MONAURAL SPEECH ORGANIZATION AND SEGREGATION

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

The Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Guoning Hu, M.S.

* * * * *

The Ohio State University

2006

Dissertation Committee:	Approved by
Dr. DeLiang Wang, Adviser	
Dr. William Mitch Masters	______________________________
Dr. Eric Fosler-Lussier	Advisor Graduate Program in Biophysics

ABSTRACT

In a natural environment, speech often occurs simultaneously with acoustic interference. Many applications, such as automatic speech recognition and telecommunication, require an effective system that segregates speech from interference in the monaural (one-microphone) situation. While this task of monaural speech segregation has proven to be very challenging, human listeners show a remarkable ability to segregate an acoustic mixture and attend to a target sound, even with one ear. This perceptual process is called auditory scene analysis (ASA). Research in ASA has inspired considerable effort in constructing computational ASA (CASA) based on ASA principles. Current CASA systems, however, face a number of challenges in monaural speech segregation.

This dissertation presents a systematic and extensive effort in developing a CASA system for monaural speech segregation that addresses several major challenges. The proposed system consists of four stages: Peripheral analysis, feature extraction, segmentation, and grouping. In the first stage, the system 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, such as periodicity, amplitude modulation, onset and offset. In the third stage, the system segments an auditory scene based on a multiscale analysis of onset and offset. The last stage includes an iterative algorithm that simultaneously estimates the pitch of a target utterance and segregates the voiced target based on a pitch estimate. Finally, our system sequentially groups voiced and unvoiced portions of the target speech for non-speech interference, and this grouping task is performed using feature-based classification.

Systematic evaluation shows that the proposed system extracts a majority of target speech without including much interference. Extensive comparisons demonstrate that the system has substantially advanced the state-of-the-art performance in voiced speech segregation, and represents the first systematic study of unvoiced speech segregation.

Dedicated to my parents, Hu, Qingyun, and Duan, Yuying, and my wife, Xu, Lei

ACKNOWLEDGMENTS

First, I owe my deepest thanks to my advisor, Dr. DeLiang Wang, for guiding the research presented in this dissertation. His insights have greatly broadened my knowledge and deepened my scientific inquiry. I have learned from him how to conduct original research. I have also learned how to share the outcome with colleagues. Without his immense support, I would not have completed this work. Working with him is one of my most cherished experiences and will continue to benefit me in my future career.

Many thanks are due to Dr. William Mitch Masters, Dr. Eric Fosler-Lussier, Dr. Osamu Fujimura, and Dr. Ashok Krishnamurthy. Their advice has been of great help to my research. Their expertise in different scientific areas has helped me to explore research problems from different angles and to apply theories and ideas from different disciplines.

I wish to thank all the friends in the Perception and Neurodynamics Laboratory at The Ohio State University for much help and care. I will remember many pleasant times when we discussed problems, exchanged opinions, or even argued with each other. In particular, I want to thank Soundararajan Srinivasan. He has been a great source of advice, help, and fun in the past few years. I also want to thank Yang Shao and Yipeng Li. They are great fellow graduate students and have made my Ph.D study more

enjoyable. I want to extend many thanks to two former labmates, Drs. Mingyang Wu and Nicoleta Roman. They helped me a lot in my research, especially in the first few years. My time with them will always be in my memory.

I am grateful to my family and friends. They have given me great support all these years. Many thanks go to my wife Lei. Her love, care, and kindness have given me tremendous assistance and comfort.

I wish to acknowledge the financial support provided to me through an NSF grant (IIS-0081058), two AFOSR grants (F49620-01-1-0027 and FA9550-04-01-0117), and an AFRL grant (FA8750-04-1-0093) to Dr. DeLiang Wang.

VITA

June 29, 1974 ……………………………..Born in Jiangsu Province, P. R. China

July, 1996 ……………............................. B.S. Physics,
Nanjing University, Nanjing, P. R. China

June, 1999 ……………………………….. M.S. Physics,
Nanjing University, Nanjing, P. R. China

PUBLICATIONS

Journal Article

Guoning Hu and DeLiang Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks,* Vol. 15, pp. 1135-1150, 2004.

Conference Papers

Guoning Hu and DeLiang Wang, "Speech segregation based on pitch tracking and amplitude modulation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA'01*), pp. 79-82, 2001.

Guoning Hu and DeLiang Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP'02*), Vol. 1, pp. 553-556, 2002.

Guoning Hu and DeLiang Wang, "Monaural speech separation," *Advances in Neural Information Processing Systems* (*NIPS'02*), Vol. 15, pp. 1221-1228, 2003.

Guoning Hu and DeLiang Wang, "Separation of stop consonants," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP'03*), Vol. 2, pp. 749-752, 2003.

Guoning Hu and DeLiang Wang, "Segregation of stop consonants from acoustic interference," *Proc. IEEE International Workshop on Neural Networks for Signal Processing* (*NNSP'03)*, pp. 647-656, 2003.

Guoning Hu and DeLiang Wang, "Auditory segmentation based on event detection," *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

Guoning Hu and DeLiang Wang, "Separation of fricatives and affricates," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP'05*), Vol. 1, pp. 1101-1104, 2005.

FIELDS OF STUDY

Major Field: Biophysics

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF ACRONYMS

Acronym	Definition
ACF	Auto-correlation function
AM	Amplitude modulation
ASA	Auditory scene analysis
ASR	Automatic speech recognition
CASA	Computational auditory scene analysis
ERB	Equivalent rectangular bandwidth
F0	Fundamental frequency
GMM	Gaussian mixture model
LPC	Linear prediction coding
MFCC	Mel-frequency cepstral coefficient
MLP	Multilayer perceptron
SNR	Signal to noise ratio
SVM	Support vector machine
T-F	Time-frequency
T-segment	Time segment

LIST OF SYMBOLS

Symbol	Definition
$A(c, m, \tau)$	ACF of the filter response in channel c at frame m
$A(m, \tau)$	Summary correlogram at frame m
$\overline{A(c,m)}$	Average of $A(c, m, \tau)$ over τ
$A_E(c, m, \tau)$	ACF of the response envelope in channel c at frame m
$\overline{A_E(c,m)}$	Average of $A_E(c, m, \tau)$ over τ
a	Order of a gammatone filter
b	Equivalent rectangular bandwidth of a gammatone filter
$C(c, m)$	Cross-channel correlation of filter responses
$C_E(c, m)$	Cross-channel correlation of response envelopes
$C_V(c, t_1, t_2, s_c, s_t)$	Cross-channel correlation of smoothed intensities at scale (s_c, s_t) (see the definition of s_c and s_t)
c	Filter channel index
$E[k, l]$	Energy within the overlapping T-F region between an ideal segment k and an estimated segment l
$E_I[k]$	Energy of an ideal segment k
$E_S[l]$	Energy of an estimated segment l
E_C	Summated energy in all the T-F regions that are labeled as correct in segmentation evaluation
E_M	Summated energy in all the T-F regions that are labeled as mismatch in segmentation evaluation
E_N	Summated energy in all the T-F regions that are labeled as missing in segmentation evaluation
E_O	Summated energy in all the T-F regions that are labeled as over-segmented in segmentation evaluation
E_U	Summated energy in all the T-F regions that are labeled as under-segmented in segmentation evaluation
f	Frequency
f_c	Center frequency of a filter channel c

$\bar{f}(c.m)$	Average instantaneous frequency of the filter response within a T-F unit u_{cm}
$\bar{f}_E(c.m)$	Average instantaneous frequency of the response envelope within a T-F unit u_{cm}
$G(0, s_c)$	Gaussian function with mean 0 and standard deviation s_c
$g(f, t)$	Impulse response of a gammatone filter centered at frequency f
H_0	Hypothesis that a T-F region is target dominant
$H_{0,a}$	Hypothesis that a T-F region is dominated by an expanded obstruent
$H_{0,b}$	Hypothesis that a T-F region is dominated by a phoneme other than an expanded obstruent
H_1	Hypothesis that a T-F region is interference dominant
$h(s_t)$	A lowpass filter with passband [0, 1/s_t]
L	Set that contains the mask labels of individual T-F units
m	Time frame index
n	Discrete time
N_c	Number that defines the size of a neighborhood along frequency, which is used in labeling a T-F unit as either target or interference with the estimated pitch
N_m	Number that defines the size of a neighborhood along time, which is used in labeling a T-F unit as either target or interference with the estimated pitch
P	Probability
P_C	Correct percentage of segmentation, which is used in segmentation evaluation
$P(H_0 \| r_{cm}(\tau))$	Probability a T-F unit u_{cm} being target dominant give target pitch τ
P_{EL}	Percentage of energy loss, which is used in segregation evaluation
P_M	Percentage of missing, which is used in segmentation evaluation
P_N	Percentage of mismatch, which is used in segmentation evaluation
P_{NR}	Percentage of noise residue, which is used in segregation evaluation
P_O	Percentage of over-segmentation, which is used in segmentation evaluation
P_U	Percentage of under-segmentation, which is used in segmentation evaluation
p	Probability density
$r_{cm}(\tau)$	Feature set for labeling individual T-F units as target or interference
$r_I[k]$	An ideal segment k
$r_S[k]$	An estimated segment l

$SP_m(\tau)$	Summation of $P(H_0\|r_{cm}(\tau))$
s_c	Scale of the smoothing over frequency. s_c is the standard deviation of the Gaussian kernel for the smoothing over frequency
s_t	Scale of the smoothing over time. $[0, 1/s_t]$ is the passband of the lowpass filter for the smoothing over time
T_m	10 ms, the time shift between consecutive time frames
T_n	Sampling time
t	Continuous time
t_{ON}	Time position of an onset candidate
t_{OFF}	Offset time of an onset candidate
u_{cm}	A time-frequency unit in channel c at frame m
$v(c, t, s_c, s_t)$	Smoothed intensity of the filer response in channel c at time t. (s_c, s_t) is the scale of smoothing (see the definition of s_c and s_t)
$X(m)$	Cochleagram at frame m
$x(t)$	Input signal
$x(c,t)$	Response of a filter channel c to input signal
$x_E(c,t)$	Envelope of the filter response in channel c
λ	Parameter for thresholding in onset and offset detection
θ_A	Threshold for labeling the T-F units that correspond to unresolved harmonics as target or interference
θ_C	0.985, threshold for considering adjacent channels as high-correlated according to their cross-channel correlation
θ_E	Parameter for evaluating the segmentation performance
θ_P	Threshold for considering a T-F unit as supporting a pitch candidate in initial pitch estimation
θ_T	Threshold for labeling the T-F units that correspond to resolved harmonics as target or interference
$\theta_V (s_c, s_t)$	Threshold for connecting simultaneous onsets or offsets according to the corresponding cross-channel correlations of smoothed intensities at scale (s_c, s_t) (see the definition of s_c and s_t)
Γ	Plausible pitch range
σ	Standard deviation of the derivative of smoothed intensity
τ	Time delay in ACF
$\tau_S(m)$	Estimated pitch at frame m

CHAPTER 1

INTRODUCTION

1.1 The problem of speech segregation

One of my wife's favorite daily routines is to walk with me in the evening. We walk on the small meadow around our apartment and talk to each other. Meanwhile, we often hear people shouting, yelling, and laughing in a playground nearby. When my wife talks, what reaches my ear is the mixture of her voice and all other sounds. Although they often get very loud, I can still hear her very well and hardly feel being interrupted. My auditory system seems to have little trouble in separating her voice from other sounds.

Above is a typical situation we face daily: When someone is talking to us, what we hear is usually not just the utterance of that person, but a mixture with other interfering sounds. Interference can be any sound, such as wind noise, music, or another speech utterance. In such situations, we need to segregate the target utterance from the mixture and extract the information carried by the utterance.

People with normal hearing are excellent at segregating target speech from various types of interference. In most situations, we do not feel bothered by interfering sounds.

