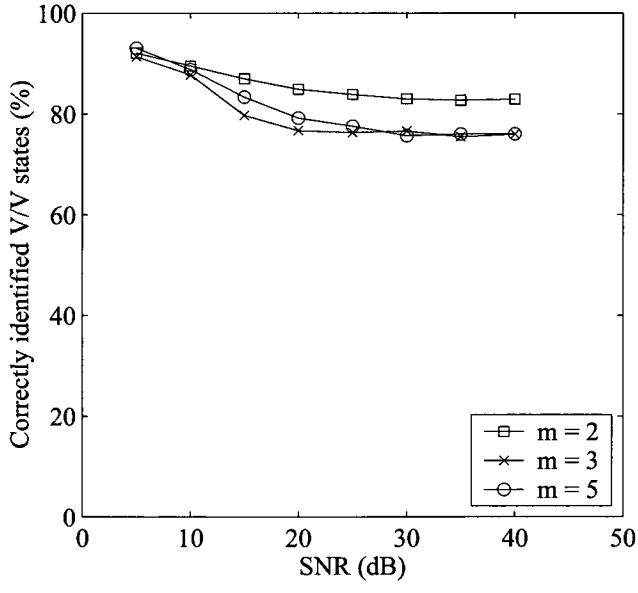
examine the performance of the state-space approach under noisy conditions. The Gaussian noise-contaminated mixed-speech experiments were set up such that the mixtures had 0 dB TIR. This means that the average power level for both speakers was the same before the speech signals were mixed. The Gaussian noise was added with SNR varying from 40 dB down to 5 dB. An important point is that the level of the SNR was determined in relation to the average power of one speaker and not the mixture of speakers. The SNR was not computed based on mixed speech since it was found that results obtained were misleading: many of the mixed speech frames actually contained only one active speaker, so basing the SNR on only one of the speakers seemed more appropriate.

The percentage of correct classifications versus SNR is plotted in figures 4.9(a) and 4.9(b) for V/U class and V/V class respectively. The case for U/U class was not considered since the U/U classification did not use state-space reconstruction, as seen in Figure 4.3. Another parameter considered in this set of experiments was the embedding dimension m. Three separate curves for $m = 2$, 3, and 5 are shown in each plot of Figure 4.9. For the V/U case shown in Figure 4.9(a), the performance of the algorithm for decreasing SNR is a decline in the percentage of correctly identified V/U frames. These results are as expected, but it is interesting to note that there is a plateau at roughly 20 dB of SNR. Below this 20 dB point, there is a linear degradation in correctly identified V/U frames versus SNR. Another anticipated result shown in Figure 4.9(a) is that increasing the embedding dimension m improves the classification results. It can be seen that increasing m is a case of diminishing returns. The percentage of correctly identified V/U frames increases only slightly as m is increased. With the additional computational cost for larger values of m, an m of 5 is likely sufficient for most applications.

From Figure 4.9(b), it can be observed that the results for the V/V class show a decline in performance when the SNR is improved. This is counterintuitive, since we would normally expect the performance to improve as the SNR is improved. Since



(a)



(b)

Figure 4.9: Percentage of correctly identified states versus SNR for different values of embedding dimension and TIR = 0 dB: (a) V/U state and (b) V/V state.

Voicing-state	Percentage of correct identifications		
	Bayesian [37]	**SAPVR** [40]	**Proposed**
U/U	85.9%	N/A	97.9%
V/U	59.8%	71%	77.3%
V/V	90.8%	N/A	80%

Table 4.5: Performance comparison at TIR = 0 dB.

frames at the decision level after state-space reconstruction can be either V/U or V/V, these results suggest that the poor performance in V/U of Figure 4.9(a) is translating into a "better" performance in V/V in Figure 4.9(b). Of course, what is really happening is that V/V frames that were previously misclassified as V/U are now being correctly classified as V/V under high levels of noise.

Table 4.5 shows a comparison of the presented results to the results given in [37] using the Bayesian approach and to the results given in [40] using the SAPVR approach. These two approaches were tested using the same TIMIT speech corpus under similar conditions to our simulations. A total increase of at least 7% in the overall percentage of correctly identified segments is achieved. Taking into consideration that the proposed algorithm does not use training data, this gives it a great advantage over the other two algorithms.

4.5 Summary

The goal of this chapter was to present a new approach to classify different voicing states of co-channel speech based on the nonlinear state-space reconstruction concept discussed in the previous chapter. The algorithm exploited three features to classify a given co-channel speech frame into one of the three voicing states: Unvoiced/Unvoiced (U/U), Voiced/Unvoiced (V/U), and Voiced/Voiced (V/V). These features were: the frame short-time energy (STE), the high-to-low frequency energy ratio (HILO), and the trajectory parallel measure (TPM). Simulation results showed that the proposed algorithm was capable of successfully classifying the voicing states of single-speaker and co-channel speech without using either *a priori* information or training data sets. This was due to the high separability characteristics of the distributions of the TPM for different classes such as the V and U states in single-speaker speech and the V/U and V/V classes in co-channel speech. However, the algorithm could not be used alone to discriminate between the S and U states or the U/U and V/V classes.

CHAPTER 5

Estimation of the Sinusoidal Model Parameters of Co-Channel Speech

5.1 Introduction

Minimizing artifacts in the processed co-channel speech is a key concern, especially if the final goal is to use the recovered speech in machine-based applications such as automatic speech recognition and speaker identification systems. Several previous studies have developed signal-processing algorithms for modelling and separating co-channel speech. The primary approaches have taken the harmonic structure of voiced speech as the basis for separation and have used either frequency-domain spectral analysis and reconstruction [11, 30, 31, 145] or time-domain filtering [146]. One promising approach to address co-channel speech separation is to exploit a speech analysis/synthesis system based on sinusoidal modelling of speech. For example, in [30] and [31], a voiced segment of co-channel speech is modelled as the sum of harmonically related sine waves with constant amplitudes, frequencies, and phases. In the sinusoidal modelling approach, the speech parameters of individual speakers are estimated by applying a high-resolution short-time Fourier transform (STFT) to the windowed speech waveform. The frequencies of underlying sine waves are assumed to be known *a priori* from the individual speech waveforms or they are determined by using a simple frequency-domain peak-picking algorithm. The amplitudes and

phases of the component waves are then estimated at these frequencies by performing a least-squares (LS) algorithm. This technique has the following drawbacks:

1. The accuracy of the estimate is limited by the frequency resolution of the STFT.

2. Error is introduced due to the edge effects of the window function used for the STFT.

This chapter presents a time-domain method to precisely estimate the sinusoidal model parameters of co-channel speech. The method does not require the calculation of the STFT or the multiplication by a window function. It incorporates a time-domain least-squares estimator and an adaptive technique to model and separate the co-channel speech into its individual speakers. The performance of the proposed method is evaluated using a database consisting of a wide variety of mixed male and female speech signals at different target-to-interference ratios (TIRs) [4].

The chapter is organized as follows. In Section 5.2, the sinusoidal model of co-channel speech consisting of K speakers is presented. A well-known frequency-domain method for estimating the sinusoidal model parameters [30] is presented in Section 5.3. Our proposed time-domain method for solving the same problem is discussed in Section 5.4. Finally, experimental results and comparisons of the two techniques are reported and discussed in Section 5.5.

5.2 Sinusoidal Modelling of Co-Channel Speech

According to the speech analysis/synthesis approach based on the sinusoidal model [147], a short segment of voiced speech (about 20 to 30 ms) can be represented (analyzed) as the sum of harmonically related sinusoidal waves with slowly varying amplitudes, frequencies, and phases. Representing speech signals by the sinusoidal parameters implies that the original signal can be reconstructed (synthesized) by summing up the sinusoidal components. The concept of sinusoidal modelling can be applied with relatively good accuracy to voiced segments of single-speaker speech

as well as co-channel speech [30]. For co-channel speech, a speech frame consisting of the addition of voiced sounds produced by K speakers can be represented using sinusoidal modelling as follows:

$$\begin{aligned} x(n) &= \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} c_\ell^{(k)} \cos\left(\ell\omega^{(k)} n - \phi_\ell^{(k)}\right) \\ &= \sum_{k=1}^{K} \sum_{l=1}^{L_k} \left[a_\ell^{(k)} \cos(\ell\omega^{(k)} n) + b_\ell^{(k)} \sin(\ell\omega^{(k)} n)\right], \end{aligned} \tag{5.2.1}$$

where $n = 0, \ldots, N-1$ is the discrete time index, $\omega^{(k)}$ is the fundamental frequency for that frame of the k^{th} speaker, and $c_\ell^{(k)}$, $\ell\omega^{(k)}$, and $\phi_\ell^{(k)}$ denote the amplitude, frequency, and phase, respectively, of the ℓ^{th} harmonic of the k^{th} speaker. The total number of harmonics in each speaker's model is referred to as L_k for $k = 1, \ldots, K$. The quadrature amplitude parameters $a_\ell^{(k)}$ and $b_\ell^{(k)}$ are related to $c_\ell^{(k)}$ and $\phi_\ell^{(k)}$ as follows [30]:

$$\begin{aligned} a_\ell^{(k)} &= c_\ell^{(k)} \cos\left(\phi_\ell^{(k)}\right) \\ b_\ell^{(k)} &= c_\ell^{(k)} \sin\left(\phi_\ell^{(k)}\right) \end{aligned} \tag{5.2.2}$$

and

$$\begin{aligned} c_\ell^{(k)} &= \sqrt{\left(a_\ell^{(k)}\right)^2 + \left(b_\ell^{(k)}\right)^2} \\ \phi_\ell^{(k)} &= \tan^{-1}\left(\frac{b_\ell^{(k)}}{a_\ell^{(k)}}\right). \end{aligned} \tag{5.2.3}$$

A precise estimate of the sinusoidal model parameters is important for separating the co-channel speech into its individual components. The basic problem addressed in this chapter can be stated as follows. Given the real observed N samples of the co-channel speech sequence $x(n)$, find the parameters $\hat{L}_k$, $\hat{\omega}^{(k)}$, $\left\{\hat{a}_\ell^{(k)}\right\}_{\ell=1}^{\hat{L}_k}$, and $\left\{\hat{b}_\ell^{(k)}\right\}_{\ell=1}^{\hat{L}_k}$ of the sequence

$$\hat{x}(n) = \sum_{k=1}^{K} \sum_{\ell=1}^{\hat{L}_k} \left[\hat{a}_\ell^{(k)} \cos(\ell\hat{\omega}^{(k)} n) + \hat{b}_\ell^{(k)} \sin(\ell\hat{\omega}^{(k)} n)\right] \tag{5.2.4}$$

that best fits $x(n)$ by minimizing the mean square error (MSE)

$$E = \frac{1}{N}\sum_{n=0}^{N-1}[x(n) - \hat{x}(n)]^2. \tag{5.2.5}$$

In the remaining sections, we will consider the case of two speakers ($K = 2$) to represent the co-channel speech without loss of generality. In this case, (5.2.1) is reduced to

$$\begin{aligned} x(n) = &\sum_{l=1}^{L_1}\left[a_\ell^{(1)}\cos(\ell\omega^{(1)}n) + b_\ell^{(1)}\sin(\ell\omega^{(1)}n)\right] \\ &+ \sum_{l=1}^{L_2}\left[a_\ell^{(2)}\cos(\ell\omega^{(2)}n) + b_\ell^{(2)}\sin(\ell\omega^{(2)}n)\right]. \end{aligned} \tag{5.2.6}$$

The solution to the minimization problem in (5.2.5) can be handled in either frequency domain or time domain. In the following two sections, we will present two methods for solving this problem. The first is a well-known frequency-domain method originated in [30] by Quatieri and Danisewicz (1990). The second method is the proposed time-domain technique that provides improved accuracy and eliminates windowing errors.

5.3 Frequency-Domain Estimation of Model Parameters

In the frequency-domain solution of (5.2.5) [11,30,31,145], the sinusoidal parameters are normally estimated by sampling the short-term spectrum of $x(n)$ at the harmonic frequencies of the individual signals and solving a system of linear equations using the LS algorithm. This can be done as follows. First, $x(n)$ is multiplied by a window function, $w(n)$, and the short-term spectrum, $X(\omega)$, is calculated by applying the STFT to (5.2.6). If the window function is real and symmetric, its spectrum will be real too. Hence, the real and imaginary parts of the short-term spectrum can be

written as

$$\mathrm{Re}[X(\omega)] = \frac{1}{2}\left[\sum_{\ell=1}^{L_1} a_\ell^{(1)} W(\omega - \ell\omega^{(1)}) + \sum_{\ell=1}^{L_2} a_\ell^{(2)} W(\omega - \ell\omega^{(2)})\right]$$
$$\mathrm{Im}[X(\omega)] = \frac{1}{2}\left[\sum_{\ell=1}^{L_1} b_\ell^{(1)} W(\omega - \ell\omega^{(1)}) + \sum_{\ell=1}^{L_2} b_\ell^{(2)} W(\omega - \ell\omega^{(2)})\right], \tag{5.3.1}$$

where $W(\omega)$ is the spectrum of the normalized window function. Here, for simplicity, we assumed that the contribution of negative frequency is neglected (i.e., one-sided spectrum).

If we assume that the fundamental frequencies of individual signals are known *a priori*, we can arrange the harmonic frequencies of both speakers in an ascending order in the vector

$$\boldsymbol{\omega} = \left\{\{\ell\omega^{(1)}\}_{\ell=1}^{L_1} \bigcup \{\ell\omega^{(2)}\}_{\ell=1}^{L_2}\right\} = [\ \omega_1,\ \omega_2,\ \cdots,\ \omega_L\ ]^T \tag{5.3.2}$$

with $\omega_1 < \omega_2 < \ldots < \omega_L$, where $L = L_1 + L_2$. Likewise, corresponding quadrature parameters can be arranged in the same order in the vectors

$$\mathbf{a} = [\ a_1,\ a_2,\ \ldots,\ a_L\ ]^T \tag{5.3.3}$$

and

$$\mathbf{b} = [\ b_1,\ b_2,\ \ldots,\ b_L\ ]^T. \tag{5.3.4}$$

Finally, these quadrature parameters can be estimated by solving the following two sets of linear equations

$$\hat{\mathbf{a}} = 2\ \mathbf{G}^{-1}\ \mathrm{Re}[\mathbf{X}(\boldsymbol{\omega})] \tag{5.3.5}$$

$$\hat{\mathbf{b}} = 2\ \mathbf{G}^{-1}\ \mathrm{Im}[\mathbf{X}(\boldsymbol{\omega})] \tag{5.3.6}$$

where $\mathbf{X}(\boldsymbol{\omega})$ is a vector containing the complex values of the sampled short-term spectrum $X(\omega)$ at the frequencies in the vector $\boldsymbol{\omega}$ and G is a symmetrical matrix

containing the real values of the sampled window function at frequency differences

$$\mathbf{G} = \begin{bmatrix} W(0) & W(\omega_1 - \omega_2) & \dots & W(\omega_1 - \omega_L) \\ W(\omega_2 - \omega_1) & W(0) & \dots & W(\omega_2 - \omega_L) \\ \vdots & \vdots & \ddots & \vdots \\ W(\omega_L - \omega_1) & W(\omega_L - \omega_2) & \dots & W(0) \end{bmatrix}. \tag{5.3.7}$$

Practically, the short-term spectrum is performed using a Hann or a Hamming window and an FFT with a large number of points (typically 4096 points) [30]. This large FFT is used to give sufficient frequency resolution for adequate separation.

5.4 Proposed Time-Domain Estimation of Model Parameters

In this section, a time-domain method to precisely estimate the sinusoidal model parameters of co-channel speech is proposed. The method does not require the calculation of the STFT or the multiplication by a window function. It incorporates a time-domain least-squares estimator and an adaptive technique to model and separate the co-channel speech into its individual speakers.

5.4.1 Estimation setup

In matrix notation, we may rewrite (5.2.4) as

$$\hat{\mathbf{x}} = \mathbf{Q}\mathbf{h}, \tag{5.4.1}$$

where $\hat{\mathbf{x}}$ is the vector

$$\hat{\mathbf{x}} = \begin{bmatrix} \hat{x}(0), & \hat{x}(1), & \dots, & \hat{x}(N-1) \end{bmatrix}^T \tag{5.4.2}$$

and $\mathbf{h}$ is given as

$$\mathbf{h} = \begin{bmatrix} \mathbf{h}^{(1)} \\ \mathbf{h}^{(2)} \end{bmatrix} \tag{5.4.3}$$

with

$$\mathbf{h}^{(k)} = \begin{bmatrix} \mathbf{a}^{(k)} \\ \mathbf{b}^{(k)} \end{bmatrix}, \tag{5.4.4}$$

where

$$\mathbf{a}^{(k)} = \begin{bmatrix} \hat{a}_1^{(k)}, & \hat{a}_2^{(k)}, & \ldots, & \hat{a}_{\hat{L}_k}^{(k)} \end{bmatrix}^T \tag{5.4.5}$$

and

$$\mathbf{b}^{(k)} = \begin{bmatrix} \hat{b}_1^{(k)}, & \hat{b}_2^{(k)}, & \ldots, & \hat{b}_{\hat{L}_k}^{(k)} \end{bmatrix}^T. \tag{5.4.6}$$

$\mathbf{Q}$ is a matrix of the form

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}^{(1)} & \mathbf{Q}^{(2)} \end{bmatrix}, \tag{5.4.7}$$

where

$$\mathbf{Q}^{(k)} = \begin{bmatrix} \mathbf{Q}_a^{(k)} & \mathbf{Q}_b^{(k)} \end{bmatrix} \tag{5.4.8}$$

with the matrices elements given as

$$Q_a^{(k)}(i,j) = \cos(ij\hat{\omega}^{(k)}) \tag{5.4.9}$$

and

$$Q_b^{(k)}(i,j) = \sin(ij\hat{\omega}^{(k)}), \tag{5.4.10}$$

for $i = 0, 1, \ldots, N-1$, $j = 1, 2, \ldots, \hat{L}_k$ and $k = 1, 2$. The MSE in (5.2.5) can now be written as

$$E = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \mathbf{x}^T\mathbf{x} + \hat{\mathbf{x}}^T\hat{\mathbf{x}} - 2\hat{\mathbf{x}}^T\mathbf{x}, \tag{5.4.11}$$

where

$$\mathbf{x} = \begin{bmatrix} x(0), & x(1), & \ldots, & x(N-1) \end{bmatrix}^T. \tag{5.4.12}$$

Substituting (5.4.1) into (5.4.11) gives

$$E = \mathbf{x}^T\mathbf{x} + \mathbf{h}^T\mathbf{Q}^T\mathbf{Q}\mathbf{h} - 2\mathbf{h}^T\mathbf{Q}^T\mathbf{x}. \tag{5.4.13}$$

The estimation criterion is to seek the minimization of (5.4.13) over the parameters $\hat{L}_k$, $\hat{\omega}^{(k)}$, $\left\{\hat{a}_l^{(k)}\right\}_{l=1}^{\hat{L}_k}$, and $\left\{\hat{b}_l^{(k)}\right\}_{l=1}^{\hat{L}_k}$.

The most important and difficult part in the estimation process is to estimate the fundamental frequencies $\{\omega^{(k)}\}_{k=1,2}$. Unfortunately, without *a priori* knowledge of the frequency parameters, direct minimization of (5.4.13) is a highly nonlinear problem that is difficult to solve. Furthermore, the MSE surface in (5.4.13) has several local minima besides the global minimum that might mislead the estimation process if the search is not performed close enough to the optimum frequencies. If the fundamental frequencies are known *a priori* or can be estimated precisely (as is done iteratively in Section 5.4.4), one can easily find the optimum values of the other parameters accordingly (as discussed in Section 5.4.3).

5.4.2 Estimating the number of harmonics

If the fundamental frequencies $\hat{\omega}^{(k)}$ are assumed to be known, the total number of harmonics in each signal can be estimated simply as

$$\hat{L}_k = \left\lfloor \frac{\pi}{\hat{\omega}^{(k)}} \right\rfloor. \tag{5.4.14}$$

Practically, $\hat{L}_k$ is chosen much smaller than the value calculated by (5.4.14), since most of the energy of voiced speech is concentrated below 2 kHz. Using this assumption can dramatically reduce the computational complexity of the system.

5.4.3 Estimating the amplitude parameters

The optimum values of the quadrature parameters $\left\{a_\ell^{(k)}\right\}_{\ell=1}^{L_k}$ and $\left\{b_\ell^{(k)}\right\}_{\ell=1}^{L_k}$ can be estimated directly (assuming the availability of the fundamental frequencies) by finding the standard linear LS solution to (5.4.13) as follows [148]:

$$\mathbf{h}_{opt} = \left(\mathbf{Q}^T\mathbf{Q}\right)^{-1}\mathbf{Q}^T\mathbf{x} = \mathbf{R}^{-1}\mathbf{P}, \tag{5.4.15}$$

where

$$\mathbf{R} = \mathbf{Q}^T\mathbf{Q} \tag{5.4.16}$$

and

$$\mathbf{P} = \mathbf{Q}^T\mathbf{x}. \tag{5.4.17}$$

The minimum MSE corresponding to $\mathbf{h}_{opt}$ is given by substituting (5.4.15) into (5.4.13) to give

$$E_{\min} = \mathbf{x}^T\mathbf{x} - \mathbf{P}^T\mathbf{R}^{-1}\mathbf{P}. \tag{5.4.18}$$

5.4.4 Estimating the fundamental frequencies

Since, in practical applications, the fundamental frequencies of the individual speech waveforms are not known *a priori*, they must be estimated from the mixed data. A direct approach to solve this problem is to search the K-dimensional MSE surface for its minimum with respect to the fundamental frequencies. Obviously, this is an exhaustive process that is not feasible in real-time applications. Furthermore, the MSE surface does not have a unique minimum as shown in Figure 5.1(b). Therefore, searching the entire MSE surface for optimum frequencies can be misleading by converging to a local minimum. An alternative approach is to narrow the search procedure around the optimum fundamental frequencies by starting with initial guesses that are close enough to the optimum values. These initial estimates can be determined either from the previous frames or by applying a simple rough multi-pitch estimation algorithm such as the one proposed in [82]. After obtaining the initial guesses for the $\omega^{(k)}$, the optimum fundamental frequencies can be estimated by searching the MSE surface of (5.4.18) using the method of the steepest descent [149]. Using weight vector $\mathbf{w} = [\ \hat{\omega}^{(1)},\ \ \hat{\omega}^{(2)}\ ]^T$, we describe the steepest descent algorithm by

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \frac{1}{2}\mu\nabla\mathbf{E}(t), \tag{5.4.19}$$

where t is the iteration time index, the gradient is given as

$$-\nabla\mathbf{E}(t) = \begin{bmatrix} -\nabla E^{(1)}(t) \\ -\nabla E^{(2)}(t) \end{bmatrix}, \tag{5.4.20}$$

and μ is a positive scalar that controls both the stability and the speed of convergence. The gradient of the MSE is calculated by differentiating (5.4.18) with respect to each

fundamental frequency as follows:

$$\begin{aligned}
-\nabla E^{(k)}(t) &= -\frac{\partial}{\partial \hat{\omega}^{(k)}}\left(E_{min}(t)\right) \\
&= \frac{\partial}{\partial \hat{\omega}^{(k)}}\left(\mathbf{P}^T\mathbf{R}^{-1}\mathbf{P}\right) \\
&= \frac{\partial}{\partial \hat{\omega}^{(k)}}\left(\mathbf{P}^T\right)\mathbf{R}^{-1}\mathbf{P} + \mathbf{P}^T\frac{\partial}{\partial \hat{\omega}^{(k)}}\left(\mathbf{R}^{-1}\right)\mathbf{P} + \mathbf{P}^T\mathbf{R}^{-1}\frac{\partial}{\partial \hat{\omega}^{(k)}}\left(\mathbf{P}\right) \\
&= \mathbf{x}^T\dot{\mathbf{Q}}\left(\mathbf{R}^{-1}\mathbf{P}\right) - \left(\mathbf{P}^T\mathbf{R}^{-1}\right)\dot{\mathbf{R}}\left(\mathbf{R}^{-1}\mathbf{P}\right) + \left(\mathbf{R}^{-1}\mathbf{P}\right)^T\dot{\mathbf{Q}}^T\mathbf{x} \\
&= \mathbf{x}^T\dot{\mathbf{Q}}\mathbf{h}_{opt} - \mathbf{h}_{opt}^T\dot{\mathbf{R}}\mathbf{h}_{opt} + \mathbf{h}_{opt}^T\dot{\mathbf{Q}}^T\mathbf{x},
\end{aligned} \tag{5.4.21}$$

where

$$\dot{\mathbf{Q}} = \frac{\partial \mathbf{Q}}{\partial \hat{\omega}^{(k)}} \tag{5.4.22}$$

and

$$\dot{\mathbf{R}} = \frac{\partial \mathbf{R}}{\partial \hat{\omega}^{(k)}}. \tag{5.4.23}$$

Note that all vectors and matrices of the above three equations (except $\mathbf{x}$) are functions of the time argument t. For simplicity, this time argument has been removed from these equations. Differentiating (5.4.7) and (5.4.16) and substituting into (5.4.21) gives

$$\begin{aligned}
-\nabla E^{(k)}(t) &= \mathbf{x}^T\dot{\mathbf{Q}}\mathbf{h}_{opt} - \mathbf{h}_{opt}^T\left(\mathbf{Q}^T\dot{\mathbf{Q}} + \dot{\mathbf{Q}}^T\mathbf{Q}\right)\mathbf{h}_{opt} + \mathbf{h}_{opt}^T\dot{\mathbf{Q}}^T\mathbf{x} \\
&= \mathbf{x}^T\dot{\mathbf{Q}}\mathbf{h}_{opt} - \mathbf{h}_{opt}^T\mathbf{Q}^T\dot{\mathbf{Q}}\mathbf{h}_{opt} - \mathbf{h}_{opt}^T\dot{\mathbf{Q}}^T\mathbf{Q}\mathbf{h}_{opt} + \mathbf{h}_{opt}^T\dot{\mathbf{Q}}^T\mathbf{x} \\
&= \left(\mathbf{x}^T - \mathbf{h}_{opt}^T\mathbf{Q}^T\right)\dot{\mathbf{Q}}\mathbf{h}_{opt} + \mathbf{h}_{opt}^T\dot{\mathbf{Q}}^T\left(\mathbf{x} - \mathbf{Q}\mathbf{h}_{opt}\right) \\
&= \left(\mathbf{x} - \mathbf{Q}\mathbf{h}_{opt}\right)^T\dot{\mathbf{Q}}\mathbf{h}_{opt} + \left(\dot{\mathbf{Q}}\mathbf{h}_{opt}\right)^T\left(\mathbf{x} - \mathbf{Q}\mathbf{h}_{opt}\right) \\
&= 2\left(\mathbf{x} - \mathbf{Q}\mathbf{h}_{opt}\right)^T\dot{\mathbf{Q}}\mathbf{h}_{opt}.
\end{aligned} \tag{5.4.24}$$

Note that

$$\dot{\mathbf{Q}}\mathbf{h}_{opt} = \frac{\partial \mathbf{Q}}{\partial \hat{\omega}^{(k)}}\mathbf{h}_{opt} = \dot{\mathbf{Q}}^{(k)}\mathbf{h}_{opt}^{(k)}. \tag{5.4.25}$$

This simplifies (5.4.24) to

$$-\nabla E^{(k)}(t) = 2\left(\mathbf{x} - \mathbf{Q}\mathbf{h}_{opt}\right)^T\dot{\mathbf{Q}}^{(k)}\mathbf{h}_{opt}^{(k)}. \tag{5.4.26}$$

The fundamental frequencies are updated iteratively using (5.4.19). After each iteration, the optimum amplitude parameters corresponding to the estimated frequencies are calculated using (5.4.15). The iteration terminates once the error calculated in a step exceeds the error calculated in the preceding step. After each termination, the iteration can be restarted with a reduced step size μ in order to obtain greater accuracy. Note that even by using (5.4.19), final estimates of fundamental frequencies may still have small inaccuracies because frequencies may vary slightly within the speech frame due to the quasi-periodic nature of voiced speech. The use of exact gradient to update the fundamental frequencies in (5.4.19) gives an advantage compared to [30], where an approximation of the gradient is used. Furthermore, the gradient calculation is an integrated process in the proposed time-domain method since the components on the right-hand side of equation (5.4.26) are already part of the previous steps in the algorithm.

An example of the MSE surface obtained for the single-speaker case ($K = 1$) is shown in Figure 5.1. Figure 5.1(a) shows a 30 ms speech frame for a single speaker, while Figure 5.1(b) shows the corresponding MSE surface using (5.4.18) as the cost function. From Figure 5.1(b), it can be observed that the optimal fundamental frequency of this speech frame is approximately 165 Hz. For the two-speaker case ($K = 2$), the MSE surface would instead be two-dimensional.

Figures 5.2 and 5.3 show an example of applying the proposed algorithm to a frame of co-channel speech. On the left-hand side of Figure 5.2, two speech frames of individual male and female speakers are mixed into one frame. The estimation process of fundamental frequencies is started by the initial values of 94 Hz for the male speech and 165 Hz for the female speech. After convergence, the speech signals are reconstructed using the estimated sinusoidal parameters as shown on the right-hand side of Figure 5.2. Figure 5.3 shows the convergence curves of the fundamental frequencies as well as the MSE. It takes the system approximately 10 iterations to converge to the optimum values.

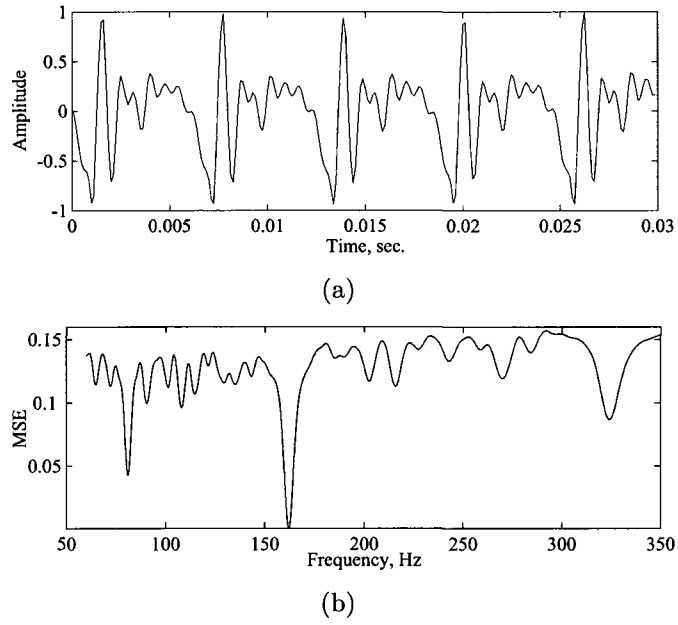


Figure 5.1: Example MSE surface for a single speaker: (a) 30 ms single voiced-speech segment in the time domain, and (b) MSE performance versus fundamental frequency based on (5.4.18).

5.4.5 The ill-conditioned estimation problem

In some instances, the harmonics of the two speakers can be very close to each other. When the harmonics overlap, the correlation matrix $\mathbf{R}$ in (5.4.16) will be singular and the parameter estimation process in (5.4.15) becomes ill-conditioned. To handle this problem, the spacing between adjacent harmonics is continuously calculated. If two adjacent harmonics are found to be closely spaced, e.g., less than 25 Hz apart, only one sinusoid is used to represent these two harmonics. The amplitude parameters of this single component are then estimated and shared equally between the two speakers [30]. It is important to notice that choosing a large number of harmonics to estimate using (5.4.14) when both speakers are in the "voiced" state, increases the chances of harmonic overlap and hence having ill-conditioned correlation matrix.

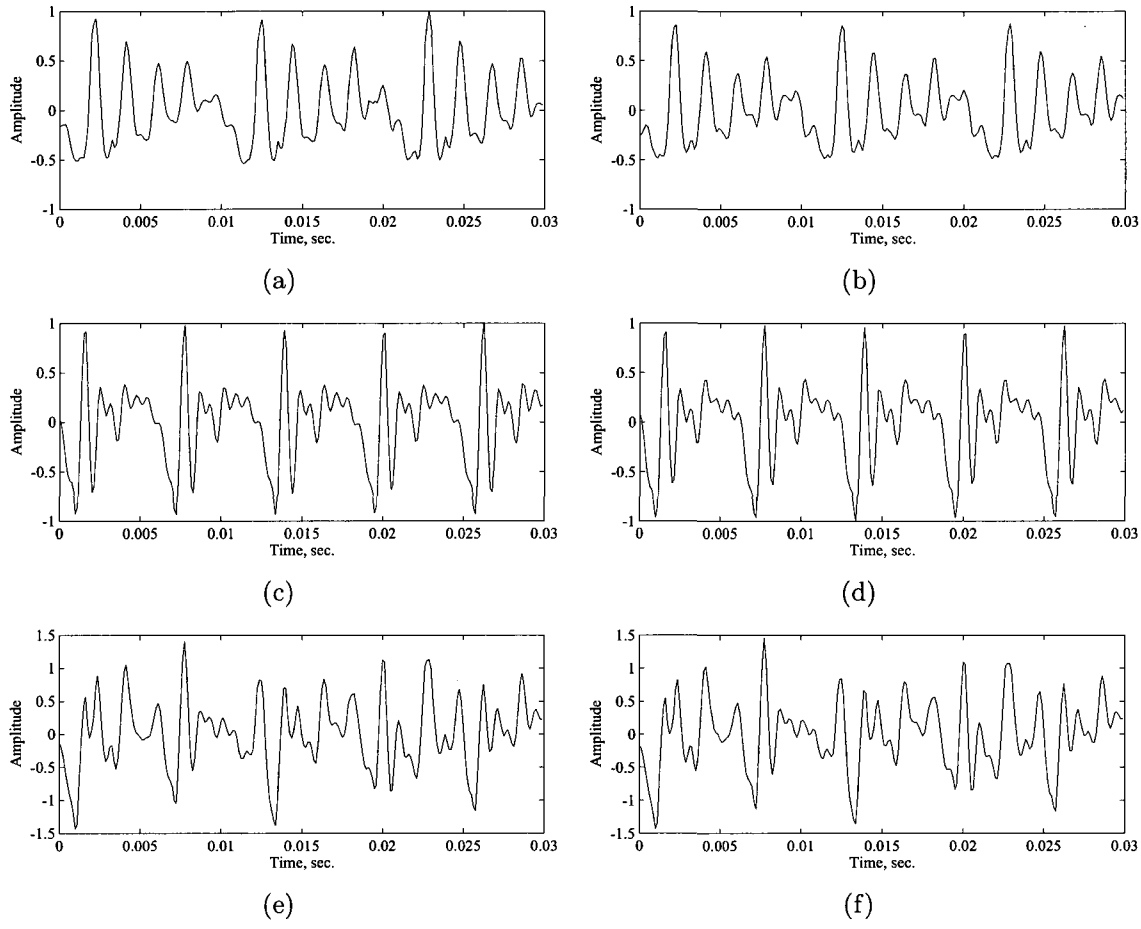


Figure 5.2: Speech recovery using the proposed method after convergence of (5.4.19). On the left-hand side, the original speech frames of (a) male speaker, (c) female speaker, and (e) mixed speech of (a) and (c). On the right-hand side, (b), (d), and (f) show the corresponding reconstructed speech frames for the waveforms in (a), (c), and (e), respectively.

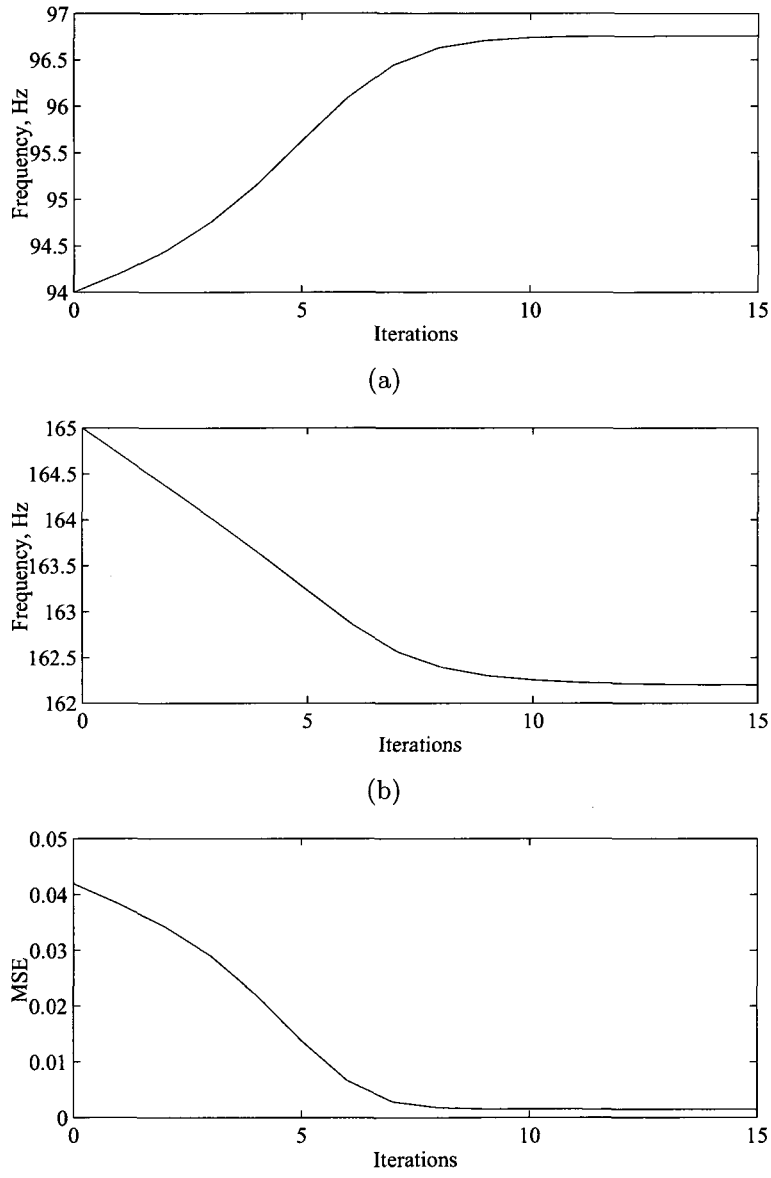


Figure 5.3: Convergence of fundamental frequencies of the proposed method applied to the co-channel speech frame of Figure 5.2(e): (a) convergence of the fundamental frequency of the first speaker, (b) convergence of the fundamental frequency of the second speaker, and (c) convergence of the MSE.

5.5 Simulation Results

The performance of the proposed method is evaluated using a speech database consisting of 200 frames of mixed speech. All voiced-speech segments of 30 ms were randomly chosen from the TIMIT data set [144] for male and female speakers and mixed at different TIRs. The speech data were sampled at a rate of 16 kHz.

Two sets of simulations were conducted to compare the performance of the proposed method with the frequency sampling approach presented in [30]. As suggested by the authors, a Hann window and a high-resolution STFT of length 4096 were used in the frequency-domain technique. To avoid errors due to multi-pitch detection algorithms, the initial guess of the fundamental frequency of each speaker was calculated directly from the original speech frames before mixing, using a simple autocorrelation method.

In the first set of simulations, the comparison was carried out in terms of the signal-to-distortion ratio (SDR) versus TIR, as shown in Figure 5.4, for TIRs ranging from -5 to 15 dB. The SDR measure is defined as [150]

$$SDR\ [\text{dB}] = 10\ \log_{10} \frac{\sum_n s(n)^2}{\sum_n [s(n) - \hat{s}(n)]^2}, \tag{5.5.1}$$

where $s(n)$ is the original target signal before mixing and $\hat{s}(n)$ is the reconstructed signal after separation from the mixture $x(n)$. Each point in Figure 5.4 represents the ensemble average of the SDRs over all 200 test frames. Two cases are considered for each algorithm. In case 1, precise estimation of the fundamental frequencies is done using (5.4.19), and in case 2 only the initial guess of the fundamental frequencies is used. Plots SDR-TD1 and SDR-TD2 are the results for the proposed algorithm in case 1 and case 2, respectively, while plots SDR-FD1 and SDR-FD2 depict the results for the frequency-domain method. As can be seen from Figure 5.4, the SDR increases monotonically for both algorithms with the increase of the TIR in all cases.

More importantly, we see from Figure 5.4 that the proposed technique outperforms

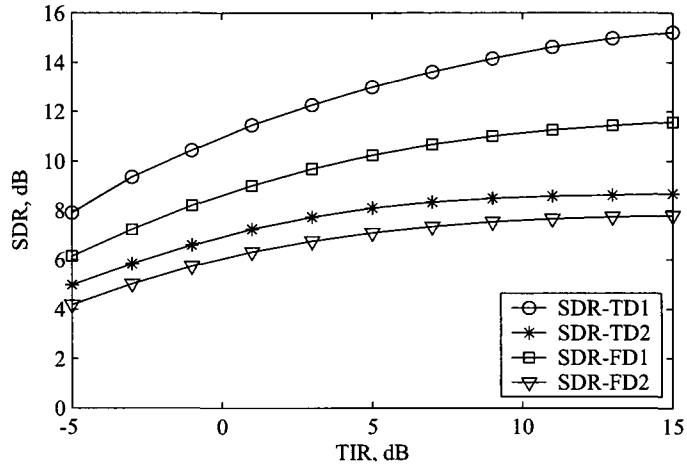


Figure 5.4: SDR results: SDR1 and SDR2 for the proposed time-domain method, and SDR3 and SDR4 for the frequency-domain method, with precise and initial frequency estimates of $\{\omega^{(k)}\}_{k=1,2}$, respectively.

the frequency-domain technique in both case 1 and case 2. At TIR = -5 dB, SDR-TD1 and SDR-TD2 are greater than SDR-FD1 and SDR-FD2 by about 2 and 1 dB, respectively. This difference is greater for larger TIRs. As suggested in Section 5.1, analysis of the resulting estimates using voiced-speech segments has revealed that the discrepancies are due to the limited frequency resolution of the STFT (even with $M = 4096$) and due to the choice of window function and resulting edge effects. Other window functions such as rectangular and Hamming windows had similar discrepancies when tested.

The robustness against background noise was examined in a second set of simulations using the MSE measure versus the signal-to-noise ratio (SNR). Speech segments were corrupted by additive white Gaussian noise (AWGN) with SNR varied from 0 to 15 dB. The results are presented in Figure 5.5. As shown in the figure, the proposed algorithm has a superior performance in low SNR compared to the frequency-domain technique. The AWGN causes additional frequency resolution problems after even a high-resolution STFT. If the proposed time-domain estimation approach is used instead, then the effect of the AWGN is not as severe.

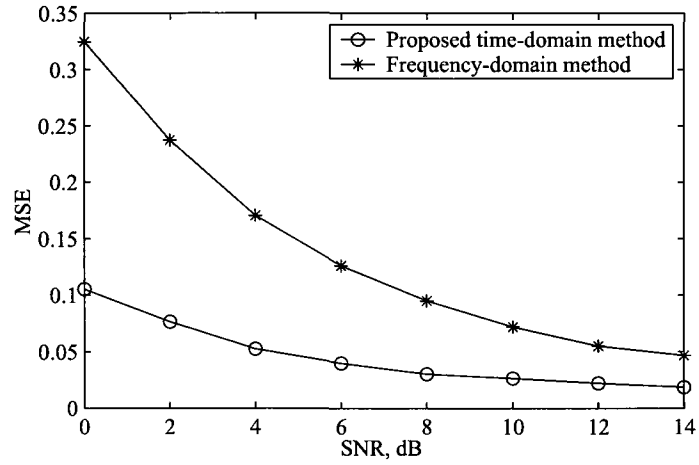


Figure 5.5: MSE results for AWGN for both the proposed time-domain technique and the standard frequency-domain method.

5.6 Summary

The chapter has focused on estimating and separating the sinusoidal model parameters of co-channel speech. A new time-domain method to precisely estimate these parameters was presented. The method did not require the calculation of the discrete Fourier transform or the multiplication by a window function which both degrade the estimate of the sinusoidal model parameters. Separation was performed on a frame-by-frame basis by incorporating a least-squares estimator and an adaptive technique to model and separate the co-channel speech into its individual speakers. The application of this method on real speech signals demonstrated the effectiveness of this method in separating co-channel speech signals with different TIRs. This method was also capable of producing accurate and robust parameter estimation in low SNR situations compared to other existing algorithms.

CHAPTER 6

System Implementation

6.1 Introduction

The algorithms presented in chapters 4 and 5 have been discussed, implemented, and tested individually in order to measure their performance. When building the overall CCSS system, numerical computation issues need to be taken into account in the final implementation. For example, when the proposed algorithms were integrated in the final system shown in Figure 1.1, the overall system encountered slow performance due to high computational complexity. However, there is still room for efficiency and optimization in both the algorithms and the implementation.

The purpose of this chapter is to suggest some modifications to implement the overall separation system in a more efficient way. These modifications are made to reduce the implementation cost as well as to improve the overall performance. Two main modifications are considered: a sample-based TPM method and a simplified sinusoidal parameter estimator. First, we present a modified version of the TPM algorithm for determining voicing states. The algorithm allows the computation of the TPM to be performed on a sample-by-sample basis that makes it reasonable for real-time applications. Afterwards, a simplified method for estimating the sinusoidal model parameters using linear LS solution is presented.

The remainder of this chapter is organized as follows. In the following section, we

describe the modified technique for calculating the TPM. Estimating the sinusoidal model parameters based on linearized LS solution is presented and discussed in Section 6.3. Section 6.4 shows how these algorithms are integrated together to form the overall CCSS system. Finally, the conclusions are given in Section 6.5.

6.2 The Sample-Based TPM Method

As discussed in chapter 4, segmenting co-channel speech into regions of single speaker and multi speaker segments is an important process in the CCSS system. This stage provides the parameter estimation algorithm with the necessary information about the number of speakers and their voicing states in order to estimate the speech model parameters. In conventional methods of co-channel speech segmentation and classification, the input speech waveform is usually analyzed: 1) using a fixed frame size, 2) once every frame, and 3) at a fixed frame shift. In fact, using a fixed frame size and frame shift for co-channel speech analysis may produce a considerable amount of error, particularly at the voicing transition periods where speech characteristics change rapidly.