However, interference is a serious problem for machines and there is a great need for an effective speech segregation system for many applications. For example, the performance of automatic speech recognition (ASR) is severely degraded by interfering sounds (Lippmann, 1997; Cooke, 2003). An automatic speech recognizer would greatly benefit from a good speech segregation system. Such a system is also very helpful in telecommunication by improving the speech quality and reducing the cost of transmitting non-speech signal. In addition, interfering sound is a serious problem for hearing impaired people, even assisted by a hearing aid, when listening to a target speaker (Dillon, 2001). To help people with this problem, one needs to design a hearing aid that is able to extract target utterances from acoustic mixtures.

There have been extensive efforts to develop computational systems that automatically separate target sound or attenuate background interference. Many of the efforts have focused on the situation that target and interference come from different spatial locations and multiple microphones are available. In such a situation, one may attenuate interference using spatial filtering (Krim and Viberg, 1996; Brandstein and Ward, 2001; Gannot et al., 2001) that extracts signals from the target direction or cancels signals from the interfering directions. This approach, unfortunately, does not apply to situations when target and interference originate from the same location or only mono-recordings are available. Blind source separation using independent component analysis (ICA) (Bell and Sejnowski, 1995; Lee et al., 1999; Hyvärinen et al., 2001) separates mixtures into components that are statistically independent. Currently, ICA works well when sound sources are from different directions and the number of microphones is greater than or equal to the number of sources, but has difficulties in dealing with one-microphone recordings.

Applications such as telecommunication and audio information retrieval need a monaural (one microphone) solution for speech segregation. To find such a solution, one must consider the intrinsic properties of target or interference in order to distinguish and separate them. Various algorithms have been proposed for monaural speech enhancement (Lim, 1983; Benesty et al., 2005; Ephraim et al., 2005), and they are generally based on some analysis of speech or interference and subsequent speech amplification or noise reduction. For example, methods have been proposed to estimate the short-time spectra of interference and then attenuate interference accordingly (McAulay and Malpass, 1980; Ephraim and Malah, 1984; Virag, 1999; Martin, 2001), or to extract speech based on speech modeling (Paliwal and Basu, 1987; Hansen and Clements, 1991; Jensen and Hansen, 2001). Another way to deal with interference is to perform eigen-decomposition on an acoustic mixture and then apply subspace analysis to remove interference (Ephraim and van Trees, 1995; Rezayee and Gazor, 2001). Hidden Markov models have been used to model both speech and interference and then separate them (Ephraim et al., 1989; Varga and Moore, 1990; Sameti et al., 1998). These methods usually assume certain properties of interference and lack the capacity for dealing with general acoustic interference, because the variety of interference makes it very difficult to model and predict (Ephraim et al., 2005).

1.2 Computational auditory scene analysis

While monaural speech segregation by machines remains a great challenge, the human auditory system shows a remarkable capacity for this task. This observation has motivated a different approach to automatic monaural speech segregation – mimicking the auditory process of source separation.

The auditory segregation process is termed by Bregman as *auditory scene analysis* (ASA) (Bregman, 1990), which is considered to take place in two main stages. The first stage, called segmentation (Wang and Brown, 1999), decomposes the auditory scene into sensory elements (or segments), each of which should originate from a single source. The second stage is called grouping, where the segments that likely arise from the same source are grouped together. Segmentation and grouping are guided by perceptual principles, or ASA cues, that determine how the auditory scene is organized (Bregman, 1990; Darwin, 1997). These cues characterize intrinsic sound properties, including harmonicity, onset and offset, location, and prior knowledge of specific sounds.

Research in ASA has inspired considerable work to build computational ASA (CASA) systems for sound segregation (for reviews see Rosenthal and Okuno, 1998; Brown and Wang, 2005; Wang and Brown, 2006). Many CASA systems are developed for binaural situations (Nakatani and Okuno, 1999; Liu et al., 2001; Shamsoddini and Denbigh, 2001; Roman et al., 2003) based on the observation that sound sources often originate from different spatial locations and human listeners use location cues to help separating sounds from different directions (Bregman, 1990; Hawleyb et al., 2003). A typical binaural system obtains directional cues by comparing signals from two ears (or two microphones)

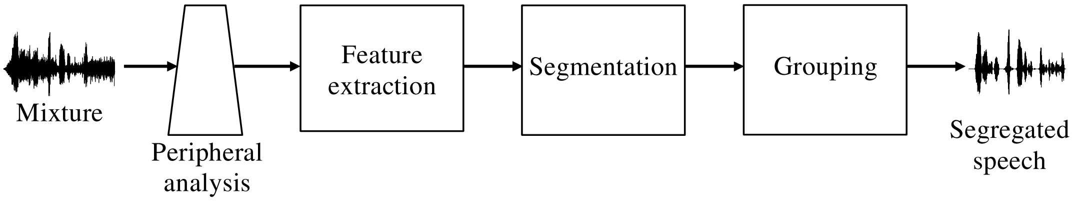

Figure 1.1. Schematic diagram of a typical CASA system

and then uses the directional cues to segregate target. In fact, binary CASA systems yield excellent result when the target and interference are from well-separated directions. However, they lack the capability to deal with the situation when sounds are from the same direction or only one-microphone recording is available.

Many studies have attempted to develop a CASA system for monaural segregation (Weintraub, 1985; Mellinger, 1992; Cooke, 1993; Brown and Cooke, 1994; Ellis, 1996; Wang and Brown, 1999). These systems aim to segregate target sound without making many assumptions about interference and tend to have a wider scope of applicability than speech enhancement methods. A typical CASA system for monaural segregation is shown in Figure 1.1. It has four stages: Peripheral analysis, feature extraction, segmentation, and grouping. The peripheral processing decomposes the auditory scene into a time-frequency (T-F) representation via bandpass filtering and time windowing. The second stage extracts auditory features corresponding to ASA cues, which are used in subsequent segmentation and grouping. In segmentation and grouping, the system generates segments for both target and interference and then groups the segments originating from the same source into a stream. The waveform of the segregated target

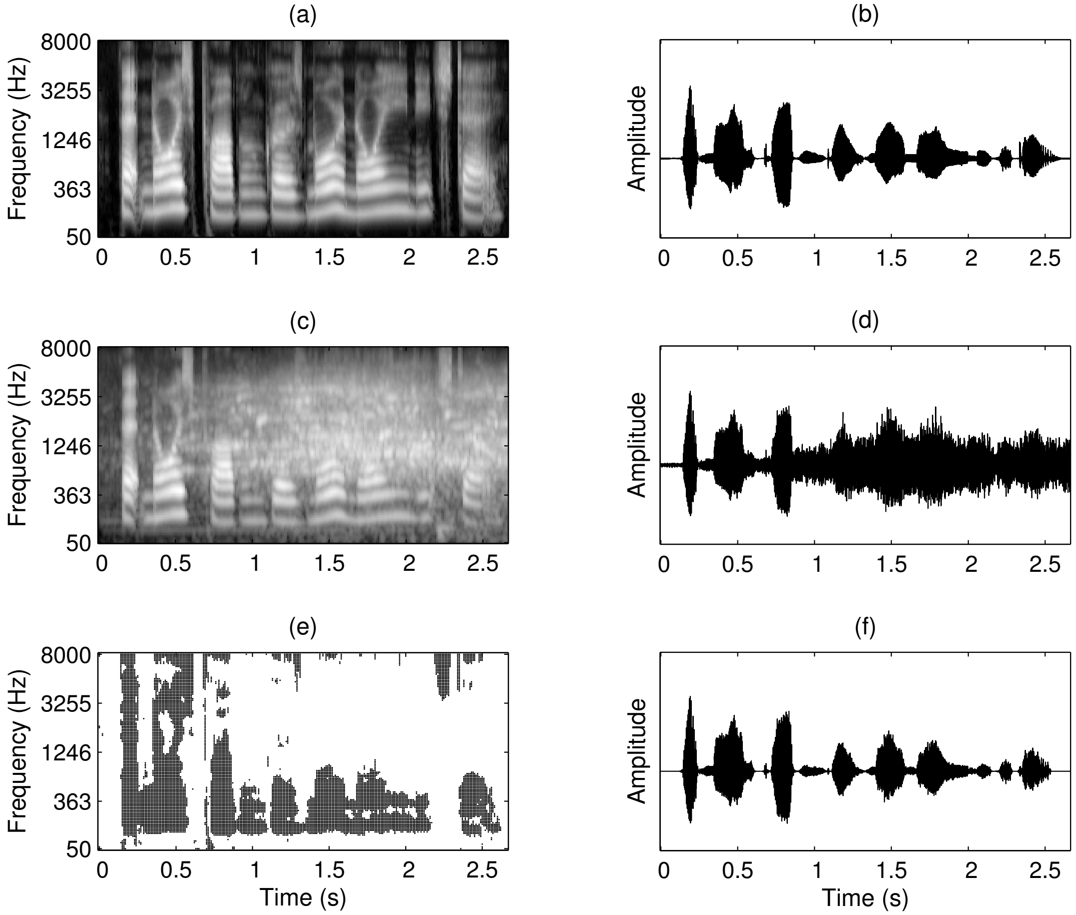

Figure 1.2. Signal representation. (a) T-F decomposion of a female utterance, “That noise problem grows more annoying each day.” (b) Waveform of the utterance. (c) T-F decomposition of the utterance mixed with a crowd noise. (d) Waveform of the mixture. (e) Target stream composed of all the T-F units (black regions) dominated by the target (ideal binary mask). (f) Waveform resynthesized from the target stream.

can then be resynthesized from the target stream (Weintraub, 1985; Brown and Cooke, 1994).

As an illustration, Figures 1.2(a) and 1.2(b) show a T-F decomposition and the waveform of a female utterance, “That noise problem grows more annoying each day,” from the TIMIT database (Garofolo et al., 1993). Figures 1.2(c) and 1.2(d) show a T-F decomposition and the waveform of the mixture of this utterance and crowd noise. The overall signal to noise ratio (SNR) of this mixture is 0 dB. For concision, we refer to this mixture as M1. Here the input is decomposed using a filterbank with 128 gammatone filters (Patterson et al., 1988) and 20-ms rectangular time windows with 10-ms window shift (see Section 3.1 for implementation details). We refer to a time window as a *time frame* and the T-F area in a filter channel and within a frame as a *T-F unit*. Figures 1.2(a) and 1.2(c) show the energy within each T-F unit, where a brighter pixel indicates stronger energy within the unit. Figure 1.2(e) shows an ideal target stream we aim to segregate, which contains all the T-F units with stronger target energy. To obtain this stream, a typical CASA system first merges neighboring T-F units into segments in the stage of segmentation. In this stage, the system generates segments for the target, shown as the contiguous black regions in the figure, as well as segments for the interference. Then in the stage of grouping, the system determines for each segment whether or not it belongs to the target and then groups it accordingly. Figure 1.2(f) shows the waveform resynthesized from the target stream in Figure 1.2(e).

1.3 Computational objective

A critical issue in developing a CASA system is to determine its computational goal (Marr, 1982). Ideally, we would like to obtain the exact target signal from the mixture. In fact, most speech enhancement systems try to obtain an estimate of the target that is as close to the true target as possible. However, in practice this goal is probably unrealistic due to the nonstationary nature of interference.

With the initial decomposition of an acoustic mixture into T-F units described in the previous section, we have suggested that the computational goal of a CASA system should be to retain the signals within the T-F units where target speech is more intense than interference and remove others (Hu and Wang, 2001; Hu and Wang, 2004a) (see also Roman et al., 2003). In other words, the goal is to identify a binary T-F mask, referred to as the *ideal binary mask*, where 1 indicates that target is stronger than interference in the corresponding T-F unit and 0 otherwise. Target speech can then be resynthesized with the ideal mask by retaining the signals within the T-F regions corresponding to 1's and rejecting the remaining signals. Figure 1.2(e) shows, in fact, the ideal binary mask for the mixture M1 in Figure 1.2(d). As shown in Figure 1.2(f), the speech resynthesized from the ideal binary mask is very similar to the original clean utterance in Figure 1.2(b).

The use of the ideal binary mask as our computational goal is supported by two important properties of the target utterance resynthesized from the mask. First, the interference in the resynthesized utterance is almost inaudible. Note that usually there is still some amount of interference in the T-F units labeled 1, i.e., with dominant target.