In chapter 4, we presented a new technique that is capable of determining the voicing-state of co-channel speech by examining the trajectory parallel measure (TPM) of the state-space embedding of fixed speech frames. The algorithm has shown successful results in classifying co-channel speech into (Voiced/Voiced) V/V and (Voiced/Unvoiced) V/U states. Indeed, the two main limitations of this frame-based technique are:

1. Transitional regions between different voicing states.

2. The high computational complexity and its effect on real-time applications.

Given the higher dynamics of state change in co-channel speech makes it more appropriate to use a running analysis technique (i.e., sample-by-sample computation) rather than segmenting the waveform into fixed-size frames. Eventually, this would

help in enhancing the overall separation process since the speech could be segmented according to speech-dependant characteristics instead of regularly spaced fixed frames.

The frame-based algorithm for voicing-state classification presented earlier also suffers from high computational complexity due to excessive calculations of distances and angles between vectors in the state space. For example, the algorithm required the computation of $N(N-1)/2$ distances for each frame of a size of N samples to determine nearest neighbors for all points in the entire m-dimensional state space. In addition, if we assume that for each point in the state space we have on average a total of L_{NN} nearest neighboring points, then we need roughly NL_{NN} angle calculations to determine the overall TPM for each speech frame. Some of these calculations might be redundant due to the fact that successive frames normally overlap to better track the changes in the speech waveform.

The work presented here tries to overcome these limitations by exploiting a sample-based technique. Figure 6.1(a) shows a schematic diagram of the modified TPM algorithm. First, the co-channel speech signal is low-pass filtered to reduce the effect of high frequencies on the TPM calculation. Typically, a cutoff frequency between 1 kHz and 1.5 kHz is used since the frequency of the first formant of most human speech falls under 1 kHz. This will, in general, preserve a sufficient number of pitch harmonics in single-speaker regions to produce a high TPM value. A very small amount of uncorrelated noise is added to the speech signal prior to processing to avoid false decision during silence periods. The resulting low-pass filtered signal, $x(n)$, is then used to reconstruct the state space of the speech data on a sample-by-sample basis. Based on the method of delays [94] described in Section 3.2.2, m-dimensional vectors, $\mathbf{x}_m(n)$, are formed using the time-delayed samples of the original signal as follows:

$$\mathbf{x}_m(n) = \Big[\; x(n),\;\; x(n+d),\;\; \ldots,\;\; x(n+(m-1)d) \;\Big], \tag{6.2.1}$$

where d is the embedding delay in samples and m is the fixed embedding dimension (number of coordinates). In the frame-based TPM method, the embedding delay

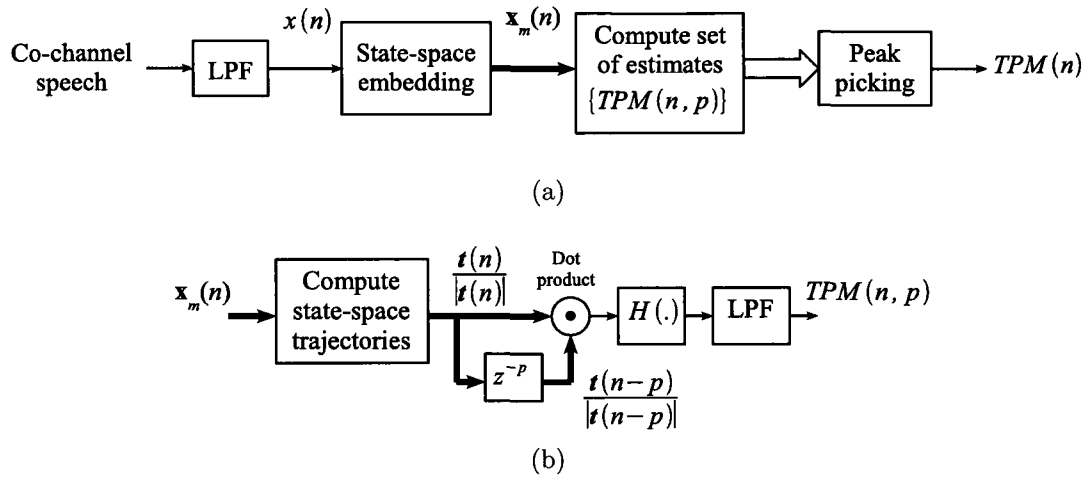


Figure 6.1: The proposed sample-based TPM algorithm: (a) the schematic diagram and (b) the computation of the sample-based TPM at a specific time delay of p samples.

was changed from frame to frame and was calculated using one of the criteria presented in chapter 3. In the sample-based method, however, it is difficult to change the embedding delay during processing and hence a fixed value of d is chosen. Experimental results showed that using a fixed embedding delay of about 1 ms (i.e., 16 samples at 16 kHz sampling rate) has little impact on the accuracy of the calculated TPM.

Any trajectory in the state-space at time index n is defined as

$$\mathbf{t}(n) = \mathbf{x}_m(n+1) - \mathbf{x}_m(n). \tag{6.2.2}$$

The only difference between (6.2.2) and (4.3.2) is that the time index in (6.2.2) extends to the entire speech waveform and is not limited to a single frame. The angle between a trajectory $\mathbf{t}(n)$ and another trajectory p samples later is given by

$$\psi(n, n-p) = \frac{\mathbf{t}(n) \cdot \mathbf{t}(n-p)}{|\mathbf{t}(n)|\,|\mathbf{t}(n-p)|}. \tag{6.2.3}$$

Next, the sample-based TPM at a time lag of p samples is calculated as

$$TPM(n,p) = H(\alpha - \psi(n, n-p)) * g(n), \tag{6.2.4}$$

where $H(\cdot)$ is the Heaviside function, α is a positive threshold close to 1 (e.g., $\alpha = 0.95$), and $g(n)$ is an averaging LPF. The computation of the sample-based TPM at a specific time delay of p samples is shown in Figure 6.1(b). Three different types of smoothing filters were examined as averaging filters: the rectangular moving average filter, the single-pole IIR filter, and the Hamming window averaging filter. The effective length of the filter was chosen to be 10 ms. The best results were obtained by the Hamming-window filter. Finally, the overall TPM at a time index n is calculated as

$$TPM(n) = \max_{P_{\min} \le p \le P_{\max}} [TPM(n,p)], \tag{6.2.5}$$

where $P_{\min}$ and $P_{\max}$ are the minimum and the maximum human pitch periods in samples, respectively.

In order to precisely locate the ending points of both onset and offset regions, the TPM can be calculated in a forward and backward manner. The overall forward-backward TPM is then determined as

$$TPM_{fb}(n) = \max\left(TPM_f(n), TPM_b(n)\right), \tag{6.2.6}$$

where the forward TPM at time n, $TPM_f(n)$, is calculated based on samples indexed earlier in time

$$TPM_f(n) = \max_{P_{\min} \le p \le P_{\max}} [TPM(n,p)], \tag{6.2.7}$$

and the backward TPM at time n, $TPM_b(n)$, is calculated based on samples indexed later in time

$$TPM_b(n) = \max_{P_{\min} \le p \le P_{\max}} [TPM(n,-p)]. \tag{6.2.8}$$

Note that, in the sample-based technique, the calculation of the TPM is not restricted to nearest neighbors as the case in the frame-based algorithm. This yields to enhancing the performance in detecting onset and offset regions as well as reducing the computational complexity of the overall algorithm. Furthermore, unlike the frame-based TPM algorithm presented in chapter 4, increasing the embedding dimension, m, will have less impact on the sample-based TPM calculation if not improving it.

Figure 6.2 shows an example of calculating the TPM for a co-channel speech segment using the sample-based technique discussed above. Comparing this figure with Figure 4.8 it can be observed that the sample-based technique has a smoother TPM with more precise spotting of onset and offset regions.

The proposed algorithm also provides a potential method for estimating pitch frequency when vocalic speech is produced by single speaker. This is important in CCSS since a good portion of the speech is single voiced. Therefore, this algorithm can provide a good initial estimate for the pitch of one of the two speakers in the double voiced regions. The pitch period, $pp(n)$, for the voiced speech of a single speaker can be calculated using the sample-based TPM as follows:

$$pp(n) = \underset{P_{\min} \leq p \leq P_{\max}}{\arg\max} \left[TPM(n,p) \right]. \tag{6.2.9}$$

Figure 6.3 shows a comparison of the calculated TPM of voiced speech for single and co-channel speech signal segments of Figure 6.2. A noticeable large peak is observed at the pitch period and its multiple in the single-speaker case in Figure 6.3(b), while no noticeable peaks are observed in the co-channel speech case in Figure 6.3(d).

6.3 Simplified Method for Sinusoidal Parameter Estimation

The most time-consuming process in the sinusoidal parameter estimation method presented in chapter 5 is the pitch frequency update using the steepest descent algorithm in (5.4.19). The steepest descent algorithm is known for its slow convergence [149]. Moreover, the proposed method involves matrix inversion per each iteration which slows down the processing even further. To solve this problem, we propose in this section a simplified method to estimate the sinusoidal parameters without the use of the steepest descent algorithm. The method takes advantage of the proximity of the initial pitch estimates to the actual pitches to linearize the

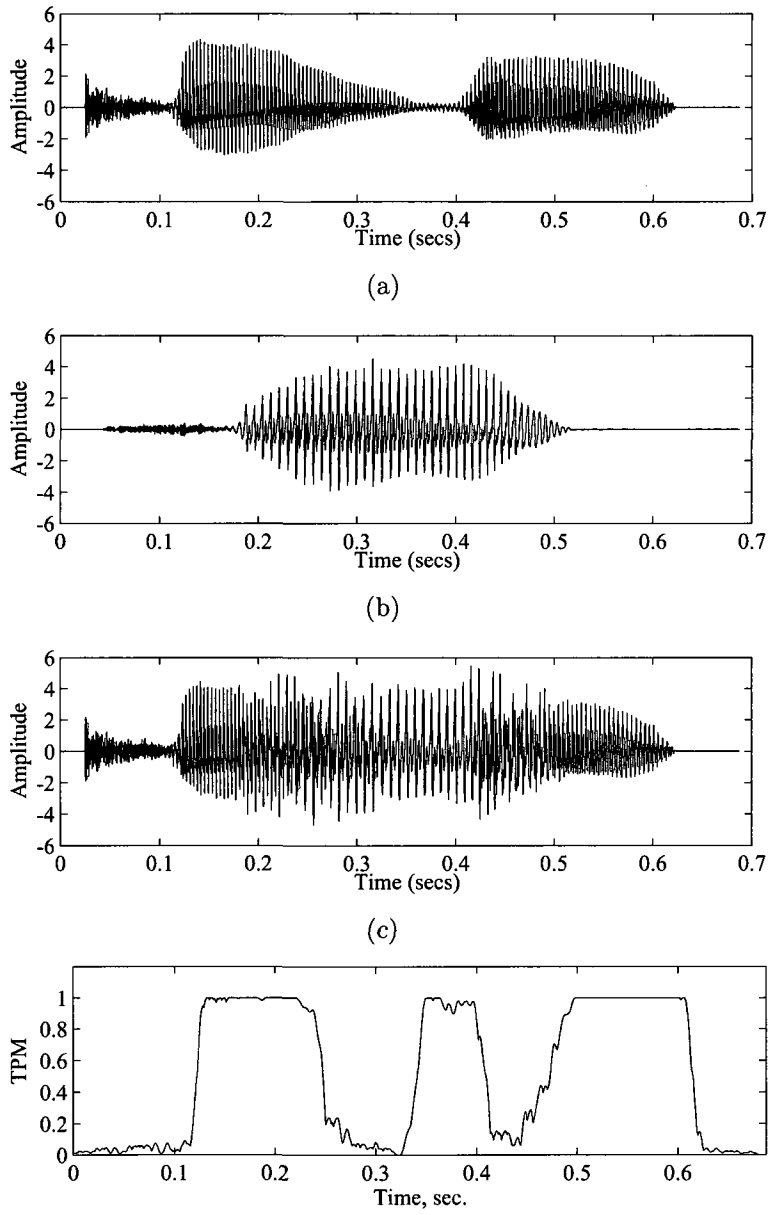


Figure 6.2: The sample-based TPM applied to co-channel speech: (a) first waveform of a female speaker uttering the phrase "toll rate", (b) second waveform of a male speaker uttering the word "she", (c) the mixed co-channel speech, and (d) the TPM plot for the waveform in (c).

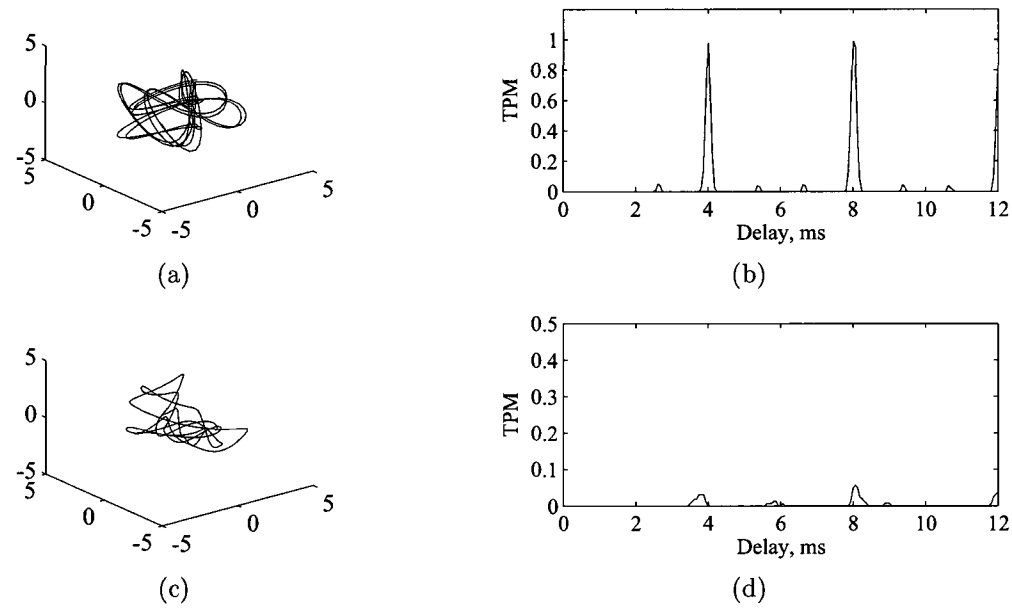


Figure 6.3: State-space visualization of (a) single-speaker voiced speech and (c) co-channel all-voiced speech of the waveforms of Figure 6.2(c) at 0.15 ms and 0.3 ms, respectively. The corresponding sample-based TPMs are shown in (b) and (d).

LS estimation problem and solve it for amplitudes, phases, and frequencies in one iteration by performing matrix inversion once.

According to the sinusoidal modelling approach, a co-channel speech frame is modeled as

$$\hat{x}(n) = \sum_{k=1}^{K} \sum_{\ell=1}^{\hat{L}_k} \left[\hat{a}_\ell^{(k)} \cos(\ell \hat{\omega}^{(k)} n) + \hat{b}_\ell^{(k)} \sin(\ell \hat{\omega}^{(k)} n) \right]. \tag{6.3.1}$$

Under the assumption that the initial estimates of the fundamental frequencies $\omega_0^{(1)}$ and $\omega_0^{(2)}$ of the two speakers are close enough to the actual pitches, final pitch estimates can be represented by

$$\hat{\omega}^{(k)} = \hat{\omega}_0^{(k)} + \Delta\hat{\omega}^{(k)}, \tag{6.3.2}$$

where $\Delta\hat{\omega}^{(k)}$ for $k = 1, 2$ are very small frequency shifts to be estimated. Substituting (6.3.2) into (6.3.1) gives

$$\hat{x}(n) = \sum_{k=1}^{2} \sum_{l=1}^{\hat{L}_k} \left[\hat{a}_l^{(k)} \cos\left(l(\hat{\omega}_0^{(k)} + \Delta\hat{\omega}^{(k)})n \right) + \hat{b}_l^{(k)} \sin\left(l(\hat{\omega}_0^{(k)} + \Delta\hat{\omega}^{(k)})n \right) \right]. \tag{6.3.3}$$

After few calculations

$$\hat{x}(n) = \sum_{k=1}^{2} \sum_{l=1}^{\hat{L}_k} \begin{array}{l} \hat{a}_l^{(k)} \cos\left(l\hat{\omega}_0^{(k)} n\right) + \hat{b}_l^{(k)} \sin\left(l\hat{\omega}_0^{(k)} n\right) \\ +\hat{c}_l^{(k)} ln \cos\left(l\hat{\omega}_0^{(k)} n\right) + \hat{d}_l^{(k)} ln \sin\left(l\hat{\omega}_0^{(k)} n\right), \end{array} \tag{6.3.4}$$

where

$$\hat{c}_l^{(k)} = \hat{b}_l^{(k)} \Delta\hat{\omega}^{(k)} \tag{6.3.5}$$

and

$$\hat{d}_l^{(k)} = \hat{a}_l^{(k)} \Delta\hat{\omega}^{(k)}. \tag{6.3.6}$$

Here, we assumed that

$$\begin{array}{l} \cos(l\Delta\hat{\omega}^{(k)} n) \approx 1 \\ \sin(l\Delta\hat{\omega}^{(k)} n) \approx l\Delta\hat{\omega}^{(k)} n. \end{array} \tag{6.3.7}$$

Equation (6.3.4) now represents a system of linear equations in the unknown parameters $\hat{a}_l^{(k)}$, $\hat{b}_l^{(k)}$, $\Delta\hat{\omega}^{(k)}$. It can be solved using the standard linear LS estimation method in one iteration. To present the problem using matrix and vector notations, let us rewrite (6.3.4) in the form

$$\hat{\mathbf{x}} = \mathbf{Q}\mathbf{h}, \tag{6.3.8}$$

where $\hat{\mathbf{x}}$ is the vector

$$\hat{\mathbf{x}} = \left[\begin{array}{cccc} \hat{x}(0), & \hat{x}(1), & \ldots, & \hat{x}(N-1) \end{array} \right]^T \tag{6.3.9}$$

and $\mathbf{h}$ is defined as

$$\mathbf{h} = \left[\begin{array}{c} \mathbf{h}^{(1)} \\ \mathbf{h}^{(2)} \end{array} \right] \tag{6.3.10}$$

with

$$\mathbf{h}^{(k)} = \left[\begin{array}{c} \mathbf{a}^{(k)} \\ \mathbf{b}^{(k)} \\ \mathbf{c}^{(k)} \\ \mathbf{d}^{(k)} \end{array} \right], \tag{6.3.11}$$

where

$$\mathbf{a}^{(k)} = \left[\begin{array}{cccc} \hat{a}_1^{(k)}, & \hat{a}_2^{(k)}, & \ldots, & \hat{a}_{\hat{L}_k}^{(k)} \end{array} \right]^T, \tag{6.3.12}$$

$$\mathbf{b}^{(k)} = \left[\hat{b}_1^{(k)}, \ \hat{b}_2^{(k)}, \ \ldots, \ \hat{b}_{\hat{L}_k}^{(k)} \right]^T, \tag{6.3.13}$$

$$\mathbf{c}^{(k)} = \left[\hat{c}_1^{(k)}, \ \hat{c}_2^{(k)}, \ \ldots, \ \hat{c}_{\hat{L}_k}^{(k)} \right]^T, \tag{6.3.14}$$

and

$$\mathbf{d}^{(k)} = \left[\hat{d}_1^{(k)}, \ \hat{d}_2^{(k)}, \ \ldots, \ \hat{d}_{\hat{L}_k}^{(k)} \right]^T. \tag{6.3.15}$$

The $\mathbf{Q}$ matrix has the form

$$\mathbf{Q} = \left[\mathbf{Q}^{(1)} \ \ \mathbf{Q}^{(2)} \right], \tag{6.3.16}$$

where

$$\mathbf{Q}^{(k)} = \left[\mathbf{Q}_a^{(k)} \ \ \mathbf{Q}_b^{(k)} \ \ \mathbf{Q}_c^{(k)} \ \ \mathbf{Q}_d^{(k)} \right] \tag{6.3.17}$$

with matrix elements given as

$$Q_a^{(k)}(i,j) = \cos(ij\hat{\omega}^{(k)}), \tag{6.3.18}$$

$$Q_b^{(k)}(i,j) = \sin(ij\hat{\omega}^{(k)}), \tag{6.3.19}$$

$$Q_c^{(k)}(i,j) = ij\cos(ij\hat{\omega}^{(k)}), \tag{6.3.20}$$

and

$$Q_d^{(k)}(i,j) = ij\sin(ij\hat{\omega}^{(k)}). \tag{6.3.21}$$

for $i = 0, 1, \ldots, N-1$ and $j = 1, 2, \ldots, \hat{L}_k$.

Similar to the algorithm presented in chapter 5, if $\hat{L}_k$ and $\hat{\omega}^{(k)}$ are estimated, the goal is to minimize the MSE

$$E = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \mathbf{x}^T\mathbf{x} + \mathbf{h}^T\mathbf{Q}^T\mathbf{Q}\mathbf{h} - 2\mathbf{h}^T\mathbf{Q}^T\mathbf{x}, \tag{6.3.22}$$

with respect to $\mathbf{h}$ (i.e., the parameters $\left\{\hat{a}_l^{(k)}\right\}_{l=1}^{\hat{L}_k}$, $\left\{\hat{b}_l^{(k)}\right\}_{l=1}^{\hat{L}_k}$, $\left\{\hat{c}_l^{(k)}\right\}_{l=1}^{\hat{L}_k}$, and $\left\{\hat{d}_l^{(k)}\right\}_{l=1}^{\hat{L}_k}$), where

$$\mathbf{x} = \left[x(0), \ x(1), \ \ldots, \ x(N-1) \right]^T \tag{6.3.23}$$

represents the original speech frame. The standard linear LS solution to (6.3.22) is calculated as

$$\mathbf{h}_{opt} = \left(\mathbf{Q}^T\mathbf{Q}\right)^{-1}\mathbf{Q}^T\mathbf{x} = \mathbf{R}^{-1}\mathbf{P}, \tag{6.3.24}$$

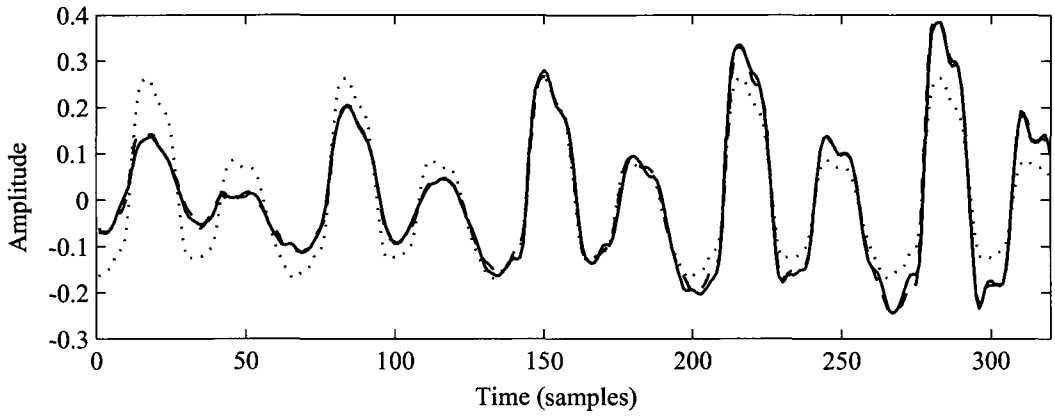


Figure 6.4: Comparison between steepest descent reconstruction (dotted line) and the simplified linear LS reconstruction (dashed line) of a single-speaker waveform (solid line).

where

$$\mathbf{R} = \mathbf{Q}^T\mathbf{Q} \tag{6.3.25}$$

and

$$\mathbf{P} = \mathbf{Q}^T\mathbf{x}. \tag{6.3.26}$$

The minimum MSE corresponding to $\mathbf{h}_{opt}$ is given by substituting (6.3.24) into (6.3.22) to give

$$E_{\min} = \mathbf{x}^T\mathbf{x} - \mathbf{P}^T\mathbf{R}^{-1}\mathbf{P}. \tag{6.3.27}$$

The simplified method not only speeds up the estimation process but also enhances the final reconstructed waveform by adjusting estimated parameters to accommodate slight changes in amplitude (envelope) and frequency along the speech segment. Figure 6.4 shows a comparison between reconstructed waveforms of a single-speaker speech frame using the steepest descent algorithm and the simplified linear LS method. Obviously, the simplified linear LS method adapts for amplitude and frequency changes more than the steepest descent method.

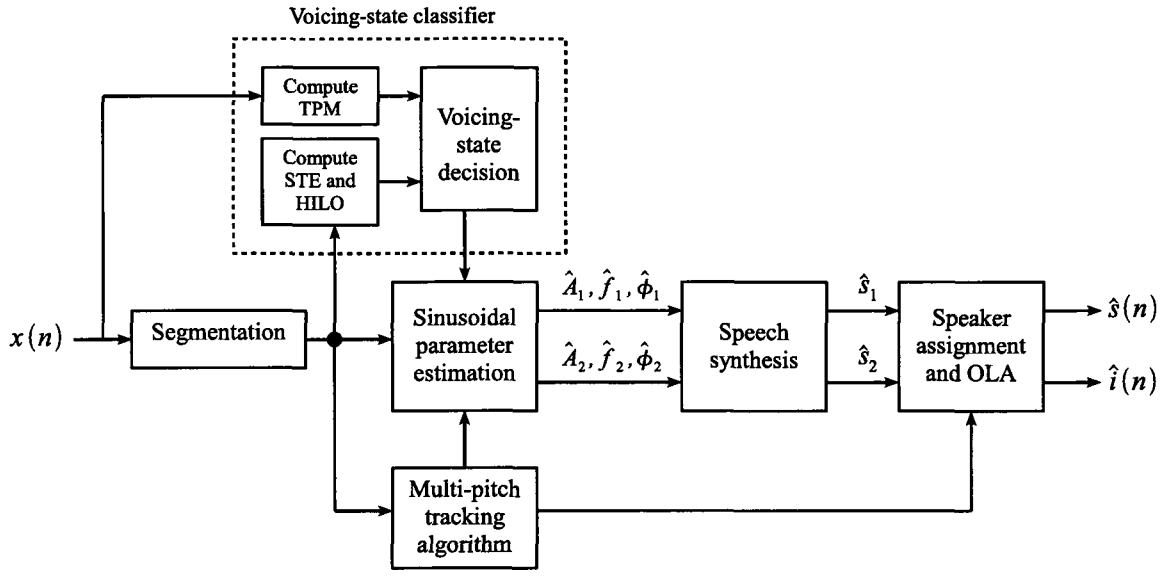


Figure 6.5: Block diagram of the implementation of the CCSS System.

6.4 The Integrated CCSS System

Figure 6.5 depicts the block diagram of the overall CCSS system in which all components are integrated together. All subsystems in the implementation work in a frame-based manner except the TPM algorithm inside the voicing-state classifier which works using a sample-based technique. In this schematic, the frame size and the frame step are not necessary assumed fixed and may vary according to the behavior of the TPM. The separation procedure starts by calculating the average short-time energy (STE) and the high-to-low frequency energy ratio (HILO) of the frame. If STE $>$ STE_{th} and HILO $<$ HILO_{th}, a voiced speech is assumed present and the frame is handled to the next stage. Otherwise, the frame is classified as U/U and no processing is made. Meanwhile, the TPM is calculated using the sample-based technique described in Section 6.2 and is sampled at the middle of each frame. The frame is classified as either single-voiced (V/U or U/V) or double-voiced V/V based on the value of the TPM and a fixed threshold TPM_{th} as shown in the flowchart of Figure 6.6.

The separation strategy is determined according to the voicing-state decision. If TPM > TPM_{th}, this suggests that only one speaker is in the "voiced" state but does not give enough information on how to identify this speaker. This is determined using the pitch information computed by the multi-pitch tracking algorithm. Since multi-pitch tracking is not the scope of this research and will be left as an area of future investigation, we assume that this information (along with pitch information of individual speakers) is known from the clean speech signals before mixing. Therefore, if the target speaker is in the "voiced" state and the interfering speaker is in the "unvoiced" state, the frame is classified as V/U. In this case, the sinusoidal model parameters of only the target speech are estimated and its waveform is reconstructed in the speech synthesis block using these parameters. This is referred to as "enhancement" of the target speech. On the other hand, if the target speaker is in the "unvoiced" state and the interfering speaker is in the "voiced" state, the frame is classified as U/V. Consequently, the waveform of only the interfering speech is estimated and reconstructed in the same way as the previous case. The reconstructed waveform of the interfering speech is then subtracted from the co-channel signal to estimate the target speech waveform. This is called "cancellation" of the interfering speech.

Otherwise, i.e., if TPM < TPM_{th} and both the target and the interferer are in the "voiced" state, the frame is classified as V/V. In this case, both waveforms are estimated and "separated" simultaneously using the sinusoidal model parameter estimation method explained in Section 6.3. For all other cases; i.e., if both individual speakers are in the same voicing state while TPM > TPM_{th}, or if one speaker is in the "voiced" state and the other speaker is in the "unvoiced" state while TPM < TPM_{th}, no processing is made and this case is considered as an error. Occasionally, this might happen, for example, if the voice of one speaker has a pitch value that is approximately an integer multiple of the other speaker's pitch. In this case, the TPM value will indicate a single-voiced frame whereas the actual state is V/V.

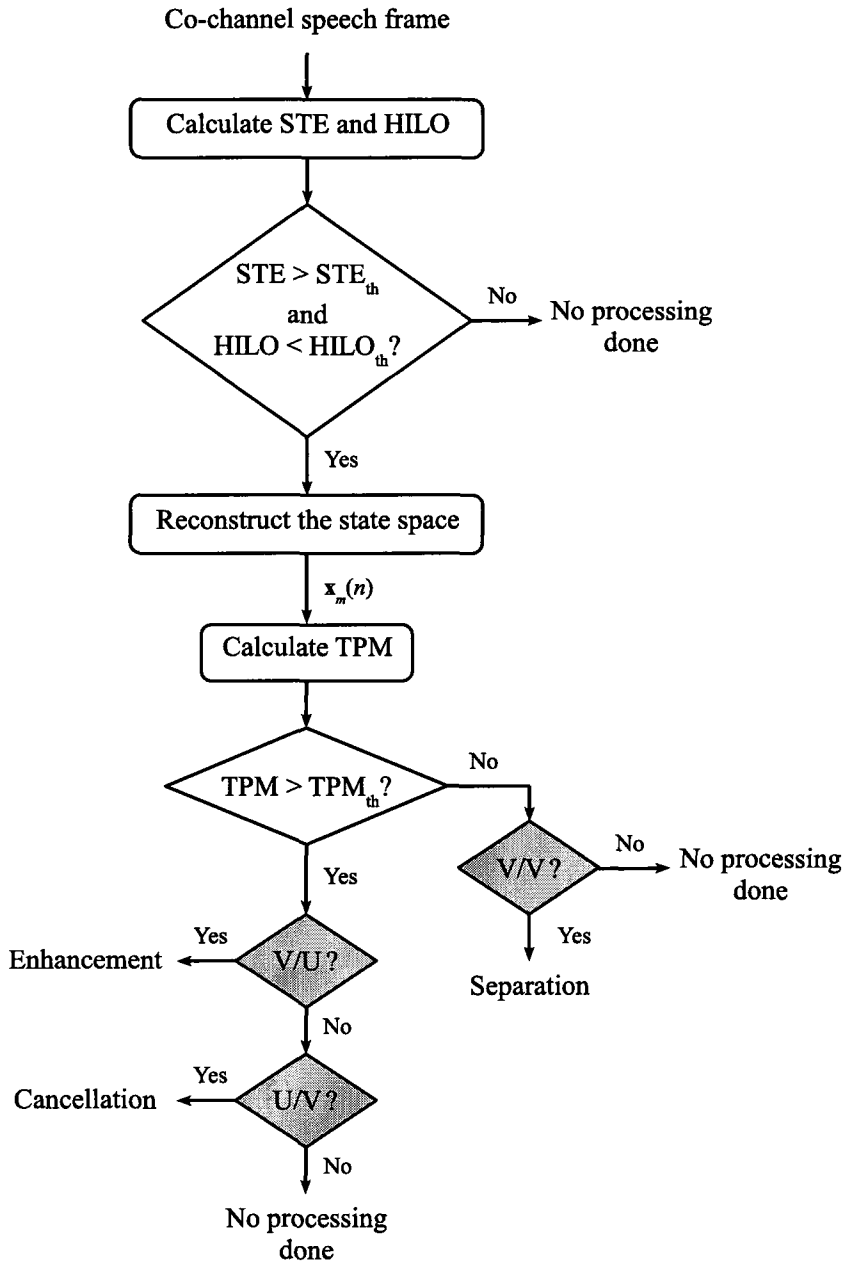


Figure 6.6: Flowchart of the separation system used in the simulations showing processing strategies of each speech frame.

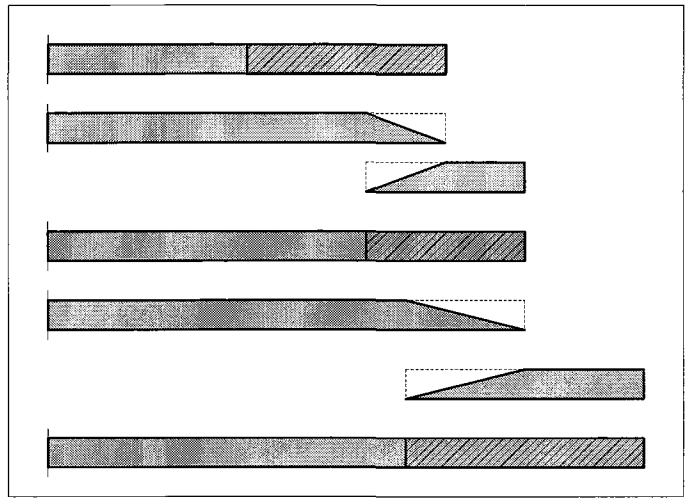


Figure 6.7: Concatenation of two speech frames with different frame sizes and frame steps using the OLA method.

The overall estimated target signal is finally constructed by concatenating synthesized speech frames using the overlap-add (OLA) method. As mentioned earlier, the choice of both frame size and frame step can be based on the information of the sample-based TPM to follow the dynamics of the co-channel speech waveform. This means that we can reduce frame sizes and frame steps during transitional-state periods and increase them during steady-state periods. To do the OLA concatenation considering the variable frame sizes and frame steps, overlapping portions of the signal and the current frame being added are multiplied by a ramp function before addition as shown in Figure 6.7. During all the simulations, a fixed frame size of 20 ms was used while a TPM-based variable frame step was chosen as follows:

$$\text{Frame step} = \begin{cases} 7.5 \text{ ms}, & \text{if } T_1 \leq \text{TPM} \leq T_2; \\ 15 \text{ ms}, & \text{otherwise.} \end{cases} \tag{6.4.1}$$

where $T_1 = 0.1$ and $T_2 = 0.9$ are thresholds optimized to keep an average frame rate of approximately 100 frames/sec. This is equivalent for comparison to a fixed frame size of 20 ms with 50% overlap.

6.5 Summary

This chapter started by describing two modified and simplified versions of the TPM algorithm and the sinusoidal model estimator for co-channel speech. First, an algorithm using the sample-based TPM in determining the voicing state of co-channel speech was presented. The algorithm worked in the state-space domain using the method of delays in a running analysis technique. Then, a simplified method for estimating the sinusoidal model parameters of co-channel speech was presented. The method was based on a linearized version of the LS estimator that allowed estimating model parameters in one iteration and by performing a single matrix inversion. Finally, the overall separation system was implemented by integrating these algorithms together. A variable frame/step size technique based on the value of the sample-based TPM was used to follow the dynamics of the co-channel speech waveform and enhance the overall separation process.

CHAPTER 7

Performance Evaluation

7.1 Introduction

An important stage for the development of any speech separation algorithm is the ability to evaluate its performance by assessing the goodness of the processed speech using measurable features. Various subjective and objective methods have been proposed to asses the output of speech processing algorithms in terms of two main attributes: quality and intelligibility. Although most of these methods have been used originally to evaluate speech coding systems and enhancing systems for speech corrupted with wide-band noise, they can also be used to quantify the performance of CCSS algorithms where the target speaker is interfered by another speaker.

This chapter begins in Section 7.2 with an overview of the quality and intelligibility features of speech. It also highlights some common methods frequently used to evaluate these two features in Sections 7.3 and 7.4. Both subjective and objective measures are then used to evaluate the performance of the proposed CCSS algorithm using some databases of real speech. The results are compared with unprocessed speech mixtures to determine the amount of enhancement achieved by separation. Furthermore, the performance of the proposed technique is compared to other state-of-the-art methods under the conditions of a speech separation challenge. These experimental results are presented and discussed in Section 7.5.

7.2 Quality and Intelligibility of Speech

Quality and intelligibility of speech are different attributes. Therefore, a processed speech requires different assessment methods to measure its quality and its intelligibility [151].

Speech quality is a highly subjective attribute that is usually referred to as the auditory impression (perception) of a human listener upon hearing the speech of other speakers. This impression is influenced by many factors such as loudness, vocal quality, speaking rate, fluency, stress, and intonation. Accordingly, the goal of speech quality measurement is to assess how "good" or "bad" a speech utterance is. This usually can be achieved through the opinion of trained listeners by comparing the test speech signal with a reference utterance of known attributes such as "natural," "raspy," "hoarse," "scratchy," and so on.

Speech intelligibility, on the other hand, refers to the degree to which spoken speech can be understood [151]. In other words, it is concerned about "what" the speaker said. Examples of acoustical factors (among many others depending on the application) affecting intelligibility include: background noise, distance between speaker and listener, reverberation time and level, early reflections, and echo interference. Unlike quality, intelligibility is not a subjective attribute and can somehow be measured using computational algorithms.

Speech can be of poor quality but highly intelligible and vice versa. For example, a speech synthesized using a small number of sine waves or speech produced by a low-quality vocoder can be highly intelligible, yet it may sound very machine-like and speaker-unidentifiable. On the contrary, a speech signal transmitted over IP networks (VoIP) with a large amount of packet loss and delay might have good quality but will be perceived as intermittent and perhaps unintelligible.

Which is more important, speech quality or intelligibility? This totally depends on the application. For example, commercial communication systems and military

applications (that may operate under adverse noise conditions and bandwidth constraints) are some areas where speech intelligibility may be considered a more important feature than the more general quality. Speech quality is important when we are dealing with multimedia applications for example.

Evaluation of quality or intelligibility of separated speech can be done using either subjective listening tests (human opinion) or objective assessment (machine-based measures). Similarly, intelligibility can be evaluated by presenting processed speech material (words, sentences, etc.) to a group of listeners or to an ASR system and counting the number of words identified correctly.

Formal subjective evaluation is usually more accurate and preferable. However, a reliable listening test requires a large number of trained listeners, and is therefore slow and expensive to conduct. Hence, objective computational measures have been more frequently used in the quality evaluation of CCSS algorithms compared to formal listening tests [66, 152]. Objective methods have the advantage of providing reproducible results and being automated. Thus, they are more economical. The main challenge to objective methods in performance evaluation techniques using speech recognition systems is the difficulty of creating reliable methods to cope easily with the complex processes done by humans in speech understanding.

7.3 Evaluating Intelligibility of Processed Co-channel Speech

7.3.1 Human listening tests

Perhaps the most reliable and accurate method to evaluate speech intelligibility is through the opinion of human listeners. This can be done by presenting the corrupted speech material (e.g., none-sense words or sentences) to a group of listeners and quantifying the intelligibility of speech in terms of percentage of words identified correctly. In CCSS systems, this type of tests is particularly utilized when speech is embedded in very low TIR levels. Test cases with TIR above 0 dB are trivial

separation tasks for human listeners with normal hearing ability [11].

Quatieri and Danisewicz [30] used vocalic sentences only to evaluate the performance of their frequency-domain sinusoidal modelling approach. They used a limited database of six sentences uttered by three males and three females. The input TIR of co-channel speech ranged from -16 to 9 dB. They reported more effective enhancement in the target speech with *a priori* information of pitch contours compared to the case with no *a priori* information.

Morgan *et al.* [11] tested their harmonic enhancement and suppression (HES) algorithm by asking ten untrained listeners to transcribe a target speech jammed by a stronger interferer at TIR of -6, -12, and -18 dB. They used these transcriptions to determine the performance difference between processed and unprocessed co-channel speech. They also tested co-channel signals with additive Gaussian noise at 10 dB and 15 dB. The co-channel speech data consisted of randomly selected and linearly added sentences of male speakers from the TIMIT database. By comparing word recognition accuracy of unprocessed and processed speech, results showed an increase from 67.5% to 88.7% at -6 dB TIR, a decrease from 62.5% to 42.5% at -12 dB TIR, and an increase from 4.6% to 12.3% at -18 dB TIR. Speaker assignment was assumed to be known in *a priori* when implementing the HES algorithm.

7.3.2 Automatic speech recognition (ASR) tests

Speech separation is not a target by itself, but a preprocessing means to another subsequent application such as ASR. Therefore, the performance of the overall application can be used to evaluate the accuracy of the speech separation stage. This is usually refereed to as an application-oriented test. ASR measures the separation success by comparing word error rate (WER) or word recognition accuracy (WRA) of the recognized speech with and without separation. WER is computed as [153]:

$$\text{WER\%} = 100 \times \frac{N_{\text{S}} + N_{\text{D}} + N_{\text{I}}}{N_{\text{T}}}, \tag{7.3.1}$$

where N_S is the number of substituted words, N_D is the number of deleted words, N_I is the number of inserted words, and N_T is the total number of words in the reference data. WRA is computed as

$$\text{WRA\%} = 100\% - \text{WER\%}. \tag{7.3.2}$$

It is worth noting that using CCSS systems to enhance speech recognition requires a careful combination between the separation algorithm and the recognition engine. A speech separation system designed to improve SNR or human intelligibility would not necessarily maximize speech recognition performance [154]. A mismatch between the separation algorithm and the speech recognizer may lead to an error rate using separated speech that is worse than when the unprocessed mixture is used.

Morgan *et al.* (1997) [11] used a keyword spotting (KWS) test to evaluate the performance of their HES algorithms. They used a vocabulary set of seven multisyllable keywords and their variants as the target speech and random utterances selected from the MIT-CBG database as the interfering speech. The two signals was linearly added and the goal of the recognizer was to correctly spot the keywords. The HES system was tested at 18, 12, and 6 dB TIRs. Performance was evaluated in terms of a figure of merit (FOM) for each keyword. Overall results showed an enhanced performance using the HES technique compared to unprocessed speech.

Recent examples of using ASR systems as a means to evaluate the performance of CCSS algorithms is the tests proposed in [62,63,66,74,155]. Their speech separation systems were evaluated using the Grid database from the 2006 speech separation challenge proposed in [156]. This database is composed of 500 sentences spoken by each of 34 different speakers, giving a total corpus size of 17000 sentences. Each sentence follows the structure: "<command:4> <color:4> <preposition:4> <letter:25> <digit:10> <adverb:4>," where the number indicates the number of choices for the associated keyword. Co-channel speech data were composed by mixing pairs of these sentences at a range of TIR varying from -9 dB to 6 dB. According to the challenge, the recognition task is to recognize the letter and the digit of the

speaker who said the color keyword "white". The ASR system is first trained using the clean speech data and then tested using the co-channel speech data. Separation accuracy is evaluated by comparing recognition results before and after separation.

7.4 Evaluating Quality of Processed Co-Channel Speech

7.4.1 Subjective quality tests

Subjective evaluation of speech quality involves a group of listeners to compare the separated speech with a reference source and grade the final quality using a predetermined scale of 5-10 levels. Similar to evaluating speech intelligibility, subjective listening tests provides the most reliable method for evaluating speech quality. These types of tests, however, can be time consuming and require, in most cases, trained listeners.

One of the widely used subjective quality measures is the mean opinion score (MOS) recommended by the IEEE and ITU [157, 158]. In this method, a number of experienced and non-experienced listeners rate the quality of the speech signal using a score from 1 (lowest perceived quality) to 5 (highest perceived quality). The overall measured quality of the signal is obtained by averaging the scores indicated by all listeners.

Li et al. (2006) [67] conducted a speech quality test for their CASA-based separation algorithm using the MOS with 10 listeners. Their results for co-channel speech separation showed an average MOS improvement from 1.27 for the unprocessed mixed speech to 2.27 for the processed speech when the pitch of the target speaker was estimated. When the true pitch was used, the MOS increased to 2.81. A MOS of 3.2 was obtained when an ideal binary mask was used to eliminate the interfering speech.

7.4.2 Objective quality measures

Objective evaluation of enhanced speech quantifies the quality by measuring a numerical distance between the original (uncorrupted) and processed (separated) signals. For an objective measure to be reliable, it needs to have a high correlation with subjective evaluation tests. Ideally, objective measures should be able to assess the quality of separated co-channel speech without the need of having access to the isolated pre-mixture sources. In practice, however, current objective measures are limited such that most of them require the original signals before mixing to be available.

In the following part of this section, we shortly discuss some objective quality measures that are commonly used to evaluate the performance of speech coding and speech enhancement systems for speech corrupted with wide-band noise. However, they can also be used to evaluate the performance of the application in hand.