Therefore, the resynthesized utterance still contains interference. However, we usually cannot hear this part of interference because of the auditory masking phenomenon: Within a critical band, a weaker signal tends to be masked by a stronger one and therefore cannot be heard by a listener (Moore, 2003). Second, the resynthesized utterance gives excellent intelligibility unless SNR is extremely low. This is supported by considerable evidence from studies that use the utterances resynthesized from ideal binary masks for human speech intelligibility tests (Roman et al., 2003; Chang, 2004; Brungart et al., 2005). In addition, an ideal binary mask also yields excellent recognition performance in recent ASR studies (Cooke et al., 2001; Roman et al., 2003). The fundamental reason for such a phenomenon is that the speech energy and interference energy tend to distribute differently in the T-F domain. Therefore the T-F units with stronger target include substantial target energy in most situations. In addition, because speech signal has significant redundancy, the speech information is well preserved even when some speech signal is lost. For an extensive discussion on ideal binary mask as the computational goal for CASA, see (Wang, 2005).

1.4 Challenges

Natural speech contains both voiced and unvoiced portions (Stevens, 1998; Ladefoged, 2001). Voiced speech consists of portions that are mainly periodic (harmonic) or quasi-periodic. Harmonicity and temporal continuity are effective ASA cues for voiced speech segregation. Harmonicity refers to the fact that a periodic sound is composed of a group of harmonics, each of which is a frequency component whose frequency is an integer multiple of the fundamental frequency (F0). Temporal continuity refers to the fact that speech signal tends to last for a certain period of time and during this period the signal usually changes smoothly across time. Most previous CASA systems have focused on segregating voiced speech based on harmonicity and temporal continuity (Weintraub, 1985; Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999). Specifically, they aim to extract the signal that has periodicity similar to that of the target. Although previous CASA systems have made significant progresses in segregating voiced speech, they face the following acute challenges:

- *Robust pitch estimation*. In order to extract the signal that has similar periodicity to that of the target, one first needs to estimate the target pitch. Accurate pitch estimation is crucial for achieving good separation between target and interference. However, target pitch estimation is difficult when there is strong interference since interference corrupts target pitch information. Various methods for robust pitch estimation have been proposed (Hess, 1983; Wu et al., 2003; de Cheveigné, 2006). However, robust pitch estimation under low SNR situations remains a substantial challenge.
- *Segregation of voiced speech in the high-frequency range*. Another key problem of previous CASA systems is that they do not segregate voiced speech well in the high-frequency range. In the high-frequency range, harmonics are generally unresolved since the corresponding auditory filters have wide passbands that respond to multiple harmonics. Psychophysical evidence suggests that the human auditory system processes resolved and unresolved harmonics differently (Carlyon

and Shackleton, 1994; Bird and Darwin, 1998). Hence, one should carefully consider the distinctions between resolved and unresolved harmonics. Most previous CASA systems employ the same strategy to segregate all the harmonics, which works reasonably well for resolved harmonics but poorly for unresolved ones.

- *Segregation of unvoiced speech.* As previous CASA systems rely heavily on harmonicity to segregate target, they are not capable of segregating unvoiced portions of speech. Compared to voiced speech segregation, unvoiced speech segregation is a more difficult problem for two reasons. First, unvoiced speech lacks the harmonic cue and is often acoustically noise-like. Second, the energy of unvoiced speech is usually much weaker than that of voiced speech; as a result, unvoiced speech is more susceptible to interference. In fact, unvoiced speech segregation has not been systematically addressed at all in previous CASA systems.
- *Sequential grouping*. The task of sequential grouping is to identify the source of each sound at different time points and then group sounds from the same source across time. It is a very challenging task and little attention was given to it in previous CASA studies. By far, there is no general solution for this task.

1.5 Dissertation organization

This dissertation presents a systematic and extensive effort in developing a CASA system for monaural speech segregation. Our endeavor is directly targeted to the challenges listed in the previous section. The developed system estimates the ideal binary mask of a target utterance from an acoustic mixture, following the general stages of a typical CASA system as those shown in Figure 1.1. The remainder of the dissertation is organized as follows.

In Chapter 2, we first survey previous monaural CASA systems, including a system of our own on voiced speech segregation that employs different methods to segregate resolved and unresolved harmonics (Hu and Wang, 2004a). We also explain CASA challenges and discuss the approaches we take to address some of them. Finally, we give an overview of our proposed system.

In Chapter 3, we describe the auditory peripheral model of our system and the features extracted by the system. These features are response envelope, correlogram, onset and offset. They are used in the subsequent segmentation and grouping stages.

In Chapter 4, we describe an auditory segmentation algorithm based on a multiscale analysis of onset and offset. Onsets and offsets are important ASA cues (Bregman, 1990) because different sound sources in an acoustic environment seldom start and end at the same time. In addition, there is strong evidence for onset detection by auditory neurons (Pickles, 1988). There are several advantages of 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. Because onset and offset are cues common to all the sounds, our algorithm handles both voiced and unvoiced speech.

Systematical evaluation shows that this algorithm segments voiced and unvoiced speech effectively.

In Chapter 5, we describe an algorithm that estimates target pitch and performs voiced target segregation in an iterative loop. Specifically, our system yields an estimate of target pitch from the segregated voiced target, and then uses the newly estimated pitch to obtain a better estimate of voiced target, and so on. This algorithm combines features for both resolved and unresolved harmonics to estimate voiced target. The output of the algorithm is the estimated pitch contours and the associated T-F masks. Our evaluation shows that this algorithm performs substantially better than existing pitch determination algorithms as well as voiced speech segregation systems.

In Chapter 6, we describe the process of grouping target signals sequentially. Since some pitch contours obtained with the iterative algorithm may correspond to interference, a sequential grouping process is needed to group voiced target together. Furthermore, we need to group the segments dominated by unvoiced target with the segregated voiced target. In this study, we consider only non-speech interference. Consequently, this grouping task becomes a classification task, i.e., to distinguish T-F regions dominated by speech utterances from those dominated by non-speech signals. The proposed system performs this task through a feature-based classifier. Systematic evaluation shows that our system extracts a majority of target speech without including much interference. It performs substantially better than previous systems in segregating unvoiced speech.

Chapter 7 summarizes the proposed system and outlines the major contributions of this dissertation. It also discusses the CASA challenges that still remain and the possible directions of future research.

CHAPTER 2

BACKGROUND AND OVERVIEW

In this chapter, we first give a review of previous CASA systems for monaural segregation and a brief introduction of voiced and unvoiced speech. We then discuss the challenging problems in monaural speech segregation, and approaches that could deal with some of the challenges, which include our earlier effort on voiced speech segregation (Hu and Wang, 2004a). In this background, we provide an overview of the novel CASA system proposed in this dissertation for monaural speech segregation.

2.1 Previous CASA research on monaural speech segregation

Most CASA systems are developed for separating spoken utterances from interference, except for a few that aim to separate other types of sounds, such as Mellinger's system (Mellinger, 1992) for separating ensemble music, and the system of Li and Wang for separating singing voice from music (Li and Wang, 2005). This section surveys previous CASA systems for monaural speech segregation.

As mentioned in Section 1.2, most monaural CASA systems apply an initial analysis of input signal that mimics auditory peripheral processing. Typically, the input is decomposed in the frequency domain through a bank of filters that have similar frequency responses as those of cochlear filtering. Several studies report that such auditory-based front-ends are more robust than traditional Fourier-based analysis in the presence of background interference (Ghitza, 1994; Jankowski et al., 1995). Some systems (Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999) further process the output of each filter to simulate auditory nerve transduction (Meddis, 1988).

With the output from the initial analysis, CASA systems extract features representing ASA cues. Most previous CASA systems for monaural speech segregation focus on segregating voiced speech using periodicity as the primary cue. A well-established representation for periodicity analysis is a correlogram (Licklider, 1951; Lyon, 1984; Slaney and Lyons, 1990), which has been adopted by many CASA systems (Weintraub, 1985; Brown and Cooke, 1994; Wang and Brown, 1999). The correlogram is a running autocorrelation of the signal within a certain period of time in each filter channel. The periodicity of the signal is represented by the corresponding autocorrelation function (ACF).

An example of the correlogram is shown in Figure 2.1. The center panel shows the correlogram of the female utterance shown in Figure 1.2(b) within the time duration from 0.79 second to 0.81 second. Each curve in the panel is the ACF of a bandpassed response with the passband centered at a particular frequency (see Section 3.3 for details on the calculation of the ACFs). The peaks of these ACFs indicate the periodicity of the

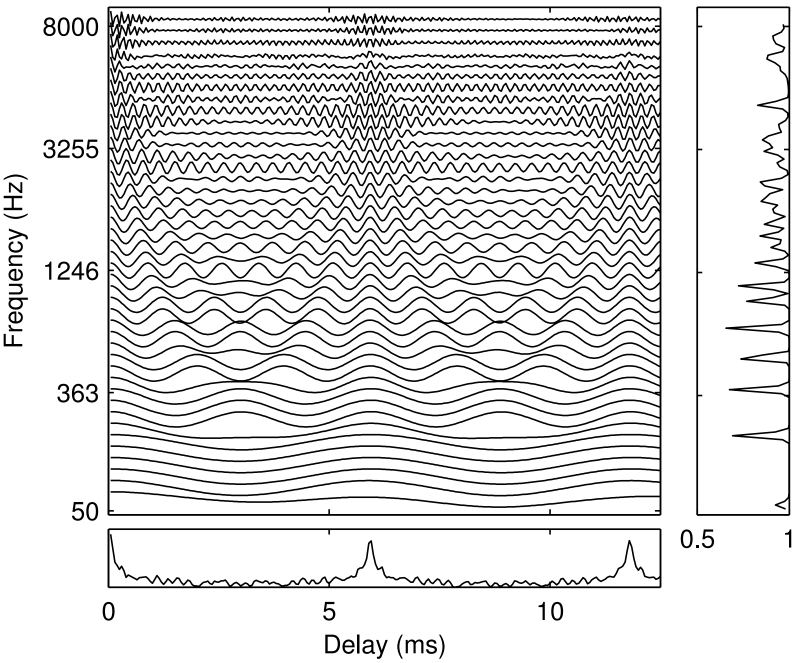

Figure 2.1. Correlogram at 0.8 second for the female utterance in Figure 1.2(b). For clarity, every third channel is displayed. The corresponding cross-channel correlation is given in the right panel, and the summary correlogram in the bottom panel.

corresponding signal. Since the utterance is voiced within this time duration, all the ACFs exhibit peaks at the delay of 5.87 ms, which corresponds to the pitch period of the utterance. This observation has motivated a well-established pitch determination algorithm, namely, summary correlogram (Licklider, 1951). The summary correlogram is the summation over frequency of all the responses in the correlogram, which is shown in the bottom panel. A dominant peak in the summary correlogram corresponds to the pitch period of the utterance.

Weintraub investigated the problem of separating two simultaneous speakers, one male and one female (Weintraub, 1985). He developed a system that first tracked the pitch contours of both utterances and then separated them by estimating the spectral amplitudes of each sound source based on periodicity and temporal continuity. His separation system contains four stages.

In the first stage, his system tracks two pitch contours, one for each speaker. In this stage, the system first uses a smoothed coincidence function, a version of autocorrelation, to capture periodicity as well as amplitude modulation (AM). Based on these functions, the system determines a dominant pitch at each 10-ms time frame. The dominant pitch may correspond to either speaker. It is assigned to the correct speaker using the knowledge that one speaker is male and the other is female. With the dominant pitch, the system tracks the pitch contour for each speaker based on the temporal continuity of pitch contours. The outcome of this stage is two estimated pitch periods at every time frame, no matter whether the corresponding utterance is voiced, unvoiced, or silent.

In the second stage, his system determines the actual state of the signal for each speaker. Weintraub considered seven states in this research: silence, periodic, nonperiodic, onset, offset, increasing-periodicity, and decreasing-periodicity. A Markov model (Rabiner, 1989; Rabiner and Juang, 1993) was trained to model the sequential relationship of these states for the situation when there are two simultaneous speakers, using the pitch and amplitude of an utterance as features. Based on the model, the system applies the Viterbi algorithm (Viterbi, 1967) to determine the state of each speaker at every time frame. The output of this stage gives a more accurate description of the signal within each frame. As a result, some estimated pitch periods are removed if the system decides that the corresponding signal is not periodic.