1. **Signal-to-noise ratio (SNR)**

   Signal-to-noise ratio (SNR) is the simplest objective measure that can be used to summarize speech quality. It is defined as the ratio of the energies of the reference speech signal and the error between the reference and separated signal. The overall SNR in dB can be calculated in time domain as:

$$\text{SNR [dB]} = 10 \log_{10} \left(\frac{\sum_n s(n)^2}{\sum_n [s(n) - \hat{s}(n)]^2} \right), \tag{7.4.1}$$

   where $s(n)$ is the original target signal before mixing and $\hat{s}(n)$ is the separated signal. It is sometimes also referred to as signal-to-distortion ratio (SDR) or signal-to-residual ratio (SRR). Segmental SNR (SNRseg) [159] calculated based on short-time frames was also used to evaluate the performance of speech separation algorithms [52,160]. Unlike the conventional SNR, SNRseg takes into account the fact that errors in low-intensity segments are usually more easily

perceived. SNRseg is defined as:

$$\text{SNRseg [dB]} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\left(\frac{\sum_{n=N_m}^{N_m+N-1} s(n)^2}{\sum_{n=N_m}^{N_m+N-1}[s(n)-\hat{s}(n)]^2}\right), \tag{7.4.2}$$

where M is the total number of frames and N is the length of each frame. Since the intervals of silence have very small energy, silent frames should be excluded from the sum in (7.4.2) to avoid the effect of large negative SNRseg values at those frames.

In the area of CCSS, many authors [59, 74, 160–162] reported an average improvement in the SNR in the range of 3-10 dB using different separation techniques. Quackenbush *et al.* [163] reported a correlation of 0.24 between this measure and the subjective quality scores done by human listeners under a variety of degradations. Despite this weak correlation, SNR remains widely used, due possibly to its simplicity.

In fact, SNR has the following drawbacks that limit its use [161]:

- It requires that the separated waveform to be near perfect.
- It requires that the original pre-mixture signal to be available for comparison.
- Distortions such as fixed phase/time delays or nonuniform gains across frequency which can have only a small effect on the perceived quality of a reconstructed speech, can have a large negative effect on SNR.
- Energy has in general only an indirect relationship to perceived quality. The same amount of energy will have a widely-varying impact on perceived quality depending on where and how it is placed in time-frequency; this is particularly significant in the case of speech, where most of the energy is below 500 Hz, yet very little intelligibility is lost when this energy is filtered out.

2. Spectral distance measures

This group includes distance measures calculated in the frequency domain or in terms of second order statistics. Examples of these measures include: the log-likelihood ratio (LLR), the Itakura-Saito (IS) measure, and Cepstral (CEP) distance [151].

The LLR measure (also called the Itakura distance) is calculated for each speech frame as follows [164]:

$$d_{\text{LLR}} = \log\left(\frac{\mathbf{a}_{\hat{x}}^T \mathbf{R}_x \mathbf{a}_{\hat{x}}}{\mathbf{a}_x^T \mathbf{R}_x \mathbf{a}_x}\right), \tag{7.4.3}$$

where $\mathbf{a}_x$ and $\mathbf{a}_{\hat{x}}$ are the LPC coefficient vectors of the clean and processed speech frames, respectively, and $\mathbf{R}_x$ represents the Toeplitz autocorrelation matrix of the clean speech frame. LLR measure can be interpreted as the logarithmic ratio between the energies of the prediction residuals of the processed and clean signals. Since the denominator in (7.4.3) represents the minimum possible residual energy, it is always smaller than the numerator. Hence, LLR measure is always positive.

In the frequency domain, (7.4.3) can be expressed as

$$d_{\text{LLR}} = \log\left(1 + \int_{-\pi}^{\pi} \left|\frac{A_x(\omega) - A_{\hat{x}}(\omega)}{A_x(\omega)}\right|^2 d\omega\right), \tag{7.4.4}$$

where $A_x(\omega)$ and $A_{\hat{x}}(\omega)$ are the Fourier transforms of $\mathbf{a}_x$ and $\mathbf{a}_{\hat{x}}$, respectively. They represent the envelopes of signal spectra that are characterized by speech formants. This means that LLR measures speech quality through the difference between spectrum envelops of the original and processed signals which is more pronounced near formants.

The main advantage of the LLR distance is that it is not sensitive to gain change nor time delay as the SNR for example. On the other hand, its main disadvantage is that it penalizes differences in short-time correlation (formants) while ignores differences in long-time correlation (pitch). This is a serious

drawback when compared to subjective measures as the human ear is very sensitive to pitch differences.

The IS distance is very similar to the LLR measure and is calculated for each speech frame as follows:

$$d_{\text{IS}} = \frac{G_{\hat{x}}}{G_x} \frac{\mathbf{a}_{\hat{x}}^T \mathbf{R}_x \mathbf{a}_{\hat{x}}}{\mathbf{a}_x^T \mathbf{R}_x \mathbf{a}_x} + \log\left(\frac{G_x}{G_{\hat{x}}}\right) - 1, \tag{7.4.5}$$

where G_x and $G_{\hat{x}}$ are the all-pole filter gains of the clean and processed frames, respectively. Unlike the LLR measure, the IS measure penalizes differences in overall spectral levels. However, this is considered a drawback as psychoacoustic studies [151] have shown that differences in spectral level have minimal effect on speech quality.

The CEP distance measure is defined as the distance between the cepstral coefficients of the clean and processed signals as follows:

$$d_{\text{CEP}} = \frac{10}{\log_e 10} \sqrt{2 \sum_{k=1}^{p} [c_x(k) - c_{\hat{x}}(k)]^2}, \tag{7.4.6}$$

where $c_x(k)$ and $c_{\hat{x}}(k)$ are the cepstrum coefficients of the clean and processed signals, respectively, and p is the order of the LPC analysis.

3. Perceptual measures

The above-mentioned objective measures are attractive to many researchers as they are simple to implement and easy to evaluate. However, they do not closely mimic the way in which human auditory system perceive speech quality. Perceptual measures, on the other hand, try to estimate the audibility of the speech distortions by calculating the difference between the reference (clean) and enhanced signal using an auditory model [165]. A common preprocessing stage in all perceptual methods is filtering the input signal using an auditory filter bank (a bank of band-pass filters known as critical-band filters with center frequencies and bandwidths increasing with frequency). This filter-bank

model is used to simulate the auditory filter of human inner ear. The critical-band frequency spacing can be approximated by bark frequency scale using a nonlinear Hertz-to-bark transformation.

The weighted slope spectral (WSS) distance [166] is an example of a perceptual measure. It evaluates speech quality based on weighted differences between the spectral slopes of the clean and processed signal in each critical frequency band.

The bark spectral distortion (BSD) measure [167] and its enhanced version [168] are also known to give a high correlation with MOS. They use the distances between loudness bark spectra to measure perceptually-significant auditory attributes. A value of zero for the BSD indicates no distortion, while a higher value indicates increasing distortion.

Another important measure that is widely used under this category is the perceptual speech quality measure (PSQM) [169] and its modified version; the perceptual evaluation of speech quality (PESQ) [170]. PESQ compares two perceptually-transformed signals and generates a value that mimics the MOS to estimate the perceived speech quality. The PESQ has been shown to have good accuracy in the following factors [171]: speech input levels to a codec, transmission channel errors, packet loss and packet loss concealment, environmental noise at the sending side, and effect of varying delays. The range of the PESQ score is 0 to 4.5.

Most perceptually-motivated techniques for speech quality measurement were originally developed to evaluate the quality of coded speech [166, 167, 169] or speech transmitted over IP networks (VoIP) [170]. They have usually been optimized for applications where signal distortion is mainly caused by quantization error and varying delays. Therefore, they may produce misleading results when applied to separated speech signals, and their use for the evaluation of CCSS systems deserves investigation.

Objective measure	Overall quality	Signal distortion	Background noise
SNRseg	0.31	0.19	0.42
LLR	0.63	0.66	0.26
IS	0.45	0.58	0.06
CEP	0.6	0.65	0.22
WSS	0.53	0.50	0.37
PESQ	0.65	0.57	0.48

Table 7.1: Estimated correlation coefficients between the subjective quality measure and some objective quality measures with overall quality, signal distortion, and background noise distortion according to [151].

Table 7.1 summarizes the estimated correlation coefficients between the subjective quality measure and some of the objective quality measures mentioned above according to [151]. The table shows the correlation coefficients for overall quality, signal distortion and background noise distortion.

7.5 Simulation Results and Discussion

This section presents the simulation results obtained in evaluating the performance of the proposed CCSS system shown in Figure 6.5. The separation system was evaluated using different subjective and objective methods to measure both the intelligibility and the quality of the processed speech. The results were compared to unprocessed co-channel speech to determine the amount of enhancement or degradation. For measuring speech intelligibility, subjective listening and ASR tests were used. For measuring speech quality, SNR, Itakura-Saito, and PESQ measures were used.

The general purpose of the first group of simulations was not to compare the performance of the proposed algorithm with other existing techniques. It aimed rather to figure out the trend and capability of the proposed algorithm in separating co-channel speech under different TIR conditions. At the end of this section, the performance of the separation system is evaluated under the conditions of the speech separation challenge in order to be compared with the performance of other recently developed systems. Throughout the experiments, speaker voice activity detection

(VAD) and speaker assignment were assumed to be known in *a priori*. These information were determined from speech waveforms before mixing.

7.5.1 Test data

Up to the author knowledge, there is no standard database created specifically for testing co-channel speech separation systems. Such a database needs to have speech recordings in a real co-channel environment where two (or more) speakers speak simultaneously at a certain distance from a single microphone. In addition, the database should contain another set of speech recordings of clean speech for the same speakers. This second set can be used, for example, in a training phase, if needed. For this reason as well as to have access to the original speech waveforms for comparison, test data used in our simulations were created using two single-speaker speech databases: the TIMIT database and the TIDIGITS database. These two databases were chosen because they are extensively used and their results are widely accepted in many speech processing researches.

The TIMIT database [144] is a speech corpus consists of 6,300 files of approximately two to four seconds of digitally recorded speech at a sampling rate of 16 kHz. Ten sentences are spoken by each of 630 male and female speakers from eight major dialect regions of the United States. All simulations in this thesis were performed using speech files from the first dialect region (New England) of the TIMIT corpus.

The TIDIGITS database [172] contains read utterances from 326 male and female speakers of different ages. Each speaker is uttering the digit sequence 0 to 9 plus the letter "O". The data were collected in a quiet environment and digitized at 20 kHz. To have consistency with the TIMIT database, speech files created from the TIDIGITS database were down-sampled to 16 kHz.

To establish a reliable testing data set of co-channel speech, target speech files were created by concatenating three digits from the TIDIGITS database to form 3 second waveforms. The interfering speech files, on the other hand, were created by

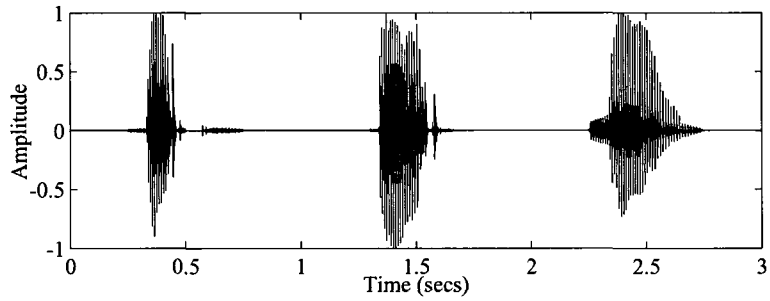
truncating speech files randomly selected from the TIMIT database at 3 seconds. These two sets of files were scaled and linearly added using the criteria explained in the next section to construct co-channel testing files. Figure 7.1 shows a co-channel speech test sample consisting of two signals of a male speaker as the target speech and a female speaker as the interfering speech. A total of 420 co-channel speech files were created in the same way using 30 sentences (15 female and 15 male) from the TIMIT database and 14 connected digits (7 female and 7 male) from the TIDIGITS database. This collection of data was found to cover a wide range of variations in the four categories of co-channel speech (female/female, female/male, male/female, and male/male). The same selected files were mixed at different TIRs from -20 dB to 20 dB with a 5 dB step.

7.5.2 Input TIR

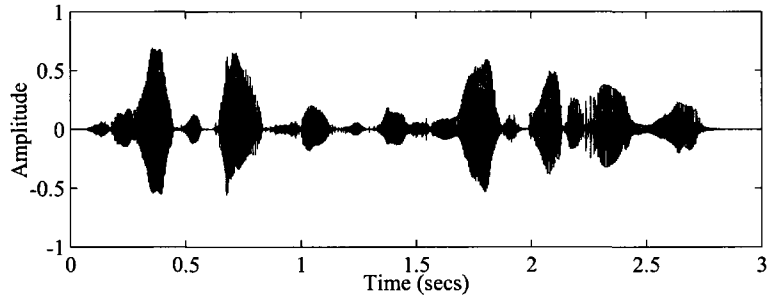
Each co-channel speech signal used in the following tests were created by scaling and linearly adding two speech waveforms: one for the target speech (selected from the TIMIT database) and the second for the interfering speech (selected from the TIDIGITS database). Since both the target speech and the interfering speech are non-stationary signals and have many periods of silence, the input TIR[1] was calculated based only on periods where both speech signals were active (i.e., overlapping portions of speech). In other words, for each speaker, periods of silence were excluded when calculating signal energy. For this purpose, a simple voice activity detection (VAD) algorithm based on speech energy was used to separate regions in which speech is present from silence regions. Only speech segments where both speakers were active (either voiced or unvoiced) were processed.

To obtain a co-channel input signal with a desired TIR, the target speech waveform, $s(n)$, and the interfering speech waveform, $i(n)$ were first segmented into 20 ms frames with 50% overlap and the energy per frame for each signal was calculated

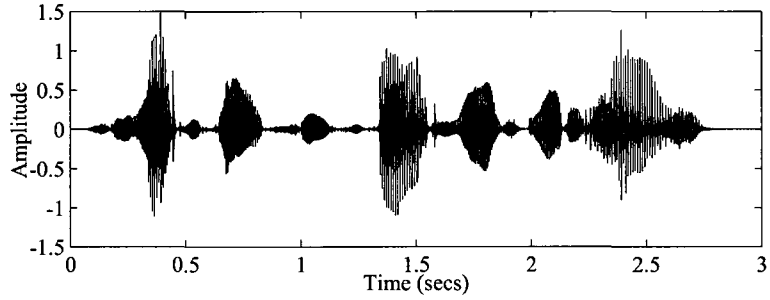
[1]In applications where both signals in the mixture are of the same interest to separate, this ratio might be referred to as signal-to-signal ratio (SSR).



(a)



(b)



(c)

Figure 7.1: Example of a constructed co-channel speech test sample: (a) the waveform of a target speech taken from the TIDIGITS database of the connected digits "six-four-nine" spoken by a male speaker, (b) the waveform of an interfering speech taken from the TIMIT data base for a female speaker saying the sentence "She had your dark suit in greasy wash water all year," and (c) the co-channel test signal resulting from adding (a) and (b) at 0 dB TIR.

as follows:

$$E_s(m) = \sum_{n=N_m}^{N_m+N-1} s(n)^2 \tag{7.5.1}$$

and

$$E_i(m) = \sum_{n=N_m}^{N_m+N-1} i(n)^2, \tag{7.5.2}$$

where m is the frame index, N_m is the index of the first sample in the current frame, and N is the frame length. Both signals were then normalized to have an average energy of unity over overlapping frames. This is done to make a plateau for the sound level of the signal with higher energy.

The TIR at this step was denoted as TIR_0 and was calculated as

$$\text{TIR}_0\ [\text{dB}] = \frac{10}{M'} \sum_{m=0}^{M-1} \log_{10}\left(\frac{E_s(m)}{E_i(m)} \cdot \text{VAD}_{ol}(m)\right), \tag{7.5.3}$$

where M is the total number of frames, M' is the total number of non-silent overlapping frames, and $\text{VAD}_{ol}(m)$ is 1 if both speakers in the m-th frame are active and 0 otherwise. To determine the value of $\text{VAD}_{ol}(m)$, segmental voice activity detections; $\text{VAD}_s(m)$ and $\text{VAD}_i(m)$ for the target and interference signals, respectively, were calculated prior to mixing using an energy threshold. $\text{VAD}_{ol}(m)$ was then calculated as follows:

$$\text{VAD}_{ol}(m) = \begin{cases} 1, & \text{if } \text{VAD}_s(m) = \text{VAD}_i(m) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{7.5.4}$$

Finally, the two speech signals were scaled as follows:

- if $\text{TIR} \geq 0$:

$$\begin{aligned} s(n) &\leftarrow s(n) \\ i(n) &\leftarrow i(n) \cdot \sqrt{\frac{\text{TIR}_0}{\text{TIR}}} \end{aligned} \tag{7.5.5}$$

- if $\text{TIR} < 0$:

$$\begin{aligned} s(n) &\leftarrow s(n) \cdot \sqrt{\frac{\text{TIR}}{\text{TIR}_0}} \\ i(n) &\leftarrow i(n) \end{aligned} \tag{7.5.6}$$

7.5.3 Intelligibility measuring tests

The first group of simulations examined the performance of the proposed CCSS algorithm in terms of intelligibility. Two types of tests were conducted under this category: subjective listening test and ASR test. Results obtained when these two tests were applied to the unprocessed speech and the processed speech are presented and discussed in this section.

1. Subjective listening test

The goal of this test was to evaluate the performance of the separation system based on human speech recognition. To perform the listening test, the 420 co-channel speech mixtures in our database were played (one at a time) to two male and two female listeners with normal hearing through a set of headphones. The tests were conducted in isolation, over a single session, in a quiet room. Listeners were asked upon hearing each speech mixture to simply identify the three digits spoken by the target speaker. Each listener was given a numeric keypad to key in the number that he/she thinks it is as close as possible to the spoken number. Since separation of co-channel speech signals with a TIR equal to or greater than 0 dB is considered a trivial task for humans with normal hearing, speech simulations for this test were conducted only at TIR values from -20 dB to -5 dB.

In order to have a baseline for comparison, each listener repeated the test using the same data files twice; one time for the unprocessed speech and one time for the processed speech. Tables 7.2 and 7.3 show the confusion matrices at TIR of -20 dB for the unprocessed data and the processed data, respectively. For each uttered number, percentage of the number most confused with it by the listeners is shown in bold. As illustrated in the tables, in general, the percentage of recognizing each number was increased in the processed data compared to the unprocessed data. This in turn was interpreted as a decrease and sometimes change of the location of most confusing numbers. For example, the number "7" was most confused in the unprocessed data

Uttered number \ Recognized number	0	1	2	3	4	5	6	7	8	9
0	93.7%	0.4%	**3.9%**	0%	0%	0%	0.8%	0.4%	0%	0.8%
1	0%	95.1%	0.9%	0%	0.9%	0.4%	0%	0%	0%	**2.7%**
2	1.6%	0.4%	91%	**3.9%**	0.4%	0%	2.7%	0%	0%	0%
3	2.8%	0%	4.6%	87%	0%	0%	0.5%	0%	**5.1%**	0%
4	**1.6%**	0%	1.2%	0%	94.8%	0.8%	0.8%	0.4%	0.4%	0%
5	0%	0%	0%	0%	1.1%	93.5%	0%	0%	0%	**5.4%**
6	0.6%	0%	**1.5%**	0%	0.3%	0%	96.4%	0%	1.2%	0%
7	0%	1.1%	0.4%	0.7%	0%	0.4%	**3.2%**	92.4%	1.1%	0.7%
8	0%	0.4%	1.8%	**3.6%**	0.4%	0%	3.6%	0%	90.2%	0%
9	0%	1.7%	0%	0%	0%	**5.4%**	0%	0%	0.3%	92.6%

Table 7.2: Confusion matrix for the subjective listening test of unprocessed speech at TIR = −20 dB. Percentages of most confusing numbers are shown in bold.

Uttered number \ Recognized number	0	1	2	3	4	5	6	7	8	9
0	96.5%	0%	**1.5%**	0%	1.2%	0%	0.4%	0.4%	0%	0%
1	0%	93.8%	0%	1.3%	0.9%	0.9%	0%	0%	0%	**3.1%**
2	1.2%	0.4%	90.6%	**4.2%**	0.8%	0%	1.2%	1.2%	0.4%	0%
3	2.3%	0%	1.4%	92.6%	0%	0%	0%	0%	**3.2%**	0.5%
4	0.4%	0.4%	0.4%	0.4%	96.1%	**2.3%**	0%	0%	0%	0%
5	0%	0%	0%	0%	0%	96.7%	0%	0.6%	0%	**2.7%**
6	0%	0.9%	**0.9%**	0%	0.6%	0%	96.6%	0.6%	0.4%	0%
7	0.4%	**1.4%**	0%	0.4%	0.4%	0.4%	1%	95%	0%	1%
8	0.9%	0.4%	0.9%	**4.9%**	2.2%	0.9%	4%	0.9%	84.5%	0.4%
9	0%	1%	0%	0%	0%	**3.4%**	0%	0%	0%	95.6%

Table 7.3: Confusion matrix for the subjective listening test of processed speech at TIR = −20 dB. Percentages of most confusing numbers are shown in bold.

with the number "6" with a percentage of 3.2%. However, the same number became most confused with the number "1" with a percentage of 1.4% in the processed data. This indicates that while the system tries to enhance the accuracy of word recognition, errors due to distortion may occur.

The final performance comparison of the recognition accuracy between unprocessed and processed data is shown in Figures 7.2(a), 7.2(b), 7.2(c), and 7.2(d) for female/female, female/male, male/female, and male/male mixtures, respectively, in terms of WER. The overall performance is shown in Figure 7.2(e). Due to the limited vocabulary of test data which were also known to the listeners in advance, the WER was calculated using (7.3.1) without considering any word deletion or insertion. In

the figures below, the WERs of incorrectly identified digits for the unprocessed speech mixtures are shown in black-colored bars while the WERs for the processed speech data are shown in gray-colored bars. The later are further divided into two segments that will be discussed shortly. Although the listening test results did not show a significant improvement (if not degradation) in the intelligibility of the processed speech, there are some interesting conclusions that can be drawn from these results.

First, the male/female case (i.e., when the target speaker was a male and the interfering speaker was a female) indicated the lowest WER (or consequently the highest WRA) for both unprocessed and processed data. There were almost no improvement made in this case. Conversely, the female/male and the male/male mixtures showed the highest WER results with some improvement in the processed data especially at low TIRs. Results for the female/female case were somewhere in between for both the accuracy and the improvement. The low WER for the female/male case was intuitively unexpected since one would expect better segregation with speakers of different gender. In fact, this most probably is due to the so called *frequency masking* phenomenon. During this phenomenon, human ears cannot hear a weak (soft) signal that is located (in the frequency domain) nearby a strong (loud) signal [5]. In this case, we say that the stronger signal have masked the weaker signal. Similarly, when a speech signal is interfered simultaneously by a second competing speech signal, the first signal might be totally or partially masked if all or part of its frequency harmonics fall below the masking threshold of the second signal. This justifies why a female target speaker can be masked more easily by an interfering male speaker and not vice versa. In the case of a male/female mixture, the high pitch of the interfering female speech leaves wider gaps between harmonic components in the frequency domain. This results in lowering the masking threshold level between these harmonics and allowing the frequency components of the target male speech to be heard. In a female/male mixture, on the other hand, the low pitch of the interfering speech causes a relatively high masking threshold level between the

dense harmonics, resulting in stronger masking to the target female speech.

The second interesting observation drawn from the results of Figure 7.2 was disclosed when the errors due to incorrectly identified digits for both the unprocessed and the processed data were examined together closely. Since all the errors in the unprocessed mixtures are due to the masking of the interfering speech to the target speech, this type of error is referred to as the *masking error* for the unprocessed data. The percentage of incorrectly identified digits in the unprocessed data that remains unrecognizable after processing the speech mixtures was identified and drawn in dark gray bars in Figure 7.2. This type of error is referred to as the masking error for the processed data. The remaining percentage of word errors in the processed data was mainly due to distortion caused by the separation algorithm itself. Therefore, this type of error is called the *distortion error*. If we compare the masking error in both cases, we can conclude that the algorithm was able to unmask a good portion of the target speech. However, at the same time, it added another type of error due to distortion. Sources of distortion error may include:

- Computational error in the matrix inversion operation of the sinusoidal parameter estimation algorithm.
- Pitch detection errors such as pitch halving and doubling.
- Errors in TPM calculation due to Pitch crossing or harmonics proximity.
- Inaccurate settings of algorithm thresholds.

2. ASR test

The second test for evaluating intelligibility of processed speech was conducted using a speaker-independent ASR of the connected digits in the test data. The test was run using the speech recognition software presented in [173]. The recognition algorithm utilizes a dynamic time warping (DTW) technique to compute the scores for a given

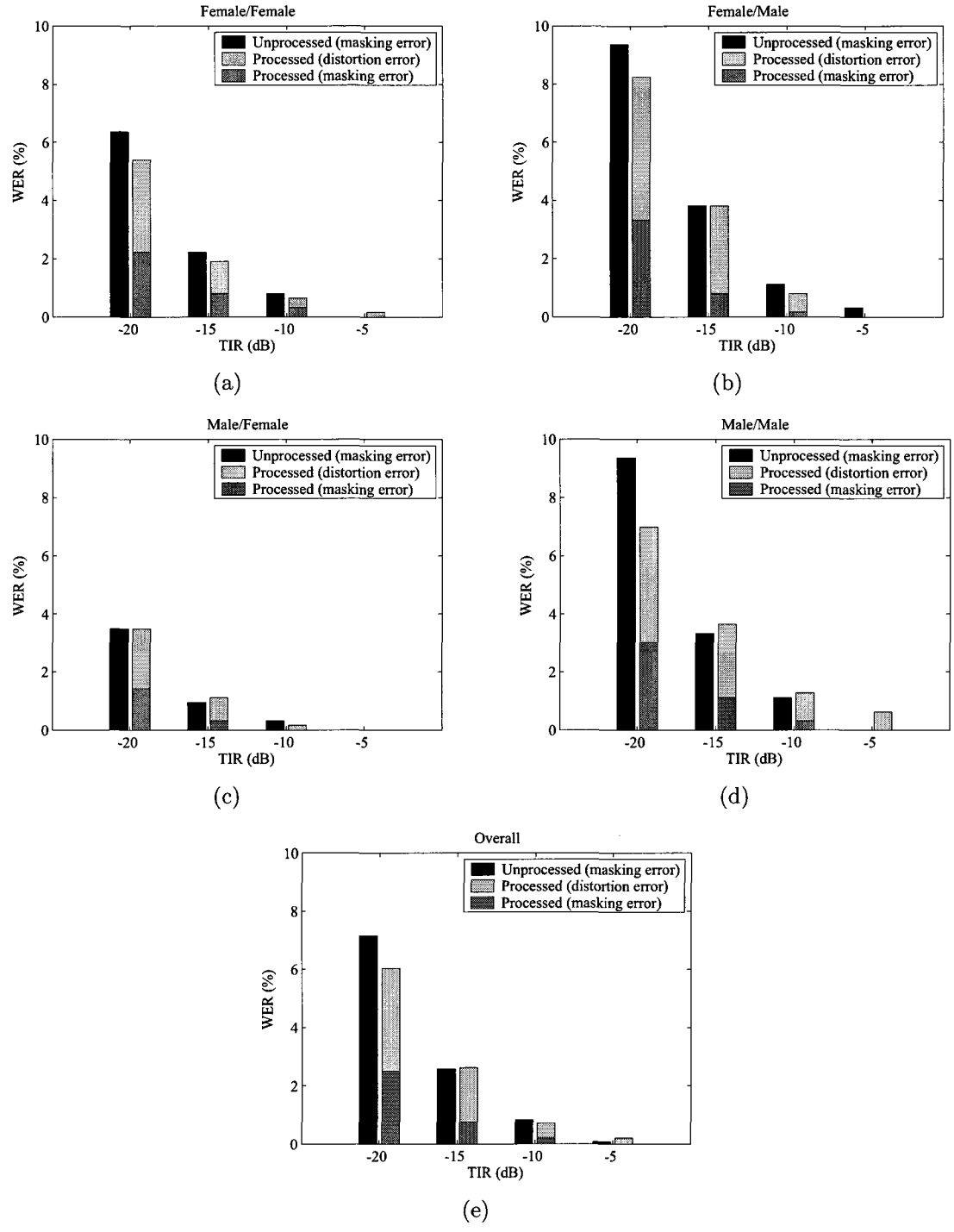


Figure 7.2: Average word error rate (WER) results of the subjective listening test using unprocessed and processed co-channel speech for the four types of mixtures: (a) female/female, (b) female/male, (c) male/female, and (d) male/male. Overall results are shown in (e). Results for processed speech show the masking errors and the distortion errors.

		Recognized number			
		1	2	3	4
Uttered number	1	95.5%	2.8%	0.3%	1.4%
	2	14%	83.2%	0.3%	2.5%
	3	20%	11%	64.8%	4.2%
	4	18.2%	11.6%	1.4%	68.8%

Table 7.4: ASR confusion matrix for unprocessed speech at 0 dB.

test signal against all reference templates using features derived from Mel frequency cepstral coefficients (MFCC). Originally, the ASR system was set up to recognize the ten digits (zero through nine). To further simplify the recognition task, the vocabulary of the training and testing sets were limited to the four digits (one through four). The training set consisted of speech files selected from the TIDIGITS database for male and female speakers different from the speakers in the testing set. Since ASR systems are much more sensitive to the input TIR than human listeners. The chosen range for input TIR was set between -10 dB to 10 dB in steps of 5 dB.

The confusion matrices for the unprocessed and processed data at 0 dB TIR are given in Tables 7.4 and 7.5 respectively. Similar to the subjective listening test, the percentage of recognizing each number was increased as a result of applying the speech separation algorithm. The WER results are shown in Figure 7.3 using the same way discussed earlier in the subjective listening test. As can be seen from the results, the use of automated speech processing at recognizing co-channel corrupted speech is not as effective as human listeners. However, the relative improvement in ASR test results was more pronounced compared to the subjective listening test results. Overall, the WER was improved by approximately 10%, 12%, and 13% at TIR = -10, -5, and 0 dB, respectively. Moreover, the recognition accuracy was approximately the same for the four types of mixtures with the exception of a slight enhancement for male target speakers than female target speakers, regardless of the gender of the interferer.

Another interesting observation is that the inserted error due to distortion is almost negligible in all cases. Compared with the results obtained in the human listening test, this supports the previously mentioned idea regarding the masking

		Recognized number			
		1	2	3	4
Uttered number	1	98.6%	0.7%	0%	0.7%
	2	3.7%	94.5%	0%	1.8%
	3	6.2%	11%	79%	3.8%
	4	2.7%	3.8%	0%	93.5%

Table 7.5: ASR confusion matrix for processed speech at 0 dB.

phenomenon. Unlike human ears, automated systems do not differentiate much between partially masked signals and distorted signals.

7.5.4 Quality measuring tests

The second group of simulations aimed to examine the performance of the proposed CCSS algorithm in terms of speech quality. Three objective measures were used in these simulations to evaluate the performance of the proposed separation algorithm: SNR measure, Itakura Saito distance, and PESQ measure. In all tests, speech quality was estimated by computing the difference between each target (clean) signal and the test signal before and after separation. Overall quality associated with a particular TIR was then calculated as the mean quality across all test signals. All objective measures were calculated for the four different speech mixtures (female/female, female/male, male/female, and male/male) at TIR values in the range from -20 dB up to 20 dB with a 5 dB increment.

1. SNR test

In this experiment, the quality of mixed and separated speech signals was compared in terms of segmental SNR (SNRsg) calculated according to (7.4.2). A plot of the resulted SNRseg versus input TIR is shown in Figure 7.4. In fact, the noise component measured in the unprocessed co-channel speech represented only the amount of interfering signal. Therefore, the measured SNRseg was exactly the same as the input TIR for all types of mixtures in this case. The noise component measured in the processed speech, on the other hand, included residual of the interfering speech

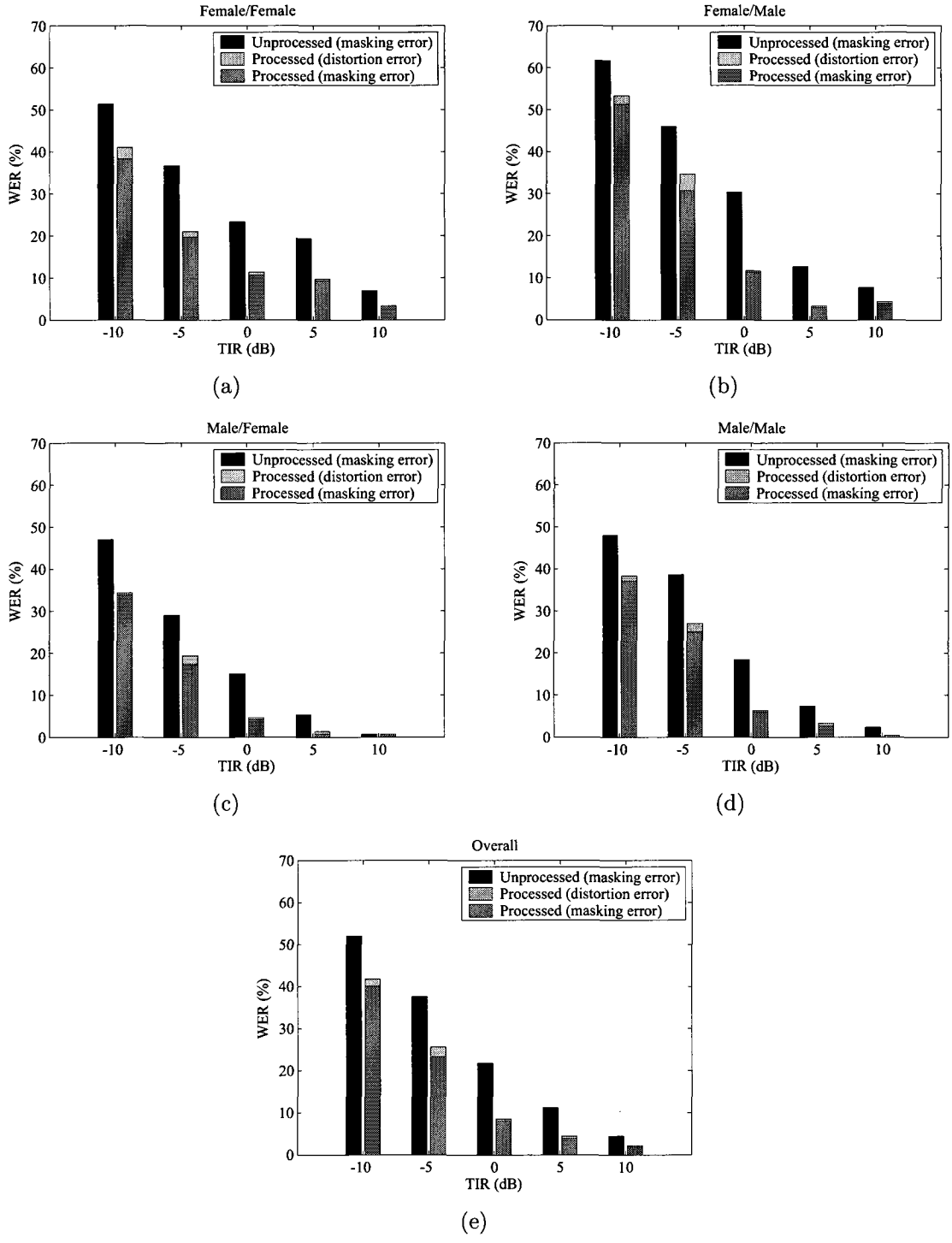


Figure 7.3: Average word error rate (WER) results of the ASR test using unprocessed and processed co-channel speech for the four types of mixtures: (a) female/female, (b) female/male, (c) male/female, and (d) male/male. Overall results are shown in (e). Results for processed speech show the masking errors and the distortion errors.

signal plus signal distortion introduced by the separation system itself. In general, it can be observed that the results obtained from the processed speech outperform the results obtained from unprocessed speech except sometimes at TIR values above 15 dB. This was expected as the signal distortion caused by the system was larger than the residual of the interfering signal at these high TIRs.

In order to view the amount of enhancement achieved at each TIR, the SNRseg gain for the same results is shown in Figure 7.5 in terms of the enhancement factor defined as

$$\mathrm{EF}_{\mathrm{SNR}} = [\mathrm{SNRseg}]_{proc} - [\mathrm{SNRseg}]_{unproc}, \tag{7.5.7}$$

where $[\mathrm{SNRseg}]_{unproc}$ and $[\mathrm{SNRseg}]_{proc}$ are the average segmental SNRs of the unprocessed and the processed speech signals, respectively. According to the plot, the maximum SNRseg enhancement was achieved at TIRs approximately between -10 dB to 0 dB. Furthermore, It can be noticed that the results for male target speaker outperformed the results for female target speaker at low TIRs while the opposite was true at high TIRs. At TIR values above 15 dB, the separated speech for male speakers was even worse than the unprocessed speech regardless of the gender of the interfering speaker.

2. Itakura Saito distance test

The IS distance was calculated according to (7.4.5) using 14 LPC coefficients. Final results for the unprocessed and the processed signals for the four types of mixtures are shown in Figure 7.6. Again, the results using IS distance were consistent with the SNRseg results. Figure 7.7 shows the amount of enhancement achieved in the IS at each TIR in terms of the enhancement factor defined as

$$\mathrm{EF}_{\mathrm{IS}} = [d_{\mathrm{IS}}]_{unproc} - [d_{\mathrm{IS}}]_{proc}, \tag{7.5.8}$$

where $[d_{\mathrm{IS}}]_{unproc}$ and $[d_{\mathrm{IS}}]_{proc}$ are the average IS distances of the unprocessed and the processed speech signals, respectively. The figure indicates that the maximum

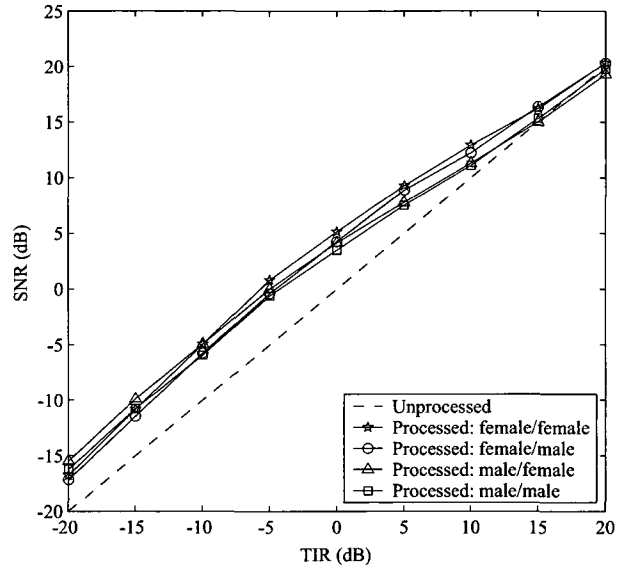


Figure 7.4: Average segmental signal-to-noise ratio (SNRseg) versus TIR for the unprocessed (dashed lines) and the processed (solid lines) signals for the four types of mixtures.

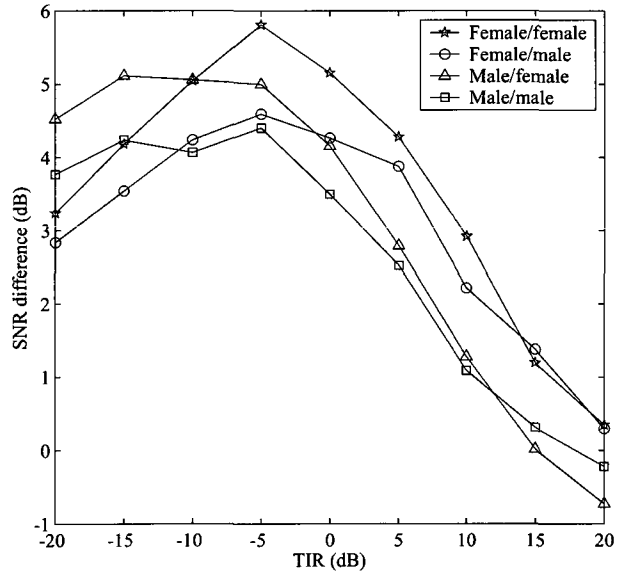


Figure 7.5: The same results of Figure 7.4 expressed in terms of the enhancement factor in (7.5.7).

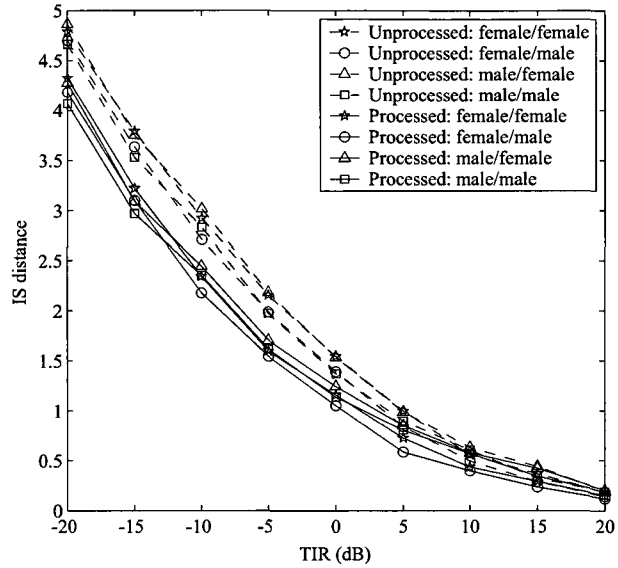


Figure 7.6: Average Itakura-Saito distance versus TIR for the unprocessed (dashed lines) and the processed (solid lines) signals for the four types of mixtures.

enhancement in the IS distance was achieved at TIRs approximately between -20 dB to -5 dB. Similar to the previous test, the results for male target speaker outperformed the results for female target speaker at very low TIRs while the opposite was true at high TIRs.

Occasionally, unpredictable errors occurred when the IS distances (or the LLR distance) was calculated. An example of this kind of error is shown in Figure 7.8. The time domain waveforms representing the clean target speech (solid line), the co-channel speech (dotted line), and the separated speech (dashed line) are plotted in Figure 7.8(a). Corresponding spectral envelops are plotted in Figure 7.8(b). It is very clear that the separated signal is much closer to the target signal than the mixed signal in both time domain waveform and low-frequency spectral envelop. However,

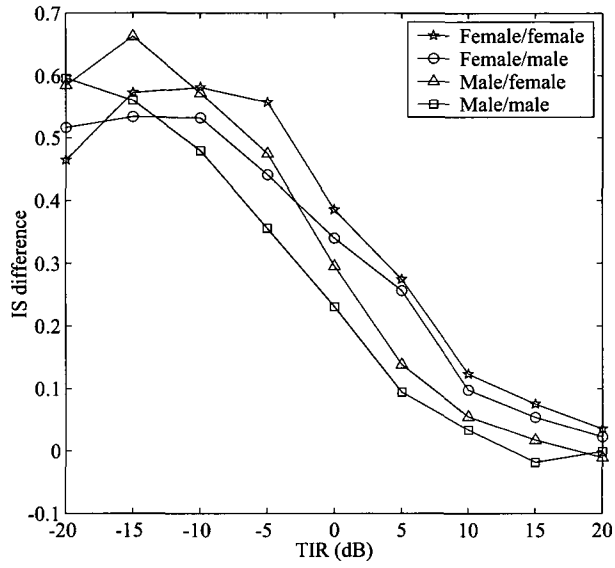
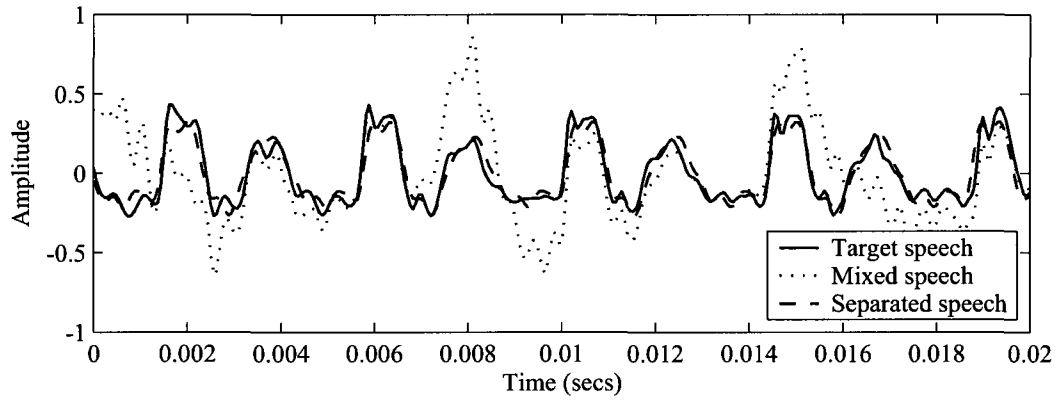


Figure 7.7: The same results of Figure 7.6 expressed in terms of the enhancement factor in (7.5.8).

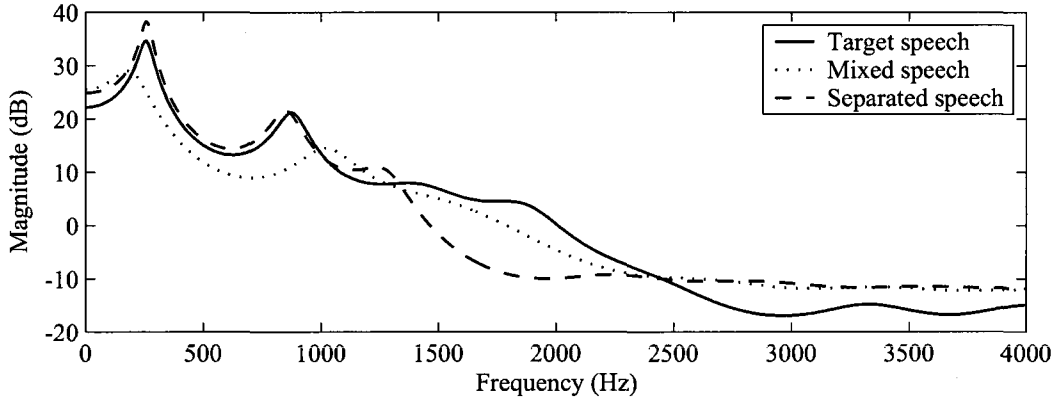
the following results were obtained for this speech segment:

$[\text{SNRseg}]_{unproc}$	$= -1.9$ dB	$[\text{SNRseg}]_{proc}$	$= 9.9$ dB
$[d_{\text{LLR}}]_{unproc}$	$= 0.5$	$[d_{\text{LLR}}]_{proc}$	$= 1.1$
$[d_{\text{IS}}]_{unproc}$	$= 0.8$	$[d_{\text{IS}}]_{proc}$	$= 4.5$

According to figure 7.8(b), the discrepancy between the IS and the LLR distances in one hand and the SNRseg in the other hand is due to spectral divergence in the frequency range from 1300 Hz to 2400 Hz. It is obvious that, in this frequency range, the area under the spectral envelop difference between the separated speech and the target speech is much bigger than the same area for the co-channel speech (see equation (7.4.4)). To partially overcome this problem, only 95% of the total number of frames with the lowest distances were considered when calculating the average.