In the third stage, his system separates the two utterances by estimating the spectral amplitude of each sound source with the estimated pitch periods and states for each speaker. His system first has an initial estimate of each sound source. In particular, if at least one source is periodic, the system determines the best amplitude of each utterance according to the smoothed coincidence function. If both sources are nonperiodic, the system simply splits the energy evenly between them. Then the system improves the amplitude estimation iteratively by forcing good temporal continuity for each source. The estimated spectral amplitude is used to resynthesize the corresponding source in the last stage.

Weintraub tested his system with a corpus of two simultaneous speakers and obtained a certain level of separation. However, since the test corpus was also used in training, it is not clear how well his system works for a more general situation. In particular, in the pitch tracking stage, his system assigns pitch points to different speakers based on the fact that the pitch periods of a male speaker are usually much longer than those of a female speaker. This scheme cannot deal with the mixture of two male speakers or two female speakers. In addition, many of the processes in his system may not be able to handle general interference since they were developed to handle acoustic mixtures of two utterances.

Nevertheless, Weintraub made several important contributions in his dissertation. First, he proposed using temporal continuity to improve the estimation of target pitch and target utterance. Second, he developed a method for resynthesizing the segregated signal from the initial T-F decomposition of input signal that mimics cochlear filtering. This resynthesis method has been adopted in several other CASA systems (Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004a) and will be employed in this dissertation as well.

Subsequently, Cooke made the first systematic effort to segregate speech from various types of interference (Cooke, 1993). His system separates different sound sources as follows. After cochlear filtering, it computes the instantaneous frequency of each filter response. Then at every time step, his system combines neighboring channels into place groups such that each place group includes adjacent filter channels where the instantaneous frequencies change smoothly across these channels. An obtained place group tends to correspond to a single harmonic or several harmonics of a single source. Place groups at consecutive time steps are connected to form segments, referred to as *strands* by Cooke. His system then combines overlapping strands into groups if they have similar periods or similar AM patterns. A pitch contour is then obtained for each group, and groups with similar pitch contours are put into the same stream. Each stream corresponds to a separated source.

Cooke collected a test corpus to evaluate his system. This corpus contains 100 samples composed of 10 target utterances mixed with 10 intrusions. Every target utterance is totally voiced and has only one pitch contour. The intrusions have a considerable variety; specifically they are: N0 – 1 kHz pure tone, N1 – white noise, N2 – noise bursts, N3 – “cocktail party” noise, N4 – rock music, N5 – siren, N6 – trill telephone, N7 – female speech, N8 – male speech, and N9 – female speech. This corpus has been subsequently used in other studies for evaluating voiced speech segregation (Cooke, 1993; Brown and Cooke, 1994; Ellis, 1996; Wang and Brown, 1999; Drake, 2001).

To evaluate the segregation performance, Cooke computed the strands of a clean utterance and compared them with the obtained stream. On average, the target segregated by his system includes more than 70% of target strands. In addition, about 10% of the segregated target corresponds to interference. Cooke's system segregates voiced speech in the low-frequency range much better than in the high-frequency range. In fact, it fails to recover many target strands in the high-frequency range even when there is no interference.

Extending Cooke's system, Brown and Cooke proposed a model for monaural speech segregation (Brown and Cooke, 1994). Their model aims to estimate a binary T-F mask for target; the target utterance can be resynthesized from the mask by adopting Weintraub's resynthesis method (Weintraub, 1985). Their system differs from Cooke's system in several aspects. Specifically, their system computes the correlogram to represent periodicity. In segmentation their system merges neighboring filter channels based on cross-channel correlation, i.e., the cross correlation of ACFs in adjacent channels (see Section 3.3 for details in calculating cross-channel correlation). The cross-channel correlation compares the similarity of ACFs in adjacent channels. A higher cross-channel correlation indicates more similar ACFs. An example of cross-channel correlation is shown in the right panel of Figure 2.1. As shown in the figure, neighboring channels corresponding to the same frequency component, such as a harmonic of the utterance, have similar ACFs and therefore have high cross-channel correlations. By clustering based on cross-channel correlations, their system merges filter channels responding to the same frequency components. Merged channels in subsequent time frames are connected into segments, referred to as *auditory objects* by Brown and Cooke, based on temporal continuity. In grouping, their system estimates a pitch contour for each auditory object using the correlogram. Objects with similar contours are grouped into streams. In addition, they considered using common onset and offset as features to group objects.

Brown and Cooke's system was evaluated with Cooke's test corpus and it yields performance similar to that of Cooke's system. In particular, their system removes most interference energy from the target stream. However, a significant amount of target energy, especially in the high-frequency range, is missed from the segregated target. The evaluation also showed that their use of common onset and offset as additional features for grouping does not yield significant performance improvement.

Wang and Brown proposed a CASA system similar to Brown and Cooke's system. The major difference between these two systems is that Wang and Brown used a two-layer oscillator network for speech segregation (Wang and Brown, 1999). In the first layer, segments are formed based on cross-channel correlation and temporal continuity. Specifically, in each time frame, neighboring T-F units are merged into segments if their cross-channel correlation is higher than a threshold. In addition, T-F units in the same channel and at consecutive frames are merged if both units have cross-channel correlations higher than the threshold. In the second layer of Wang and Brown's network, segments are grouped into two streams, one for target and the other for interference, using an estimated global pitch in each time frame as follows. First, their system finds the longest segment as a seed segment. Then for each segment, their system determines at each frame whether or not the segment has a period similar to that of the seed segment, according to the global pitch period. If a segment has periods similar to those of the seed segment at more than half of their overlapping frames, it is grouped with the seed segment into a seed stream. Otherwise, this segment is grouped into a competing stream. Overall, Wang and Brown's system is computationally much simpler than Brown and Cooke's system, but yields a similar performance on Cooke's test corpus.

The above systems can be characterized as data-driven approaches, i.e., they rely mainly on features derived from the data, such as target pitch, onset, and offset, to separate sound sources, though in Weintraub's work, prior information of the temporal continuity of speech was used. Besides the data-driven CASA systems, some CASA systems depend primarily on prior information of sound sources to achieve sound separation, which may be characterized as model-based or top-down. Ellis developed a prediction-driven system for sound separation (Ellis, 1996). His system uses a world model to describe acoustic input. The world model includes three types of sound elements: Noise cloud, transient click, and harmonic sound. By decomposing the current input signal into different elements, the system generates predictions for future input. The optimal sound separation satisfies the condition that the model prediction best matches the actual input. Ellis tested his system on several intrusions, such as city-street ambient noise and a competing talker. The outcome of his system was assessed by human listeners. Ellis reported that his system extracts major sound objects from the mixtures, but the extracted sounds have notable distortions. Also, his evaluation is not extensive. Roweis proposed to separate two speech utterances using trained models of utterances (Roweis, 2001). He first trained a hidden Markov model (HMM) for each speaker and then combined these trained models factorially to form a model for the mixture of two speakers. His system separates the mixtures by fitting the separated utterances with the combined model. Roweis extended his system to deal with mixtures of speech and babble noise in a later study (Roweis, 2003). Jang and Lee proposed to decompose a mixture using *a priori* sets of basis functions of individual sound sources (Jang and Lee, 2003). The separation between different sources is achieved by maximizing the likelihood of the decomposition coefficients. The prior sets of basis functions and the probability density functions of the associated coefficients are determined using ICA on each source separately. These above model-based CASA systems rely heavily on *a priori* information of sound sources. As a result, they lack the capacity to deal with novel interference.

In addition to the above systems, there have been other CASA studies on monaural speech segregation. Many of these systems also explore the harmonicity cue. For example, Parson proposed to segregate two simultaneous utterances by tracking pitch contours of both utterances from the spectra of the mixture and selecting the harmonics of the desired voice according to the estimated pitch contours (Parsons, 1976). Several other studies use harmonicity to segregate two concurrent vowels (Meddis and Hewitt, 1992; Brown and Wang, 1997; de Cheveigné, 1997), aiming to model the behavior of human listeners in identifying double vowels (Assmann and Summerfield, 1990). Besides harmonicity, other ASA cues have also been utilized. Abe and Ando proposed to segregate two harmonic sounds based on common frequency modulation and common AM (Abe and Ando, 1998). Masuda-Katsuse and Kawahara proposed a CASA system that generates streams for different sound sources by tracking the changes in spectral shapes (Masuda-Katsuse and Kawahara, 1999). Unoki and Akagi proposed a system that extracts harmonic signals from noise using the following cues: Common onset and offset, gradualness of change, harmonicity, and changes occurring in the acoustic event (Unoki and Akagi, 1999). These systems have been tested with only certain types of interference and it is not clear how well they handle general interference.

2.2 Voiced and unvoiced speech

Natural speech contains both voiced and unvoiced portions (Stevens, 1998; Ladefoged, 2001). Different ASA cues are involved in segregating voiced and unvoiced speech since they have distinctive acoustic-phonetic properties. In this section, we give a brief introduction to these properties.

Voiced speech refers to the part of speech signal that is periodic (harmonic) or quasi-periodic. In spoken English, voiced speech includes all vowels, approximants, and nasals, and certain stops, fricatives, and affricates (Stevens, 1998; Ladefoged, 2001). It comprises a majority of spoken English. Unvoiced speech refers to the part that is mainly aperiodic. In spoken English, unvoiced speech comprises a subset of stops, fricatives, and affricates. These three consonant categories contain the following phonemes:

- Stops: /t/, /d/, /p/, /b/, /k/, and /g/.
- Fricatives: /s/, /z/, /f/, /v/, /ʃ/, /ʒ/, /θ/, /ð/, and /h/.
- Affricates: /tʃ/ and /dʒ/.

In phonetics, all these phonemes except /h/ are called obstruents. For the sake of concision, we refer to the above phonemes as expanded obstruents. Eight of expanded obstruents, /t/, /p/, /k/, /s/, /f/, /ʃ/, /θ/, and /tʃ/, are categorically unvoiced. In addition, /h/ may be pronounced either in the voiced or the unvoiced manner. Other phonemes are categorized as voiced, although in practice they often contain unvoiced sounds. Note that an affricate can be treated as a composite phoneme, composed of a stop followed by a fricative. Hence, stops and fricatives are the two main phonetic categories comprising unvoiced speech.

Dewey conducted an extensive analysis of the relative frequencies of individual phonemes in written English (Dewey, 1923) and concluded that unvoiced sounds account for 21.0% of the total phonemes. For spoken English, a similar analysis by French, Carter, and Koenig on 500 telephone conversations containing a total of about 80,000 words (Fletcher, 1953) concluded that unvoiced phonemes account for about 24.0% of the total phonemes. Another extensive, phonetically labeled corpus is the TIMIT database, which contains 6,300 sentences read by 630 different speakers from various dialect regions in America (Garofolo et al., 1993). Many of the same sentences in the TIMIT are read by multiple speakers and there are a total of 2,342 different sentences. We have performed an analysis of relative phoneme frequencies for distinct sentences in the TIMIT corpus, and found that unvoiced phonemes account for 23.1% of the total phonemes**.** Table 2.1

Phoneme types	**Conversational (Fletcher, 1953)**	**Written (Dewey, 1923)**	**TIMIT**
Voiced Stop	6.7	6.9	7.9
Unvoiced Stop	15.1	11.9	12.8
Voiced Fricative	7.5	9.5	7.7
Unvoiced Fricative	8.6	8.6	9.8
Voiced Affricate	0.3	0.4	0.6
Unvoiced Affricate	0.3	0.5	0.5
Total	38.5	37.8	39.3

Table 2.1. Occurrence percentages (by token count) of six consonant categories

shows the occurrence percentages of six phoneme categories from these studies. It is remarkable that these percentages are quite comparable despite the fact that written, read, and conversational speech are different in many ways. In particular, the total percentages of the six consonant categories are nearly the same for the three different kinds of speech. Note that the TIMIT database is constructed to be phonetic balanced and it is not as indicative as the data used in the other two studies.

A related question is the relative duration of unvoiced speech in spoken English. Unfortunately, the data reported on the telephone conversations (Fletcher, 1953) do not contain durational information. To get an estimate, we use the durations obtained from a phonetically transcribed subset of the Switchboard corpus (Greenberg et al., 1996), which also consists of phone conservations. The amount of labeled data in the switchboard corpus, i.e. seventy-two minutes of conversation, is much smaller than that in the

Phoneme types	Conversational (Fletcher, 1953; Greenberg et al., 1996)	TIMIT
Voiced Stop	5.6	5.2
Unvoiced Stop	16.2	12.9
Voiced Fricative	5.3	5.8
Unvoiced Fricative	9.6	12.0
Voiced Affricate	0.3	0.6
Unvoiced Affricate	0.4	0.7
Total	37.4	37.2

Table 2.2. Duration percentages of six consonant categories

telephone conservations (Fletcher, 1953). Hence we do not use the labeled Switchboard corpus for phoneme frequency analysis; instead we assign the median durations from the transcription to the occurrence frequency data in the telephone conservations (Fletcher, 1953) to estimate the relative durations of unvoiced sounds. Table 2.2 shows the resulting duration percentages of six phoneme categories, along with those from the TIMIT corpus. Once again, the percentages from the conversational speech are comparable to those from the read speech. In terms of overall time duration, unvoiced speech accounts for 26.2% of the total duration in phone conversations and 25.6% in the TIMIT corpus.

The above two tables show that unvoiced sounds account for more than 20% of spoken English in terms of both occurrence frequency and time duration. In addition, voiced obstruents are often not totally voiced. Therefore, unvoiced speech may occur more often than suggested by the data shown above.

Both voiced and unvoiced speech carry information crucial for human listeners to fully understand speech (Stevens, 1998), and therefore both need to be segregated from interference. As discussed in Section 1.4, unvoiced speech is much more difficult to segregate than voiced speech. In particular, unvoiced speech is often noise-like. It is difficult to distinguish some unvoiced speech from environmental noise. For example, the sound of a fricative /s/ is very similar to colored noise.

2.3 Preliminary research on voiced speech segregation

A common problem in early CASA systems that aim to segregate speech from various types of interference (Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999) is that they do not separate voiced speech well in the high-frequency range from broadband interference. This problem is closely related to the peripheral analysis of the input scene with a bank of auditory filters: The bandwidth of an auditory filter increases quasi-logarithmically with its center frequency. These filters are derived from psychophysical data to mimic cochlear filtering (Patterson et al., 1988). An important observation is that the structure of cochlear filtering limits the ability of human listeners to resolve harmonics (Plomp, 1964; Plomp and Mimpen, 1968; Carlyon and Shackleton, 1994). In the low-frequency range, harmonics are resolved since the corresponding auditory filters have narrow passbands that include only one harmonic. In the high-frequency range, harmonics are generally unresolved since the corresponding auditory filters have wide passbands that include multiple harmonics. A basic fact of acoustic interaction is that the filter responses to multiple harmonics are amplitude-modulated and the response

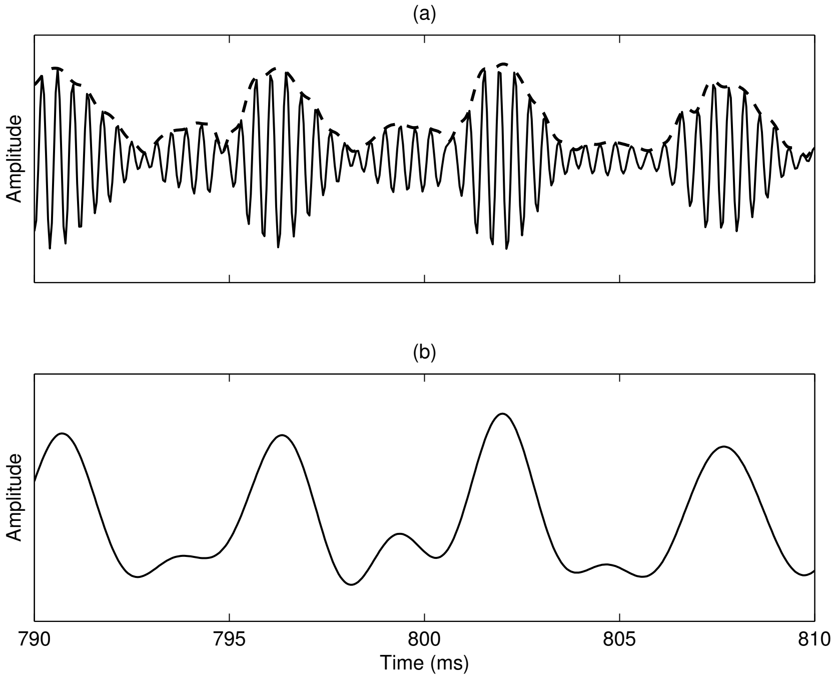

Figure 2.2. AM of a filter response. (a) Response (solid line) of a filter centered at 2.5 kHz to the female utterance shown in Figure 1.2(b) and the response envelope (dashed line). (b) Corresponding bandpass filtered envelope.

envelopes fluctuate at the F0 of target speech (Helmholtz, 1863). Figure 2.2(a) shows the response and the response envelope of a gammatone filter centered at 2.5 kHz within a time frame (from 0.79 s to 0.81 s). The input is the clean utterance shown in Figure 1.2(b). The response in Figure 2.2(a) is strongly amplitude-modulated, and its envelope fluctuates at the F0 rate at this frame.

Psychophysical evidence suggests that the human auditory system processes resolved and unresolved harmonics differently and AM is an important cue for unresolved

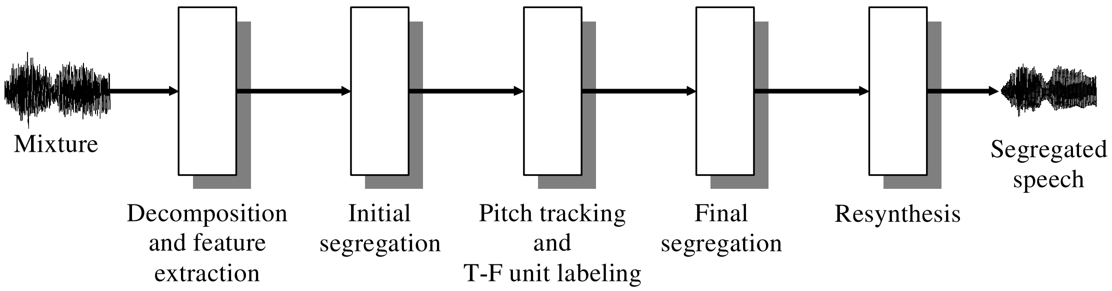
Mixture
Decomposition and feature extraction
Initial segregation
Pitch tracking and T-F unit labeling
Final segregation
Resynthesis
Segregated speech

Figure 2.3. Schematic diagram of a preliminary CASA system.

harmonics (Carlyon and Shackleton, 1994; Bird and Darwin, 1998). Based on this analysis, we have proposed a system that employs different methods to segregate resolved and unresolved harmonics of target speech (Hu and Wang, 2004a). More specifically, our system generates segments for resolved harmonics based on temporal continuity and cross-channel correlation, and groups them according to common periodicity among filter responses, similar to Brown and Cooke's system (Brown and Cooke, 1994) and Wang and Brown's system (Wang and Brown, 1999). Meanwhile, our system generates segments for unresolved harmonics based on common AM of filter responses in addition to temporal continuity. These segments are further grouped based on AM rates, which are obtained from the temporal fluctuations of the corresponding response envelopes. The AM cue was explored by Weintraub (1985) and Cooke (1993). Both of them used the AM cue primarily for grouping, whereas we used it to deal with unresolved harmonics in both segmentation and grouping.

Our system contains multiple stages, as shown in Figure 2.3. In the first stage, our system decomposes the input mixture into T-F units, the same as the Wang-Brown system (Wang and Brown, 1999). Then the following features are extracted: correlogram of filter responses, correlogram of response envelopes, cross-channel correlation, and dominant pitch at each time frame. In the second stage, T-F units that respond to resolved harmonics are merged into segments and these segments are then grouped into an initial foreground stream and a background stream based on the dominant pitch extracted in the previous stage. The processing in this stage is essentially the same as that of the Wang-Brown system. In the third stage, the pitch of target speech is estimated from the initial foreground stream**,** and is then used to label units as speech dominant or interference dominant. In the final segregation stage, according to unit labels, segments formed in the initial segregation stage are regrouped into foreground and background stream. This stage corrects some errors of initial grouping due to the inaccuracy of the dominant pitch. In addition, some T-F units are merged into segments that correspond to unresolved harmonics of target speech, and these segments are added to the foreground stream. Then the foreground stream expands to include neighboring T-F units labeled as speech dominant. Finally, a speech waveform is resynthesized from the resulting foreground stream using Weintraub’s method (Weintraub, 1985).

Our system was evaluated with Cooke’s test corpus. Figure 2.4 shows the SNR of the segregated speech for each intrusion averaged across 10 utterances. The SNR is computed using the resynthesized speech from the ideal binary mask as ground truth. The figure also shows the SNR of the original mixtures and the result from the Wang-Brown

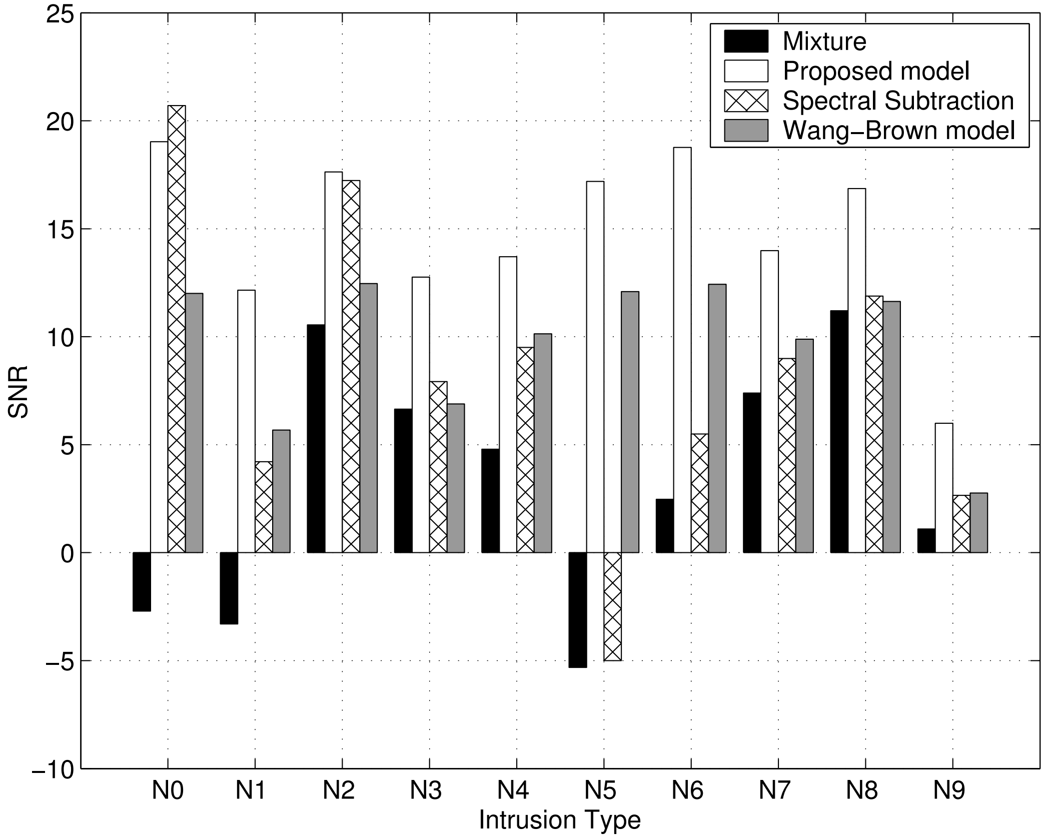

Figure 2.4. SNR results for segregated speech and original mixtures. White bars show the results from our previous system, gray bars those from the Wang-Brown system, cross bars those from the spectral subtraction method, and black bars those of original mixtures.

system. Our system yielded much better performance than the Wang-Brown system. In particular, it segregated much more target energy in the high-frequency range. We also compared our system with several speech enhancement techniques, including spectral subtraction (Huang et al., 2001) and comb filtering (Deller et al., 2000), and our system yielded the best performance (for more details, see Hu and Wang, 2004a).