(a)



(b)

Figure 7.8: Unpredictable error in calculating the LLR and IS measures: (a) 20 ms co-channel speech frame in the time domain and (b) corresponding spectral envelops.

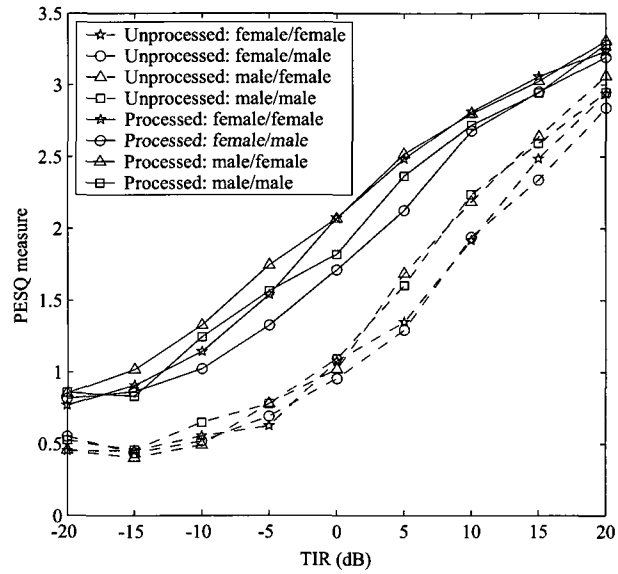


Figure 7.9: Average PESQ measure versus TIR for the unprocessed (dashed lines) and the processed (solid lines) signals for the four types of mixtures.

3. PESQ measure test

In the final test, performance of the separation system was evaluated in terms of the PESQ measure and the results are plotted in Figure 7.9. As can be seen from the figure, the results of the processed speech outperform the unprocessed speech at all TIRs. Figure 7.10, also, shows the amount of enhancement achieved in the PESQ measure at each TIR in terms of the enhancement factor defined as

$$\mathrm{EF}_{\mathrm{PESQ}} = \mathrm{PESQ}_{proc} - \mathrm{PESQ}_{unproc}, \tag{7.5.9}$$

where PESQ_{unproc} and PESQ_{proc} are the average PESQ measures of the unprocessed and the processed speech signals, respectively.

7.5.5 Performance evaluation on the speech separation challenge

In the next group of simulations, the performance of our proposed system was evaluated and compared with other systems using the speech separation challenge

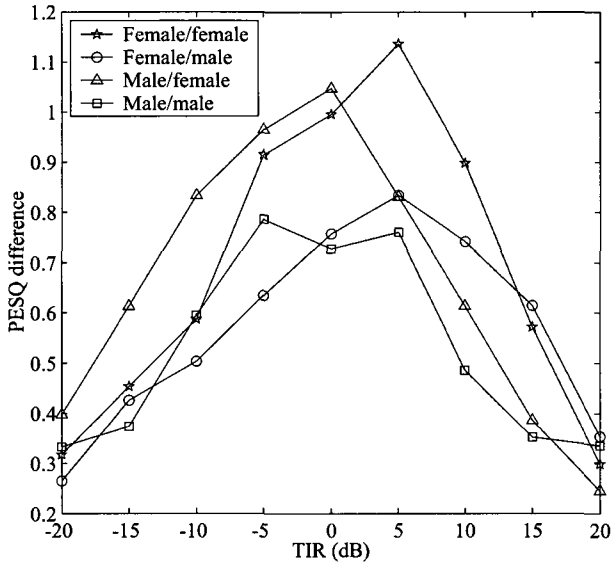


Figure 7.10: The same results of Figure 7.9 expressed in terms of the enhancement factor in (7.5.9).

proposed by Cooke and Lee (2006) [156]. This challenge aims to automatically recognize a target speech in the presence of another competing speaker. All speech files in this challenge are single-channel "wav" data sampled at 25 kHz. Every uttered sentence follows the 6-keyword syntax: "<command:4> <color:4> <preposition:4> <letter:25> <digit:10> <adverb:4>," where the number associated with each keyword indicates the number of available choices. Table 7.6 summarizes all the possible choices for each keyword. Speech database of this challenge consists of two sets of speech files; one set for training data and one set for test data. The training data set is drawn from a closed set of 34 speakers and consists of 17,000 utterances. Co-channel test data set contains pairs of sentences mixed at 6 different TIRs: -9, -6, -3, 0, 3 and 6 dB with a total of 600 sentences under each TIR condition. One third of this data consists of same talker (ST) mixtures, another third comprises of mixtures of different talkers of the same gender (SG), and the remaining third consists of different gender (DG) mixtures. The clean utterances for the test data are also provided. The co-channel speech corpus also contains a development data

Command	Color	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	1-9 and	again
lay	green	by	excluding W	Zero	now
place	red	in			please
set	white	with			soon

Table 7.6: Possible choices in each position for a speech utterance in the speech separation challenge. An example of an utterance could be "place white at L 3 now."

set. This set was used in the simulations to tune the various parameters of our CCSS system. According to the challenge, the recognition task is to recognize the letter and the digit of the speaker who said the color keyword "white". The ASR system is first trained using the clean speech data and then tested using the co-channel speech data. Separation accuracy is evaluated by comparing recognition results before and after separation.

The recognition test was first applied to the test data without any processing to generate a baseline performance for comparison. Table 7.7 shows the recognition accuracy results obtained at different TIRs using the unprocessed test data [156]. The results are tabulated separately for interfering speakers of the ST, the SG and the DG categories, as well as for the overall average score (Avg.).

Next, the recognition test was conducted using the same speech data after being processed by our proposed system shown in Figure 6.5. Similar to the previous simulations, multi-pitch information of individual speakers was assumed to be known from the clean speech signals before mixing. The obtained results are shown in Table 7.8. As can be concluded by comparing Tables 7.7 and 7.8, the separation system was able to enhance the recognition accuracy especially at TIRs equal to or greater than -3 dB. Greater improvements were observed under the SG and DG conditions. However, the system did not perform nearly as well under the ST condition which is not a realistic condition. This is mainly due to the frequent proximity of pitch contours of the two speech signals in the mixture.

TIR	ST	SG	DG	Avg.
Clean	—	—	—	98.56%
6dB	62.44%	64.25%	64.25%	63.58%
3dB	46.15%	44.13%	46.75%	45.75%
0dB	29.64%	32.96%	33.50%	31.92%
-3dB	18.10%	20.95%	19.50%	19.42%
-6dB	9.73%	14.53%	11.50%	11.75%
-9dB	5.66%	7.26%	7.50%	6.75%

Table 7.7: Recognition accuracy obtained using unprocessed test data [156].

TIR	ST	SG	DG	Avg.
Clean	—	—	—	92.25
6dB	59.40%	62.5%	70.44%	64.07%
3dB	45.25%	50.15%	59.30%	51.35%
0dB	27.51%	46.68%	56.15%	43.63%
-3dB	18.25%	29.25%	39.47%	28.91%
-6dB	8.13%	12.30%	13.16%	11.14%
-9dB	5.10%	12.57%	24.12%	13.86%

Table 7.8: Recognition accuracy of the test data processed by the speech separation algorithm.

The recognition accuracies of the speech data processed by our proposed method were then compared with the recognition accuracies of:

1. The unprocessed test data (baseline performance) [156].
2. Eighteen human listeners' test of the unprocessed data [156].
3. The CASA-based algorithm presented by Runqiang *et al.* (2006) in [62].
4. The sparse non-negative matrix factorization (SNMF) algorithm proposed by Schmidt *et al.* (2006) in [74].

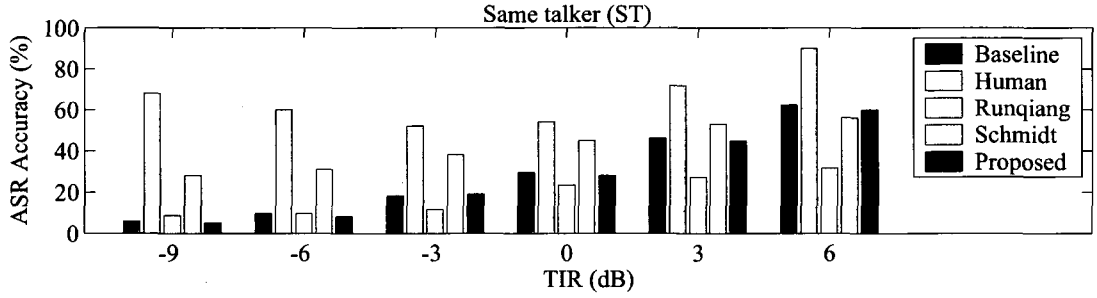
Runqiang's algorithm employed a CASA technique to separate the target speech from the interfering speech. It used the training data set of the speech corpus to train the models for speaker recognition as well as the models for speech reconstruction. Schmidt's algorithm, on the other hand, utilized an unsupervised SNMF method in the training phase to learn sparse representations of the data. This was applied to the

learning of personalized dictionaries from the training speech corpus on a phoneme level, which in turn were used to separate the audio stream into its components.

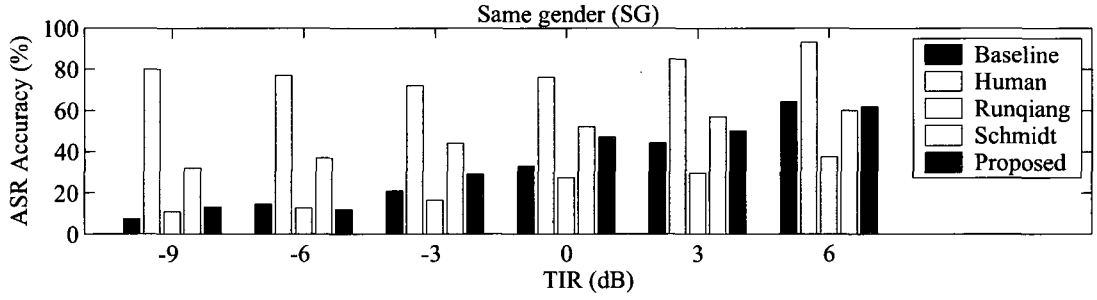
Final results using test data sets are summarized in the plots shown Figure 7.11. It can be observed from the plots that, under almost all mixing conditions, the human listening test introduced the highest recognition accuracies whereas the automated recognition test of the unprocessed speech data and Runqiang's algorithm introduced the lowest accuracies. From the figure, Runqiang's algorithm was slightly successful in enhancing the recognition accuracy only at low TIRs of -6 and -9 dB. In most cases, our algorithm and Schmidt's algorithm performed better then Runqiang's algorithm with Schmidt's algorithm outperforming our algorithm. On average, the proposed approach gave approximately the same results as the Schmidt's at 6 dB TIR. However, the results at this TIR showed a slightly better performance in our algorithm than Schmidt's under the ST and SG conditions compared to the DG condition.

7.6 Summary

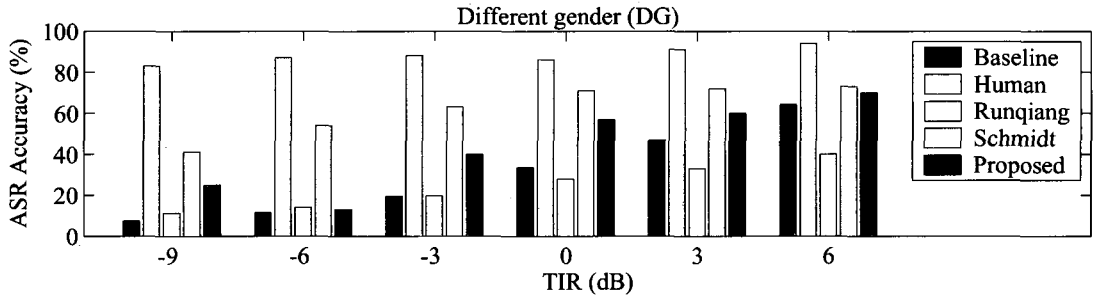
The overall separation system implemented in the previous chapter was tested and its performance was evaluated using several subjective and objective measures and metrics such as human listening tests, ASR tests, segmental SNR measure, Itakura-Saito distance, and PESQ measure. The results showed that the proposed CCSS front-end was successful in separating the co-channel speech for the four different types of mixture and at different TIRs. The proposed system was also successful in enhancing the recognition accuracy of the speech separation challenge especially under the conditions of different speakers (i.e., SG and DG conditions).



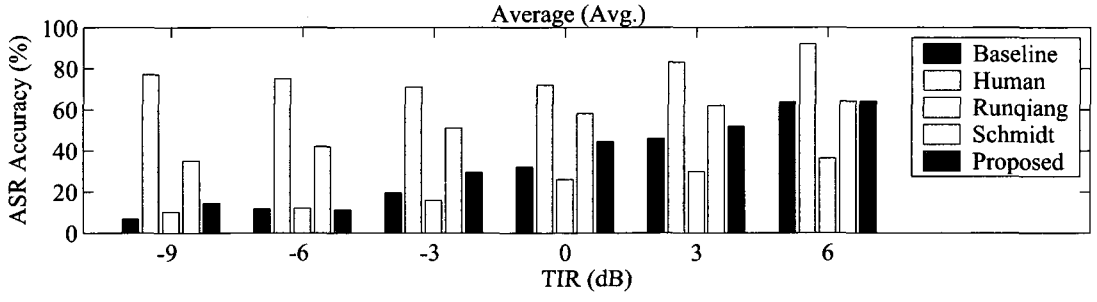
(a)



(b)



(c)



(d)

Figure 7.11: Recognition accuracy comparison in terms of word recognition rate versus TIR under: (a) same-talker (ST), (b) same-gender (SG), and (c) different-gender (DG) conditions. Overall average (Avg.) results are plotted in (d).

CHAPTER 8

Conclusion and Future Work

8.1 Conclusion

A general challenge in many speech applications is the separation or extraction of the underlying speech sources from a mixture. A classical example of this challenge is the so called cocktail-party problem in which the task is to recognize or isolate what is being said by an individual speaker in a mixture of speech from various speakers. A particular difficult situation of the cocktail-party problem occurs when only a single-channel recording is available. This is commonly known as the co-channel speech separation problem.

The objective of this thesis was to present some contributions towards the design and implementation of a robust and enhanced co-channel speech separation system based on state-space reconstruction and sinusoidal modelling in time domain. To achieve this objective, new algorithms for the main three stages (*Analysis*, *Separation*, and *Reconstruction*) of the separation system were introduced. As opposed to conventional frequency-domain techniques, the main analysis and processing approaches exploited in this work were based on the time-domain representation of speech.

In the first contribution, a frame-based classification algorithm that is capable of determining the voicing-state of co-channel speech was presented. The algorithm was based on nonlinear state-space reconstruction of speech data. Using the TPM of

the reconstructed data along with other features from the input speech such as the average STE and the HILO ratio, the algorithm classified the co-channel speech into the three voicing-states; Unvoiced/Unvoiced (U/U), Voiced/Unvoiced (V/U), and Voiced/Voiced (V/V). The proposed method required neither *a priori* information nor speech training data.

The performance of the voicing-state classifier was also investigated. Compared to other techniques [37, 40], the proposed algorithm gave superior results at different TIR values as well as different levels of background noise. An overall success rate of 85% was achieved in the correct classification of speech segments at 0 dB TIR. The largest improvement was realized in identifying V/U segments.

The sample-based modified version of the classification method was more reasonable for real-time implementation by reducing computational complexity and helping in segmenting the speech waveform according to the characteristics of speech instead of regularly-spaced frames.

Secondly, a time-domain method to precisely estimate the sinusoidal model parameters of co-channel speech was presented. The method did not require the calculation of the STFT nor the multiplication by a window function. The proposed method incorporated a least-squares estimator and an adaptive technique to model and separate the co-channel speech into its individual speakers, all in the time domain. By avoiding the STFT and windowing effects, imprecisions were eliminated that would otherwise be introduced by those components.

The application of this new time-domain method on real and synthetic data demonstrated the effectiveness of this method in separating co-channel speech signals at different TIRs. Overall, an improvement of 1-3 dB in MSE was obtained over the frequency-domain method. We also noted that the time-domain method was not as sensitive to additive white Gaussian noise as the frequency-domain methods. This result was particularly true for lower SNR situations.

Finally, these algorithms were integrated together along with other necessary

components to implement the overall CCSS system. This system was tested and its separation performance was evaluated using numerous subjective and objective measures under different speech mixing conditions. Conducted simulation results have shown the effectiveness of the proposed system as a front end in solving the CCSS problem to a good extent.

8.2 Summary of Contributions

The primary contribution of this thesis was the development of a set of new algorithms to implement an efficient co-channel speech separation system with two interfering speeches as shown in Figure 6.5. The main contributions throughout this research can be summarized as follows:

1. Development of a new frame-based voicing-state classification method to classify the co-channel speech segment into unvoiced, single-voiced, or double-voiced [2].

2. Testing the performance of the proposed voicing-state classifier at different levels of interfering speech and background noise [3].

3. Development of a time-domain method to precisely estimate the sinusoidal model parameters of co-channel speech [4].

4. Modifying the above algorithms using a sample-based voicing-state classification method and a simplified method for estimating the sinusoidal model parameters.

5. Integrating and implementing a CCSS system that was successful in separating co-channel speech under different input conditions.

6. Testing the intelligibility and quality of the speech separated by the proposed algorithm using different subjective and objective methods.

8.3 Future Work

The contributions made by this study have shown that nonlinear techniques offer a good potential in the area of co-channel speech separation. However, there is still a room to further investigate and refine the ideas presented in this thesis. Some of these areas are discussed below.

8.3.1 Multi-pitch tracking

Accurate pitch determination and tracking for both speakers are crucial operations in most speaker separation algorithms. Developing a reliable multi-pitch tracking algorithm is important to improve the quality and the intelligibility of the processed speech. It also leads to the enhancement of the speaker assignment stage and, consequently, the overall performance of the separation system.

8.3.2 CASA-oriented techniques

As indicated in the literature review, CASA-based approaches were among the most promising techniques to solve the CCSS problem. This is due the fact that the essential idea in the CASA system is the attempt to mimic the human auditory system for speech separation. Merging the proposed methods with a CASA-based techniques could lead to a better separation performance.

8.3.3 Speech masking

Auditory masking occurs when the "perception" of a weaker sound is affected (made inaudible) by the presence of a louder sound. This commonly happens in speech applications when a loud noise-like sound masks the desired speech signal. However, auditory masking of speech can also occur due to another speech sound with higher intensity. Investigating the masking effect on speech corrupted by another speech sound and utilizing this effect to improve the perceptual quality of the separated speech is another interesting area of research. For example, exploiting speech

masking effects when calculating the sinusoidal model parameters of the target speech could dramatically reduce the perceptual artifacts of the reconstructed speech. This technique can be very useful if the output speech is to be listened to by humans.

8.3.4 Quality measurements

It is important to find new techniques to measure the quality of separated co-channel speech that works better for speech corrupted by another speech. For example, CCSS system evaluation methods can be extended to other application-oriented tests such as SID and listening tests of hearing-impaired subjects. Furthermore, it is also important to develop a standard speech database created specifically for testing co-channel speech separation systems. Such a database needs to have speech recordings in a real co-channel environment where two (or more) speakers speak simultaneously at a certain distance from a single microphone. In addition, the database should contain another set of speech recordings of clean speech for the same speakers. This second set can be used, for example, in a training phase, if needed.

References

[1] O. M. Mitchell, C. A. Ross, and G. H. Yates, "Signal processing for a cocktail party effect," *Journal of the Acoustical Society of America*, vol. 50, pp. 656–660, August 1971.

[2] Y. A. Mahgoub and R. M. Dansereau, "Voicing-state classification of co-channel speech using nonlinear state-space reconstruction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2005)*, vol. 1, (Philadelphia, PA, USA), pp. 409–412, March 2005.

[3] Y. A. Mahgoub and R. M. Dansereau, "Performance of phase-space voicing-state classification for co-channel speech," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, vol. 1, (Ottawa, ON, Canada), pp. 742–747, May 2005.

[4] Y. A. Mahgoub and R. M. Dansereau, "Time domain method for precise estimation of sinusoidal model parameters of co-channel speech," *Research Letters in Signal Processing*, vol. 2008, Article ID 364674, p. 5 pages, 2008.

[5] G. A. Miller, "The masking of speech," *Psychological Bulletin*, vol. 44, pp. 105–129, March 1947.

[6] J. P. L. Brokx and S. G. Nooteboom, "Intonation and the perception of simultaneous voices," *Journal of Phonetics*, vol. 10, pp. 23–26, 1982.

[7] E. C. Cherry, "Some experiments on the recognition of speech with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, September 1953.

[8] R. E. Yantorno, "Co-channel speech and speaker identification study," tech. rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, Rome, New York, September 1998.

[9] P. F. Assmann and A. Q. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System* (S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. Fay, eds.), ch. 5, pp. 231–308, New York, NY, USA: Springer-Verlag, 2004.

[10] M. L. Seltzer, "Automatic detection of corrupt spectrographic features for robust speech recognition," Master's thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 2000.

[11] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 407–424, September 1997.

[12] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," tech. rep., Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA, 2007.

[13] R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Effects of co-channel interference on speaker identification," in *Proceedings of the SPIE International Symposium on Technologies for Law Enforcement*, November 2000.

[14] J. M. Lovekin, R. E. Yantorno, K. R. Krishnamachari, D. S. Benincasa, and S. J. Wenndt, "Developing usable speech criteria for speaker identification technology," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2003)*, vol. 1, (Hong Kong, China), pp. 421–424, May 2001.

[15] Y. Shao and D. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2003)*, vol. 2, (Hong Kong, China), pp. 205–208, April 2003.

[16] A. N. Iyer, B. Y. Smolenski, R. E. Yantorno, J. K. Shah, E. J. Cupples, and S. J. Wenndt, "Speaker identification improvement using usable speech concept," in *Proceedings of the 12th European Signal Processing Conference (EUSIPCO-2004)*, (Vienna, Austria), pp. 349–352, September 2004.

[17] P. Divenyi, ed., *Speech separation by humans and machines.* Boston, MA, USA: Kluwer Academic Publishers, 2005.

[18] V. C. Shields, "Separation of added speech signals by digital comb filtering," Master's thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, 1970.

[19] R. H. Frazier, "An adaptive filtering approach toward speech enhancement," Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, June 1975.

[20] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1976)*, vol. 1, (Philadelphia, PA, USA), pp. 251–253, April 1976.

[21] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 354–358, August 1978.

[22] Y. Perlmutter, L. Braids, R. Frazier, and A. Oppenheim, "Evaluation of a speech enhancement system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1977)*, vol. 2, (Hartford, CT, USA), pp. 212–215, May 1977.