This study shows that AM is an effective cue for segregating unresolved harmonics. A CASA system should use both periodicity and AM of filter responses to segregate the voiced portions of speech. In the above system, we first divide all the T-F units into two groups: those with high cross-channel correlations are considered as responding to resolved harmonics and others responding to unresolved harmonics. T-F units of the first group are segregated using the periodicity cue and others are segregated using the AM cue. Though the above method is simple and intuitive, it may not be the optimal way to utilize these cues. Later in this dissertation, we explore supervised learning to determine the optimal integration of these cues for voiced speech segregation.

2.4 Challenges and proposed strategies

As discussed in Section 2.1, model-based CASA systems rely on prior information or specific models of sound sources to achieve sound separation. However, in practice, interference is generally unknown or unpredictable. Therefore, these CASA systems are limited by an inability to model various types of interference. Although feature-based models have been proposed to deal with general interference (Weintraub, 1985; Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999), they face the following challenges: voiced speech segregation in the high-frequency range, robust pitch estimation, unvoiced speech segregation, and sequential grouping. Our previous system (Hu and Wang, 2004a) presented in the previous section has addressed the problem of segregating voiced speech in the high-frequency range. In this section, we discuss the remaining challenges and ideas that would help to deal with these challenges.

A. *Pitch estimation*

As discussed in Section 1.4, voiced speech is periodic or quasi-periodic. Harmonicity has proven to be a effective cue for segregating voiced target speech (Weintraub, 1985; Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004a). As shown by Hu and Wang (2004a), accurate pitch information helps to greatly improve the segregation result.

In many situations one needs to estimate target pitch from the mixture, which is a difficult task because the interference often corrupts target pitch information. Various methods for robust pitch estimation have been proposed (Hess, 1983; Wu et al., 2003; de Cheveigné, 2006), including several CASA systems (Weintraub, 1985; Brown and Cooke, 1994; Hu and Wang, 2004a). However, robust pitch estimation under low SNR situations still poses a substantial challenge. Since the difficulty of robust pitch estimation stems from interfering sounds, it is desirable to remove or attenuate interference before pitch estimation (Weintraub, 1985; Rouat et al., 1997; Wu et al., 2003). As a result, the problem of pitch estimation for sound separation becomes a "chicken and egg" problem: We want to segregate speech or remove interference using target pitch, but before estimating target pitch, we want to have speech segregated or interference attenuated (de Cheveigné, 2006).

We believe that a key to resolve the above dilemma is the observation that one does not need the entire target signal to estimate target pitch. Usually, several harmonics are sufficient for pitch estimation. Conversely, without perfect target pitch, one is still able to segregate some target. Therefore, we suggest a strategy that estimates target pitch and segregates targets in tandem. The idea is that we first obtain a rough estimate of target pitch, then use this estimate to segregate target. With the segregated target, we can generate a better estimate of target pitch, then a better segregation of target with better pitch information, and so on. This algorithm performs in an iterative manner. In each iteration, the algorithm estimates target pitch from the segregated target and then updates the segregated target with the current pitch estimate. We achieve *both* pitch estimation and speech segregation simultaneously when the iterative process converges. This iterative pitch estimation and speech segregation were present in rudimentary form in our previous system (Hu and Wang, 2004a) in which two iterations were used for pitch estimation and target segregation. In this dissertation, we develop this idea fully.

B. Segmentation of unvoiced speech

Previous CASA systems that aim to segregate speech rely mainly on harmonicity and therefore cannot handle unvoiced speech. As discussed in Section 2.2, unvoiced speech segregation must be addressed since unvoiced portions are essential for human listeners to understand speech. Yet, no systematic method has been proposed to either segment or group unvoiced speech. We discuss the general problem of segmentation, including unvoiced speech segmentation, in this subsection, and unvoiced speech grouping in the next subsection.

Segmentation is recognized as an important conceptual stage in ASA. A segment as a region of T-F units contains more global information about the source, such as spectral and temporal envelope, than that provided by individual T-F units. Such global information could be the key for distinguishing sounds from different sources. One may skip the stage of segmentation by grouping individual T-F units directly. However, grouping based on local information is unlikely very robust. We believe that auditory segmentation provides a useful foundation for grouping and is essential for successful CASA.

Previous CASA systems generally form segments according to two assumptions (Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004a). First, signals from the same source likely generate responses with similar temporal or periodic structure in neighboring auditory filters**.** Second, signals with good continuity in time likely originate from the same source. The first assumption works well for harmonic sounds, but not for noise-like signals, such as unvoiced speech. The second assumption is problematic when target and interference have significant overlap in time.

From a computational standpoint, auditory segmentation is analogous to image segmentation, which has been extensively studied in computer vision (Forsyth and Ponce, 2002). In image segmentation, the main task is to find bounding contours of visual objects. These contours correspond to sudden changes of certain local image properties, such as luminance and color. In auditory segmentation, the corresponding task is to find onsets and offsets of individual auditory events, which correspond to sudden changes of acoustic energy. Therefore we approach the problem of auditory segmentation based on onset and offset analysis of auditory events (Hu and Wang, 2004b). As discussed in

Section 1.5, onsets and offsets are important ASA cues that can be used to segment both voiced and unvoiced speech.

C. Sequential grouping

As mentioned in Section 1.4, the task of sequential grouping is to group the T-F regions corresponding to the same sound source across time. Temporal continuity is an effective cue for grouping T-F regions neighboring in time. However, it cannot handle T-F regions that do not overlap in time. Sequential grouping of such T-F regions is a very challenging problem. In CASA research, little attention was given to the problem of sequential grouping until recently. Barker et al. proposed to organize target segments from a mixture of speech and factory noise based on recognizing the phonetic content of the corresponding speech utterance (Barker et al., 2005). An alternative approach proposed by Shao and Wang groups temporal segments from a mixture of two utterances by recognizing the speaker of each T-F segment (Shao and Wang, 2005). Both studies targeted a specific type of interference and obtained some success. However, the general problem of sequential grouping is not solved.

A systematic study of sequential grouping is beyond the scope of this dissertation. Instead of finding a general solution, we focus on a situation that is common in practice, i.e., when speech signal is corrupted by non-speech interference. In such a situation, we may formulate the problem of sequential grouping as a classification task, i.e., to classify T-F regions as speech or interference. A T-F region dominated by the speech signal likely has acoustic-phonetic characteristics similar to those of clean speech, whereas a T-F region dominated by interference does not. Therefore, we investigate sequential grouping of T-F regions by classifying acoustic-phonetic features derived from each T-F region (Hu and Wang, 2005).

2.5 Overview of the proposed system

Based on the above discussion, we propose a system for monaural speech segregation that implements the strategies suggested in the previous section in order to address several CASA challenges. Our system adopts the typical stages of CASA, as shown in Figure 1.1. The details of the system are described in Chapter 3, 4, 5, and 6. Major innovations of our system are listed below.

In the stage of auditory segmentation, we apply a multiscale analysis, motivated by scale-space theory widely used in image segmentation (Romeny et al., 1997). The advantage of using a multiscale analysis is to provide different levels of detail for an auditory scene. Many acoustic signals consist of auditory events with varied sizes in the T-F domain. For example, speech signal consists of a series of phonemes that have different durations. With a multiscale analysis, we can detect and localize auditory events at proper scales. Our multiscale segmentation takes place in three steps. First, an auditory scene is smoothed to different degrees (scales). Second, the system detects onsets and offsets at certain scales, and forms segments by matching individual onset and offset fronts. Third, the system generates segments by integrating analysis at different scales.

In the stage of grouping, our system first segregates voiced speech with an iterative algorithm that estimates target utterance and target pitch simultaneously. We first generate a rough estimate of target utterance and target pitch, and then improve the estimation in an iterative manner until the estimates converges. This iterative algorithm yields several pitch contours and a T-F region corresponding to each pitch contour. Note that some pitch contours may correspond to interference. Then our system groups target-dominant T-F regions into a target stream as follows. It first classifies a T-F region associated with a pitch contour as speech or interference. Those T-F regions classified as speech are grouped into a target stream, which forms the segregated voiced target. Finally, our system segregates unvoiced target by identifying segments dominated by unvoiced speech. In the above two classification processes, our system distinguishes T-F regions dominated by speech from those dominated by interference with a Bayesian classifier using acoustic-phonetic features derived from individual T-F regions.

CHAPTER 3

AUDITORY PERIPHERY AND FEATURE EXTRACTION

This chapter describes the first two stages of our proposed system. In the first stage, our system decomposes the input in the T-F domain. In the second stage, it extracts following auditory features corresponding to ASA cues: envelope, correlogram, cross-channel correlation, onset, and offset. Most of the processes described here have been applied in previous CASA systems (see Wang, 2006, for a comprehensive review of auditory features and their extraction).

3.1 Auditory periphery

Our system first models cochlear filtering by decomposing the input in the frequency domain with a bank of gammatone filters. Gammatone filters are derived from psychophysical observations of the auditory periphery and the gammatone filterbank is a standard model of cochlear filtering (Patterson et al., 1988). The impulse response of a gammatone filter centered at frequency *f* is:

$$g(f,t)=\begin{cases} b^a t^{a-1} e^{-2\pi bt}\cos(2\pi ft) & t\ge 0 \\ 0 & \text{else} \end{cases} \tag{3.1}$$

where $a=4$ is the order of the filter, and b is the equivalent rectangular bandwidth (ERB). According to Glasberg and Moore (1990), the general relationship between ERB and the center frequency *f* is described by the following equation:

$$ERB(f)=24.7(4.3f/1000+1)\,. \tag{3.2}$$

The bandwidth increases as the center frequency f increases. For a fourth order gammatone filter, a bandwidth correction of 1.019 is suggested by Patterson et al. (1992):

$$b=1.019ERB(f) \tag{3.3}$$

In addition, the gain of the filter is adjusted by a factor $(2\pi b)^4/3$ so that the gain of the filter response at the center frequency is equal to 1.

Our periphery model uses a bank of 128 gammatone filters with center frequencies ranging from 50 Hz to 8000 Hz. The center frequencies of these filters are equally distributed on the ERB scale. Figure 3.1(a) shows the individual frequency responses of these filters in decibels. It is clear from the figure that filters with higher center frequencies respond to wider frequency ranges. Figure 3.1(b) shows the frequency response of the entire filterbank, i.e., the summation of the frequency responses of individual filters. This filterbank has a fairly flat frequency response within the range of the passband.

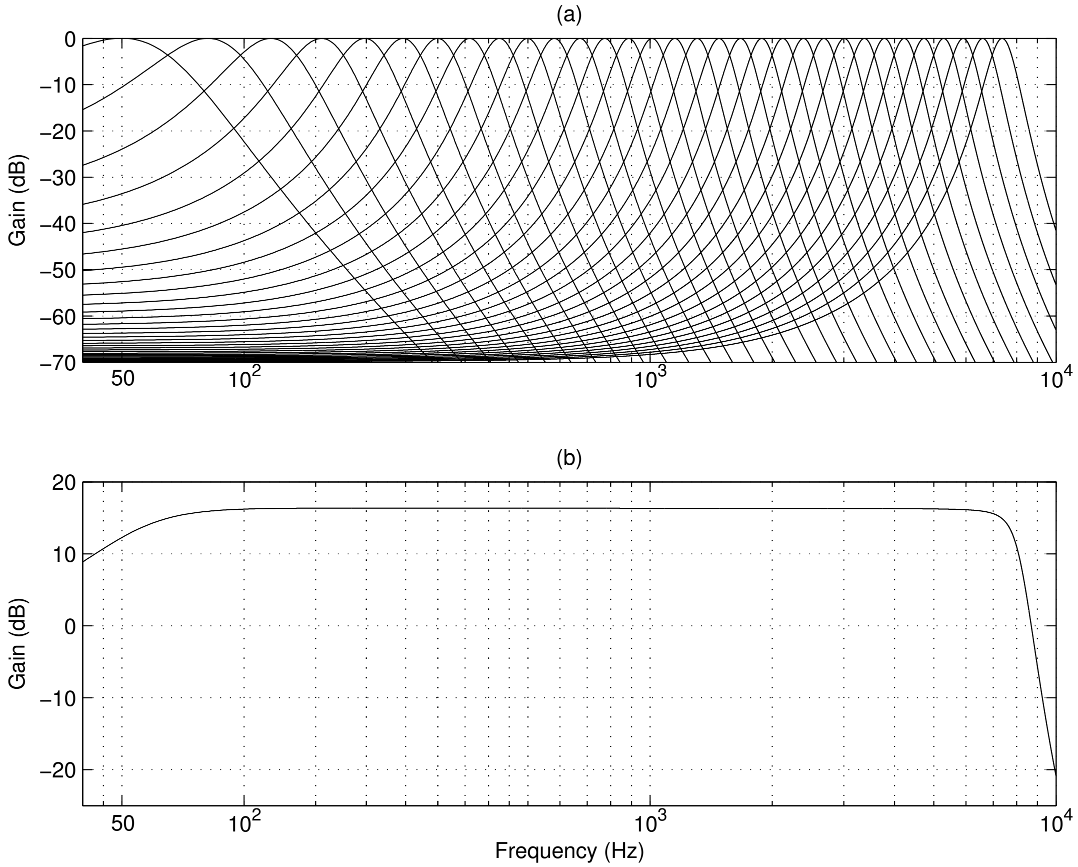

Figure 3.1. Frequency response of a gammatone filterbank with 128 channels centered from 50 Hz to 8000 Hz. (a) Frequency response of individual filters. For clarity, every fourth filter is shown in the figure. (b) Frequency response of the entire filterbank.

Let $x(t)$ be the input signal. The response from a filter channel c, $x(c, t)$, is

$$x(c,t) = x(t) * g(f_c, t) \tag{3.4}$$

where "$*$" denotes convolution, and f_c the center frequency of this filter. Because each filter introduces a delay to the filter response, the response is shifted backwards by $(a-1)/(2\pi b)$ to compensate for the filter delay (Holdsworth et al., 1988), which aligns the peak of the impulse response of each filter at time point 0.

3.2 Envelope extraction

As discussed in Section 2.3, when the input contains a periodic signal, some filter channels respond to multiple harmonics. Such a filter response is amplitude-modulated and the response envelope fluctuates at the F0 of the periodic signal (Helmholtz, 1863). Therefore, such a response envelope carries important AM information. A general way to obtain the response envelope is to perform half-wave rectification and then lowpass filtering. Since we are interested in the envelope fluctuations corresponding to target pitch, we perform bandpass filtering instead, where the passband corresponds to the plausible F0 range of target speech, i.e., [70 Hz, 400 Hz], the typical pitch range for adults (Nooteboom, 1997). The resulting bandpassed envelope in channel c is represented by $x_E(c, t)$.

As an illustration, Figure 2.2(a) shows the response and the response envelope of a gammatone filter centered at 2.5 kHz within a 20-ms time frame (from 790 ms to 810 ms). The input is the female utterance shown in Figure 1.2(b). This response is strongly amplitude-modulated, and the corresponding bandpass filtered envelope, shown in Figure 2.2(b), fluctuates at the F0 rate of the input at this time frame.

3.3 Correlogram and cross-channel correlation

In each filter channel, the output is divided into 20-ms time frames with 10-ms overlap between consecutive frames.

As discussed in Section 2.1, a correlogram is a commonly used periodicity representation, and it consists of ACFs of filter responses across all the filter channels. Let u_{cm} denote a T-F unit for frequency channel c and time frame m, the corresponding ACF of the filter response is given by

$$A(c,m,\tau)=\frac{\sum_n x(c,mT_m-nT_n)x(c,mT_m-nT_n-\tau T_n)}{\sqrt{\sum_n x^2(c,mT_m-nT_n)\sum_n x^2(c,mT_m-nT_n-\tau T_n)}} \tag{3.5}$$

Here, τ is the delay and n denotes discrete time. T_m = 10 ms is the time shift from one frame to the next and T_n is the sampling time. The above summation is over 20 ms, the length of a time frame. The periodicity of the filter response is indicated by the peaks in the ACF, and the corresponding delays are the periods. Here we calculate the ACF within the following range: $\tau T_n \in$ [0, 15 ms]. As a result, the ACFs are able to indicate any period within this range. Note that the plausible pitch range is [70 Hz, 400 Hz], corresponding to the period within [2.5 ms, 14.29 ms]. Equation (3.5) computes a normalized version of ACF. The purpose of normalization is to remove the influence of intensity fluctuations of filter response so that the resulting ACF represents the periodicity of the filter response more accurately.

As shown by Brown and Cooke (1994) and Wang and Brown (1999), cross-channel correlation measures the similarity between the responses of two adjacent filter channels and indicates whether the filters are responding to the same sound component or not. For a T-F unit u_{cm}, its cross-channel correlation with unit $u_{c+1,m}$ is given by

$$C(c,m) = \frac{\sum_\tau [A(c,m,\tau) - \overline{A(c,m)}][A(c+1,m,\tau) - \overline{A(c+1,m)}]}{\sqrt{\sum_\tau [A(c,m,\tau) - \overline{A(c,m)}]^2 \sum_\tau [A(c+1,m,\tau) - \overline{A(c+1,m)}]^2}} \tag{3.6}$$

where $\overline{A}$ denotes the average of A.

Similar to Equations (3.5) and (3.6), our system computes a normalized envelope autocorrelation:

$$A_E(c,m,\tau) = \frac{\sum_n x_E(c,mT_m - nT_n) x_E(c,mT_m - nT_n - \tau T_n)}{\sqrt{\sum_n x_E^2(c,mT_m - nT_n) \sum_n x_E^2(c,mT_m - nT_n - \tau T_n)}} \tag{3.7}$$

and cross-channel correlation of response envelopes,

$$C_E(c,m) = \frac{\sum_\tau [A_E(c,m,\tau) - \overline{A_E(c,m)}][A_E(c+1,m,\tau) - \overline{A_E(c+1,m)}]}{\sqrt{\sum_\tau [A_E(c,m,\tau) - \overline{A_E(c,m)}]^2 \sum_\tau [A_E(c+1,m,\tau) - \overline{A_E(c+1,m)}]^2}} \tag{3.8}$$

Figures 3.2(a) and 3.2(b) illustrate the correlogram and the envelope correlogram as well as the cross-channel correlation at a time frame (from 790 ms to 810 ms) for the female utterance shown in Figure 1.2(b). Figure 3.2(a) is the same as Figure 2.1. As shown in the figure, in the low-frequency range where harmonics are resolved, the autocorrelation of filter response generally reflects the periodicity of a single harmonic. Channels corresponding to the same harmonic have high cross-channel correlations. In the high-frequency range where harmonics are unresolved, the autocorrelation of filter response is amplitude-modulated. Adjacent channels in the high-frequency range are not

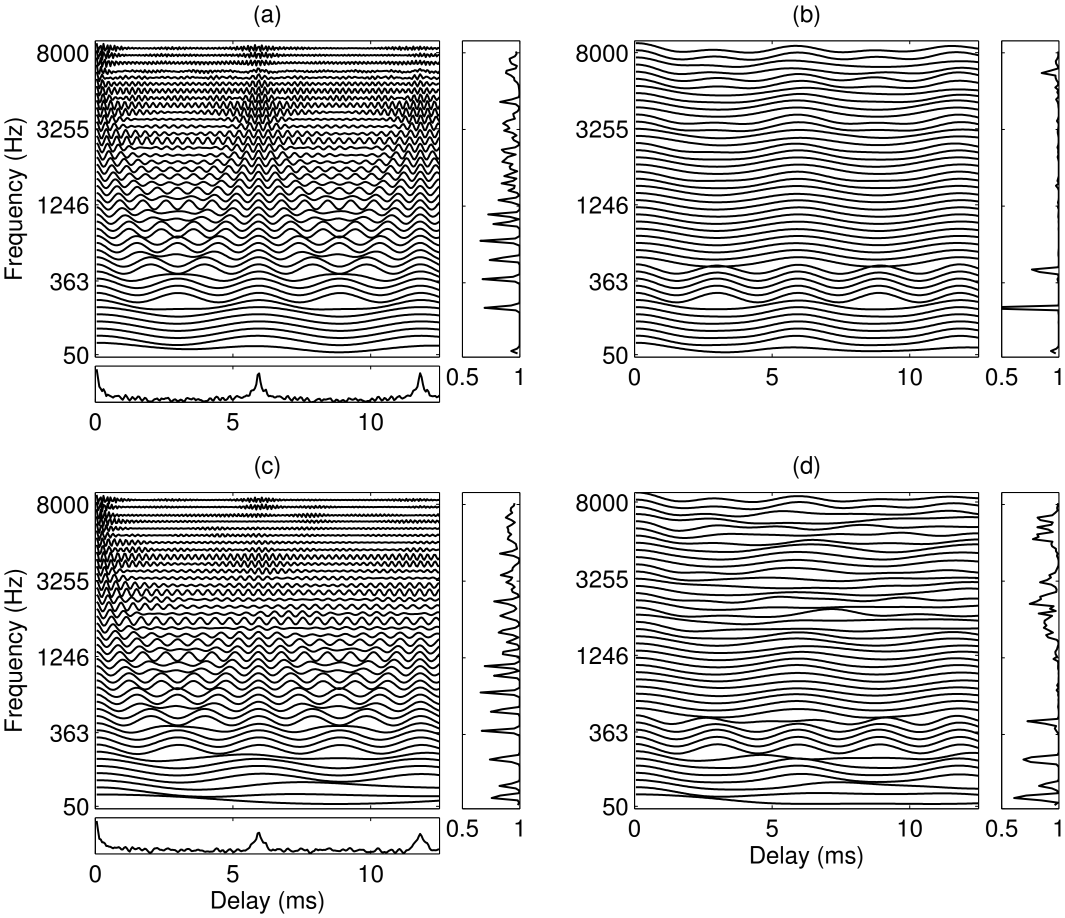

Figure 3.2. Auditory features. (a) Correlogram at a time frame (from 790 ms to 810 ms) for the female utterance show in Figure 1.2(b). For clarity, every third channel is displayed. The corresponding cross-channel correlation is given in the right panel, and the summary correlogram in the bottom panel. (b) Envelope correlogram for the utterance. The corresponding cross-channel envelope correlation is shown in the right panel. (c) Correlogram, cross-channel correlation, and summary correlogram for the mixture M1 shown in Figure 1.2(d). (d) Envelope correlogram and cross-channel envelope correlation for the mixture.

as highly correlated as those in the low-frequency range. On the other hand, the autocorrelations of the response envelopes in the high-frequency channels have similar fluctuation patterns that correspond to the pitch of the female utterance. Figures 3.2(c) and 3.3(d) show the correlograms for the mixture M1 shown in Figure 1.2(d). As shown in the figure, the interference corrupts the periodicity of ACFs, especially in the high-frequency channels. The bottom panels of Figures 3.2(a) and 3.2(c) show the summary correlogram, which are the summation of ACFs across all the channels. As discussed in Section 2.1, a summary correlogram exhibits peaks at delays corresponding to the pitch period of the utterances. We will come back to this point in Chapter 5 when we discuss the problem of pitch estimation.

3.4 Onset and offset

Onsets and offsets correspond to sudden amplitude increases and decreases, respectively. In particular, the positions of onsets and offsets are indicated by the most significant relative changes of intensities (Klapuri, 1999). Since the relative change of intensity is measured by the first-order derivative of its logarithm with respect to time, a standard way to identify such intensity changes is to find the peaks and valleys of the derivative. We calculate the intensity of filter response as the square of the response envelope, which is extracted using half-wave rectification and low-pass filtering. The lowpass filter used here is a filter with a 74.5-ms Kaiser window and a transition band from 30 Hz to 50 Hz.