[23] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *Journal of the Acoustical Society of America*, vol. 60, pp. 911–918, October 1976.

[24] R. J. Stubbs and Q. Summerfield, "Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 84, pp. 1236–1249, October 1988.

[25] R. J. Stubbs and Q. Summerfield, "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 87, pp. 359–372, January 1990.

[26] B. A. Hanson and D. Y. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1984)*, vol. 9, (San Diego, CA, USA), pp. 65–68, March 1984.

[27] J. A. Naylor and S. F. Boll, "Techniques for suppression of an interfering talker in co-channel speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1987)*, vol. 12, (Dallas, TX, USA), pp. 205–208, April 1987.

[28] C. K. Lee and D. G. Childers, "Cochannel speech separation," *Journal of the Acoustical Society of America*, vol. 83, pp. 274–280, January 1988.

[29] M. A. Zissman, C. J. Weinstein, L. D. Braida, R. M. Uchanski, and W. M. Rabinowitz, "Speech-state adaptive simulation of co-channel talker interference suppression," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1989)*, vol. 1, (Glasgow, Scotland), pp. 361–364, May 1989.

[30] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 56–69, January 1990.

[31] F. M. Silva and L. B. Almeida, "Speech separation by means of stationary least-squares harmonic estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1990)*, vol. 2, (Albuquerque, NM, USA), pp. 809–812, April 1990.

[32] Y. H. Gu and W. M. G. van Bokhoven, "Co-channel speech separation using frequency bin non-linear adaptive filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1991)*, vol. 2, (Toronto, ON, Canada), pp. 949–952, April 1991.

[33] J. Naylor and J. Porter, "An effective speech separation system which requires no *a priori* information," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1991)*, vol. 2, (Toronto, ON, Canada), pp. 937–940, May 1991.

[34] H. G. M. Savic and J. Sorensen, "Co-channel speaker separation based on maximum-likelihood deconvolution," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1994)*, vol. 1, (Adelaide, Australia), pp. 25–28, April 1994.

[35] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Co-channel speaker separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1995)*, vol. 1, (Detroit, MI, USA), pp. 828–831, May 1995.

[36] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 418–423, October 1976.

[37] D. S. Benincasa and M. I. Savic, "Co-channel speaker separation using constrained nonlinear optimization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1997)*, vol. 2, (Munich, Germany), pp. 1195–1198, April 1997.

[38] R. E. Yantorno, B. Y. Smolenski, and N. Chandra, "Usable speech measures and their fusion," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS-2003)*, vol. 3, (Bangkok, Thailand), pp. 734–737, May 2003.

[39] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS-2000)*, (Honolulu, HI, USA), pp. 710–713, November 2000.

[40] N. Chandra and R. E. Yantorno, "Usable speech detection using the modified spectral autocorrelation peak to valley ratio using the LPC residual," in *Proceedings of the 4th IASTED International Conference on Signal and Image Processing*, (Hawaii, USA), pp. 146–150, August 2002.

[41] K. R. Krishnamachari, R. E. Yantorno, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2001)*, vol. 1, (Salt Lake City, UT, USA), pp. 649–652, May 2001.

[42] J. M. Lovekin, K. R. Krishnamachari, and R. E. Yantorno, "Adjacent pitch period comparison (APPC) as a usability measure of speech segments under co-channel conditions," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS-2001)*, (Nashville, TN, USA), pp. 139–142, November 2001.

[43] M. T. Johnson, A. C. Lindgren, R. J. Povinelli, and X. Yuan, "Co-channel speech detection approaches using cyclostationarity or wavelet transform," in *Proceedings of the 4th IASTED International Conference on Signal and Image Processing (SIP-2002)*, (Kauai, USA), August 2002.

[44] A. N. Iyer, M. Gleiter, B. Y. Smolenski, and R. E. Yantorno, "Structural usable speech measure using LPC residual," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS-2003)*, (Awaji Island, Japan), pp. 236–240, December 2003.

[45] N. Sundaram, R. E. Yantornoa, B. Y. Smolenski, and A. N. Iyer, "usable speech detection using linear predictive analysis a model-based approach," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS-2003)*, (Awaji Island, Japan), pp. 231–235, December 2003.

[46] U. O. Ofoegbu, B. Smolenski, and R. Yontorno, "Structure-based voiced/usable speech detection using state space embedding," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS-2004)*, (Seoul, Korea), pp. 811–815, November 2004.

[47] J. K. Shah, B. Y. Smolenski, and R. E. Yantorno, "Decision level fusion of usable speech measures using consensus theory," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS-2003)*, (Awaji Island, Japan), pp. 245–250, December 2003.

[48] Z. Heming, Z. Xudong, and Y. Yibiao, "Co-channel speech separation based on sinusoidal model for speech," in *Proceedings of the 5th International Conference on Signal Processing*, vol. 2, (Beijing, China), pp. 815–818, August 2000.

[49] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America*, vol. 83, pp. 257–264, January 1988.

[50] A. S. Bregman, *Auditory scene analysis: the perceptual organization of sound.* Cambridge, MA, USA: MIT Press, 1990.

[51] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, pp. 35–50, July 1992.

[52] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[53] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement* (J. Benesty, S. Makino, and J. Chen, eds.), ch. 16, pp. 371–402, New York: Springer, 2005.

[54] G. Hu and D. L. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control* (E. Hänsler and G. Schmidt, eds.), ch. 12, pp. 485–515, New York, NY, USA: Springer, 2006.

[55] M. Weintraub, "A computational model for separating two simultaneous talkers," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1986)*, vol. 11, (Tokyo, Japan), pp. 81–84, April 1986.

[56] R. Meddis and M. J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *Journal of the Acoustical Society of America*, vol. 91, pp. 233–245, January 1992.

[57] M. P. Cooke, *Modeling auditory processing and organization.* PhD thesis, University of Sheffield, UK, 1993.

[58] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis.* PhD thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, USA, June 1996.

[59] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, pp. 684–697, May 1999.

[60] L. Ottaviani and D. Rocchesso, "Separation of speech signal from complex auditory scenes," in *Proceedings of the COST G-6 Conference on Digital Audio Effects*, (Limerick, Ireland), pp. 87–90, December 2001.

[61] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1135–1150, September 2004.

[62] H. Runqiang, Z. Pei, G. Qin, Z. Zhiping, W. Hao, and W. Xihong, "CASA based speech separation for robust speech recognition," in *International Conference on Spoken Language Processing (INTERSPEECH 2006 - ICSLP)*, (Pittsburgh, PA, USA), pp. 77–80, September 2006.

[63] S. Srinivasan, Y. Shao, Z. Jin, and D. Wang, "A computational auditory scene analysis system for robust speech recognition," in *International Conference on Spoken Language Processing (INTERSPEECH 2006 - ICSLP)*, (Pittsburgh, PA, USA), pp. 73–76, September 2006.

[64] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 708–716, November 2000.

[65] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 396–405, February 2007.

[66] L. Gu and R. M. Stern, "Single-channel speech separation based on modulation frequency," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2008)*, (Las Vegas, NV, USA), pp. 25–28, April 2008.

[67] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 2014–2023, November 2006.

[68] M. Wu, D. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 229–241, May 2003.

[69] D. L. Wang and G. J. Brown, eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Hobbken, New Jersey: Wiley-IEEE Press, 2006.

[70] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, May-June 2000.

[71] A. J. W. van der Kouwe, D. Wang, and G. J. Brown, "A comparison of auditory and blind separation techniques for speech segregation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 189–195, March 2001.

[72] Y. Li, S. Amari, A. Cichocki, D. W. C. Ho, and X. Shengli, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 423–437, February 2006.

[73] G. Jang, T. Lee, and Y. Oh, "Single-channel signal separation using time-domain basis functions," *IEEE Signal Processing Letters*, vol. 10, pp. 168–171, June 2003.

[74] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *International Conference on Spoken Language Processing (INTERSPEECH 2006 - ICSLP)*, (Pittsburgh, PA, USA), pp. 2614–2617, September 2006.

[75] G. Jang and T. Lee, "A probabilistic approach to single channel blind signal separation," in *Advances in Neural Information Processing Systems (NIPS-2002)* (S. Becker, S. Thrun, and K. Obermayer, eds.), vol. 15, pp. 1197–1204, Cambridge, MA, USA: MIT Press, December 2002.

[76] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A joint probabilistic-deterministic approach using source-filter modeling of speech signals for single channel speech separation," in *Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing (MSLP-2006)*, vol. 1, (Maynooth, Ireland), pp. 47–52, September 2006.

[77] A. M. Reddy and B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in *International Conference on Spoken Language Processing (INTERSPEECH 2004 - ICSLP)*, (Jeju Island, Korea), pp. 2445–2448, October 2004.

[78] H. A. T. Kristjansson and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, vol. 2, (Montreal, QC, Canada), pp. 817–820, May 2004.

[79] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems (NIPS-2001)* (T. K. Leen, T. G. Dietterich, and V. Tresp, eds.), vol. 13, pp. 793–799, Cambridge, MA, USA: MIT Press, December 2001.

[80] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH-2003)*, (Geneva, Switzerland), pp. 1009–1012, September 2003.

[81] D. E. M. J. Reyes-Gomez and N. Jojic, "Multiband audio modeling for single channel acoustic source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, vol. 5, (Montreal, Canada), pp. 641–644, May 2004.

[82] A. de Cheveigné, "A mixed speech F0 estimation algorithm," in *Proceedings of the 2nd European Conference on Speech Communication and Technology (EUROSPEECH-1991)*, (Genova, Italy), pp. 445–448, September 1991.

[83] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, pp. 353–362, October 1974.

[84] A. de Cheveigné and H. Kawahara, "Multiple period estimation and pitch perception model," *Speech Communication*, vol. 27, pp. 175–185, April 1999.

[85] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1993)*, vol. 2, (Minneapolis, MN, USA), pp. 728–731, April 1993.

[86] M. R. Schoeder, "Period histogram and product spectrum: new methods for fundamental frequency measurements," *Journal of the Acoustical Society of America*, vol. 43, no. 4, pp. 829–834, 1968.

[87] Y. H. Kwon, D. J. Park, and B. C. Ihm, "Simplified pitch detection algorithm of mixed speech signals," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS-2000)*, vol. 3, (Geneva, Switzerland), pp. 722–725, May 2000.

[88] F. Sha and L. K. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS 2004)* (L. K. Saul, Y. Weiss, and L. Bottou, eds.), vol. 17, pp. 1233–1240, Cambridge, MA, USA: MIT Press, December 2004.

[89] M. H. Radfar, A. Sayadiyan, and R. M. Dansereau, "A new algorithm for two-speaker pitch tracking in single channel paradigm," in *Proceedings of the 8th International Conference on Signal Processing (ICSP-2006)*, vol. 1, (Guilin, China), November 2006.

[90] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128–133, April 1951.

[91] A. Khurshid and S. L. Denham, "A temporal-analysis-based pitch estimation system for noisy speech with a comparative study of performance of recent systems," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1112–1124, September 2004.

[92] H. D. I. Abarbanel, *Analysis of observed chaotic data.* New York: Springer, 1st ed., 1996.

[93] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Physical Review Letters*, vol. 45, pp. 712–716, September 1980.

[94] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Lecture Notes in Mathematics* (D. A. Rand and L. S. Young, eds.), vol. 898, pp. 366–381, Springer-Verlag, 1981.

[95] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the Atmospheric Sciences*, vol. 20, pp. 130–141, March 1963.

[96] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, pp. 579–616, November 1991.

[97] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical Review A*, vol. 45, pp. 3403–3411, March 1992.

[98] C. Rhodes and M. Morari, "False-nearest-neighbors algorithm and noise-corrupted time series," *Physical Review E*, vol. 55, pp. 6162–6170, May 1997.

[99] T. Aittokallio, M. Gyllenberg, J. Hietarinta, T. Kuusela, and T. Multamäki, "Improving the false nearest neighbors method with graphical analysis," *Physical Review E*, vol. 60, pp. 416–421, July 1999.

[100] R. Hegger and H. Kantz, "Improved false nearest neighbor method to detect determinism in time series data," *Physical Review E*, vol. 60, pp. 4970–4973, October 1999.

[101] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, "The analysis of observed chaotic data in physical systems," *Reviews of Modern Physics*, vol. 65, pp. 1331–1392, October 1993.

[102] A. M. Fraser and H. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, vol. 33, pp. 1134 – 1140, February 1986.

[103] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modeling* (W. Hardcastle and A. Marchal, eds.), vol. 55, pp. 241–261, Boston, MA, USA: Kluwer Academic Publishers, 1990.

[104] S. McLaughlin and A. Lowry, "Nonlinear dynamical systems concepts in speech analysis," in *Proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH-1993)*, (Berlin, Germany), pp. 377–380, September 1993.

[105] M. Banbrook and S. McLaughlin, "Is speech chaotic?: invariant geometrical measures for speech data," in *IEE Colloquium on Exploiting Chaos in Signal Processing*, vol. 8, (London, UK), pp. 1–10, June 1994.

[106] G. Kubin, "Nonlinear processing of speech," in *Speech coding and synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), ch. 16, pp. 557–610, Amsterdam, The Netherlands: Elsevier Science Inc., 1995.

[107] A. Kumar and S. Mullick, "Nonlinear dynamical analysis of speech," *Journal of the Acoustical Society of America*, vol. 100, pp. 615–629, July 1996.

[108] L. R. Rabiner and R. W. Shafer, *Digital processing of speech signals*. Englewood Cliffs, NJ: Prentice Hall, 1978.

[109] J. Thyssen, H. Nielsen, and S. Hansen, "Nonlinear short-term prediction in speech coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1994)*, vol. 1, (Adelaide, SA, Australia), pp. 185–188, April 1994.

[110] M. Faúndez-Zanuy, G. Kubin, W. B. Kleijn, P. Maragos, S. McLaughlin, A. Esposito, A. Hussain, and J. Schoentgen, "Nonlinear speech processing: overview and applications," *International Journal of Control and Intelligent Systems*, vol. 30, no. 1, pp. 1–10, 2002.

[111] B. Townshend, "Nonlinear prediction of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1991)*, vol. 1, (Toronto, Canada), pp. 425–428, April 1991.

[112] A. Wang, Z. Sun, and X. Zhang, "A non-linear prediction speech coding system based on ANN," in *Proceedings of the 4th World Congress on Intelligent Control and Automation*, vol. 1, (Shanghai, China), pp. 607–611, June 2002.

[113] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 1–17, January 1999.

[114] I. Mann and S. McLaughlin, "Stable speech synthesis using recurrent radial basis functions," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH-1999)*, vol. 5, (Budapest, Hungary), pp. 2315–2318, September 1999.

[115] M. Faúndez-Zanuy, "On the usefulness of linear and nonlinear prediction residual signals for speaker recognition," in *Proceedings of the ITRW on Non-Linear Speech Processing (NOLISP-2007)*, (Paris, France), pp. 19–22, May 2007.

[116] A. C. Lindgren, M. T. Johnson, and R. J. Povinelli, "Speech recognition using reconstructed phase space features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2003)*, vol. 1, (Hong Kong, China), pp. 60–63, April 2003.

[117] A. C. Lindgren, M. T. Johnson, and R. J. Povinelli, "Joint frequency domain and reconstructed phase space features for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, vol. 1, (Montreal, QC, Canada), pp. 533–536, May 2004.

[118] A. Petry and D. A. C. Barone, "Fractal dimension applied to speaker identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2001)*, vol. 1, (Salt Lake City, UT, USA), pp. 405–408, May 2001.

[119] A. M. Chan and H. Leung, "Equalization of speech and audio signals using a nonlinear dynamical approach," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 356–360, MAY 1999.

[120] R. Hegger, H. Kantz, and L. Matassini, "Noise reduction for human speech signals by local projections in embedding spaces," *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, vol. 48, pp. 1454–1461, December 2001.

[121] M. T. Johnson, A. C. Lindgren, R. J. Povinelli, and X. Yuan, "Performance of nonlinear speech enhancement using phase space reconstruction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2003)*, vol. 1, (Hong Kong, China), pp. 920–923, April 2003.

[122] V. Pitsikalis and P. Maragos, "Speech analysis and feature extraction using chaotic models," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2002)*, vol. 1, (Orlando, FL, USA), pp. 533–536, May 2002.

[123] X. Liu, R. J. Povinelli, and M. T. Johnson, "Detecting determinism in speech phonemes," in *Proceedings of the IEEE 10th Digital Signal Processing Workshop and the 2nd Signal Processing Education Workshop*, (Pine Mountain, GA, USA), pp. 41–46, October 2002.

[124] D. E. Terez, "Robust pitch determination using nonlinear state-space embedding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2002)*, vol. 1, (Orlando, FL, USA), pp. 345–348, May 2002.

[125] M. T. Johnson, R. J. Povinelli, A. C. Lindgren, J. Ye, X. Liu, and K. M. Indrebo, "Time-domain isolated phoneme classification using reconstructed phase spaces," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 458–466, July 2005.

[126] I. Kokkinos and P. Maragos, "Nonlinear speech analysis using models for chaotic systems," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 1098–1109, November 2005.

[127] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, F. M. Roberts, and J. Ye, "Statistical models of reconstructed phase spaces for signal classification," *IEEE Transactions on Signal Processing*, vol. 54, pp. 2178–2186, June 2006.

[128] D. S. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data," *Physica D*, vol. 20, pp. 217–236, June-July 1986.

[129] T. Quatieri, *Discrete-time speech signal processing: principles and practice.* Upper Saddle River, NJ: Prentice Hall, 2002.

[130] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 201–212, June 1976.

[131] L. R. Rabiner and M. R. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, pp. 338–343, August 1977.

[132] L. J. Siegel and A. C. Bessey, "Voiced/unvoiced/mixed excitation classification of speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, pp. 451–460, June 1982.

[133] D. G. Childers, M. Hahn, and J. N. Larar, "Silent and voiced/unvoiced/mixed excitation (four-way) classification of speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 1771–1774, November 1989.

[134] R. Ahn and W. H. Holmes, "Harmonic-plus-noise decomposition and its application in voiced/unvoiced classification," in *Proceedings of the IEEE Conference on Speech and Image Technologies for Computing and Telecommunications*, vol. 2, (Brisbane, QLD, Australia), pp. 587–590, December 1997.

[135] Z. Xiong and T. Huang, "Boosting speech/non-speech classification using averaged Mel-frequency cepstrum," in *Proceedings of the Third IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, vol. 2532, (Hsinchu, Taiwan), pp. 573–580, December 2002.

[136] D. Arifianto and T. Kobayashi, "IFAS-based voiced/unvoiced classification of speech signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2003)*, vol. 1, (Hong Kong, China), pp. 812–815, April 2003.

[137] Z. Lachiri and N. Ellouze, "Speech classification in noisy environment using subband decomposition," in *Proceedings of the 7th International Symposium on Signal Processing and its Applications (ISSPA-2003)*, vol. 1, (Paris, France), pp. 409–412, July 2003.

[138] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 314–323, February 1988.

[139] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, pp. 293–309, February 1967.

[140] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 28, pp. 96–103, March 1985.

[141] D. S. Benincasa and M. I. Savic, "Voicing state determination of co-channel speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1998)*, vol. 2, (Seattle, WA, USA), pp. 1021–1024, May 1998.

[142] M. A. Zissman and C. J. Weinstein, "Automatic talker activity labeling for co-channel talker interference suppression," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1990)*, vol. 2, (Albuquerque, New Mexico), pp. 813–816, April 1990.

[143] J. McNames, "A nearest trajectory strategy for time series prediction," in *Proceedings of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*, (Leuven, Belgium), pp. 112–128, July 1998.

[144] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[145] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using multipitch analysis and iterative parameter estimation," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY, USA), pp. 83–86, October 2001.

[146] A. Bánhalmi, K. Kovács, A. Kocsor, and L. Tóth, "Fundamental frequency estimation by least-squares harmonic model fitting," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH-2005)*, (Lisbon, Portugal), pp. 305–308, September 2005.

[147] R. J. McAulay and T. F. Quatiery, "Speech analysis/synthesis based on sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, August 1986.

[148] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Transactions on Signal Processing*, vol. 48, pp. 338–352, February 2000.

[149] S. Haykin, *Adaptive filter theory.* Upper Saddle River, NJ: Prentice-Hall, 3rd ed., 1996.

[150] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, July 2006.

[151] P. C. Loizou, *Speech enhancement: theory and practice.* Boca Raton, FL: CRC Press, 2007.

[152] W. T. Hicks, B. Y. Smolenski, and R. E. Yantorno, "Testing the intelligibility of corrupted speech with an automated speech recognition system," in *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI-2003*, vol. 4, (Orlando, FL, USA), July 2003.

[153] I. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition," tech. rep., IDIAP Research Institute, Martigny, Switzerland, March 2005.

[154] L. Heck and M. Mao, "Automatic speech recognition of co-channel speech: integrated speaker and speech recognition approach," in *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP-2004)*, (Jeju Island, Korea), pp. 829–832, October 2004.

[155] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech and Language*, March 2008.

[156] M. P. Cooke and T. W. Lee, "Speech separation challenge." http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm, 2006.

[157] ITU-T, "Subjective performance assessment of telephone-band and wideband digital codecs," *ITU-T Recommendation P.830*, 1996.

[158] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T Recommendation P.835*, 2003.

[159] P. Mermelstein, "Evaluation of a segmental SNR measure as an indicator of the quality of ADPCM coded speech," *Journal of the Acoustical Society of America*, vol. 66, pp. 1664–1667, December 1979.

[160] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2299–2310, November 2007.

[161] D. P. W. Ellis, "Evaluating speech separation systems," in *Speech separation by humans and machines* (P. Divenyi, ed.), ch. 20, pp. 295–304, Springer US, 2004.

[162] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1766–1776, August 2007.

[163] S. R. Quackenbush, T. P. Barnwell, , and M. A. Clements, *Objective measures of speech quality.* Englewood Cliffs, NJ: Prentice Hall, 1st ed., 1988.

[164] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 67–72, February 1975.

[165] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II psychoacoustic model," *Journal of the Audio Engineering Society*, vol. 50, pp. 765–778, October 2002.

[166] D. H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1982)*, (Paris, France), p. 12781281, May 1982.

[167] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, pp. 819–829, June 1992.

[168] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1998)*, vol. 1, (Seattle, WA, USA), pp. 541–544, May 1998.

[169] ITU-T, "Objective quality measurement of telephone-band (300-3400 Hz) speech codecs," *ITU-T Recommendation P.861*, 1998.

[170] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, 2001.

[171] J. D. Gibson, B. Wei, and H. Dong, "Speech signal processing," in *Circuits, Signals, and Speech and Image Processing* (R. C. Dorf, ed.), ch. 15, pp. 793–799, Boca Raton, FL, USA: CRC Press, December 2005.

[172] R. G. Leonard and G. Doddington, "TIDIGITS." Linguistic Data Consortium, Philadelphia (http://www.ldc.upenn.edu), 1993.

[173] O. Omogbenigun, "A complete bundle of Matlab files for isolated word speech recognition." http://www.mathworks.com/matlabcentral/fileexchange/19298, 2008.