Because of the intensity fluctuation within individual events, many peaks and valleys of the derivative do not correspond to real onsets and offsets. We smooth the intensity over time before onset and offset detection to reduce such fluctuations. In addition, we perform smoothing over frequency. Since an acoustic event tends to have synchronized onset and offset across frequency, smoothing helps to enhance such synchronies of onsets and offsets in neighboring frequency channels. This procedure is similar to the standard Canny edge detector in image processing (Canny, 1986). In this study, we smooth the intensity over time with a lowpass filter and over frequency with a Gaussian kernel. Let $v(c, t, 0, 0)$ denote the log-intensity at time t in filter channel c:

$$v(c,t,0,s_t) = v(c,t,0,0) * h(s_t) \tag{3.9}$$

$$v(c,t,s_c,s_t) = v(c,t,0,s_t) * G(0,s_c) \tag{3.10}$$

where $h(s_t)$ is a low-pass filter with passband $[0, 1/s_t]$ in Hz, and $G(0,s_c)$ is a Gaussian function with zero mean and standard deviation s_c. A lowpass filter with a 182.5-ms Kaiser window and a 10-Hz transition band is applied for smoothing over time. The parameter pair (s_c, s_t) indicates the degree of smoothing and is referred to as the 2-dimensional (2-D) scale. The larger the scale is, the smoother the intensity is. The smoothed intensities at different scales form the so-called scale space (Romeny et al., 1997).

As an example, Figure 3.3 shows the initial and smoothed intensities for the mixture M1 shown in Figure 1.2(d). Figure 3.3(a) shows the initial intensity. The smoothed intensities at three scales, (2, 1/14), (6, 1/14), and (6, 1/4) are shown in Figures 3.3(b), 3.3(c), and 3.3(d), respectively. To display more details, Figures 3.3(e), 3.3(f), 3.3(g), and

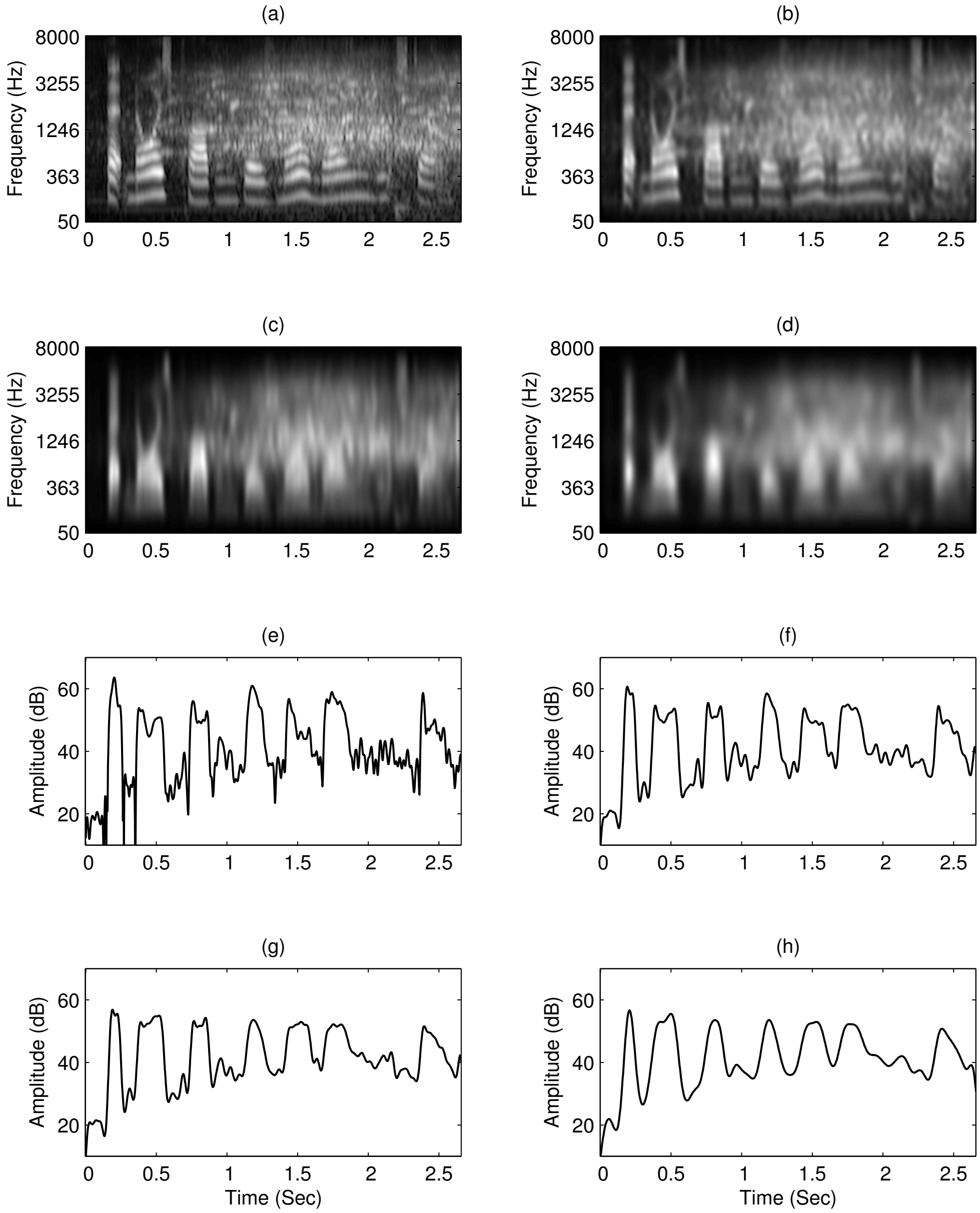

Figure 3.3. Smoothed intensity values at different scales. (a) Initial intensity for all the channels. (b) Smoothed intensity at the scale (2, 1/14). (c) Smoothed intensity at the scale (6, 1/14). (d) Smoothed intensity at the scale (6, 1/4). (e) Initial intensity for a channel centered at 600 Hz. (f) Smoothed intensity for the channel at the scale (2, 1/14). (g) Smoothed intensity for the channel at the scale (6, 1/14). (h) Smoothed intensity for the channel at the scale (6, 1/4). The input is the mixture M1 shown in Figure 1.2(d).

3.3(h) show the initial and smoothed intensities at these three scales for a single frequency channel centered at 600 Hz, respectively. As shown in the figure, the smoothing process gradually reduces the intensity fluctuations. Local details at onsets and offsets also become blurred, but the major intensity changes are preserved.

At a certain scale (s_c, s_t), onset and offset candidates are detected by marking peaks and valleys of the time derivative of the smoothed intensity:

$$\begin{aligned}\frac{d}{dt}v(c,t,s_c,s_t) &= \frac{d}{dt}[v(c,t,0,0)*h(s_t)*G(0,s_c)]\\ &= v(c,t,0,0)*[\frac{d}{dt}h(s_t)]*G(0,s_c)\end{aligned} \tag{3.11}$$

Onsets correspond to the peaks of the derivative above a certain threshold, and offsets the valleys below a certain threshold. The purpose of thresholding is to remove peaks and valleys corresponding to insignificant intensity fluctuations. Figure 3.4 shows an example of the above onset and offset detection on the mixture M1 shown in Figure 1.2(d) in a filter channel centered at 600 Hz. As shown in the figure, most detected onsets and offsets correspond to the true onsets and offsets of the female utterance in the mixture. Some true onsets and offsets of the female utterance are not detected due to the strong coarticulation between neighboring phonemes, influence of the interference, and the smearing effect of smoothing.

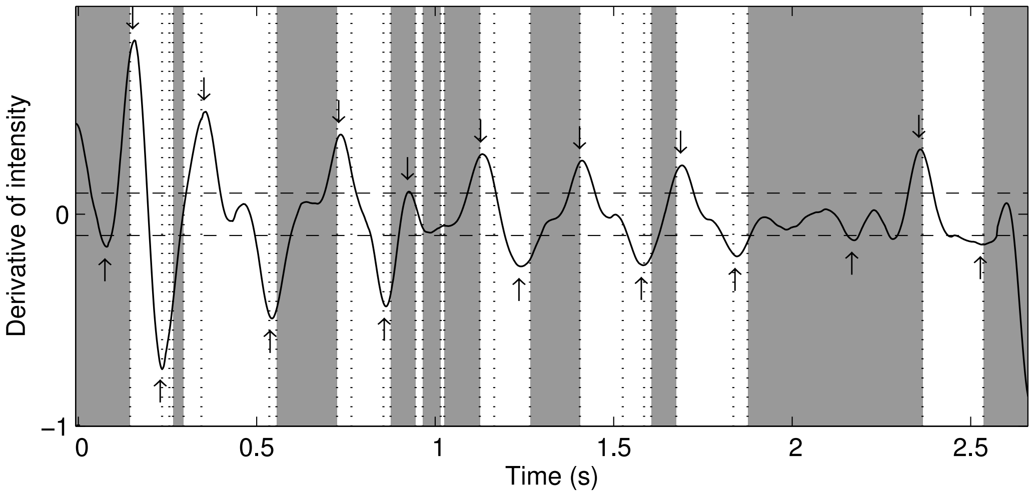

Figure 3.4. Onset and offset detection. The input is the response of a gammatone filter (centered at 600 Hz) to the mixture M1 shown in Figure 1.2(d). The scale is (6, 1/4). The corresponding intensity is shown in Figure 3.3(h). The threshold for onset detection is 0.1 and for offset detection is -0.1, indicated by the two dash lines. Detected onsets are marked by downward arrows and offsets by upward arrows. Vertical dotted lines indicate the boundaries of the auditory events of the female utterance (see Section 4.1 for details in determining these boundaries). The time durations where the intrusion is stronger are indicated by grey.

CHAPTER 4

AUDITORY SEGMENTATION

This chapter describes the second stage of our proposed system – auditory segmentation. In this chapter, we first discuss the computational goal of auditory segmentation. We then give a detailed description of this stage and a systematic evaluation of segmentation performance. Our preliminary studies of auditory segmentation have been published in the *Proceeding of ISCA Tutorial and Research Workshop on Statistical & Perceptual Audio Processing* (*SAPA*) (Hu and Wang, 2004b), and accepted by the *IEEE Transactions on Audio, Speech, and Language Processing* (Hu and Wang, 2006b).

4.1 Computational goal of auditory segmentation

The signal from one source, e.g., a speech utterance, contains a series of acoustic events, such as phonemes. One may consider that the computational goal of auditory segmentation is to find the onsets and offsets of individual acoustic events. In practice, one must limit the focus of CASA to the local acoustic environment of a listener; in other words, only acoustic events audible to a listener should be considered. Therefore, we set the task of auditory segmentation as to determine audible portions of each acoustic event instead of finding the physical onset and offset of the event. We refer to the collection of the audible portions of an acoustic event as an *auditory event*.

To determine the audibility of a sound, two perceptual effects need to be considered. First, a sound must be audible on its own, i.e. its intensity must exceed a certain level, referred to as the absolute threshold in a frequency band (Moore, 2003). Second, when there are multiple sounds in the same environment, a weaker sound tends to be masked by a stronger one (Moore, 2003). Hence, we consider a sound to be audible in a small local T-F region if it satisfies the following two criteria:

- Its intensity is above the absolute threshold.
- Its intensity is higher than the summated intensity of all other signals in that region. Accordingly, a local T-F region can only have one audible event.

The absolute threshold of a sound depends on frequency and varies among listeners (Moore, 2003). For young adults with normal hearing, the absolute threshold is about 15 dB sound pressure level (SPL) within the frequency range of [300 Hz, 10 kHz] (Killion, 1978). Therefore, we take 15 dB SPL as a constant absolute threshold for the sake of simplicity. To make this threshold meaningful, the overall intensity of any input signal is normalized to 60 dB SPL before being processed by our system.

By applying the above criteria to individual T-F units, we specify an auditory event as the collection of all the T-F units where an acoustic event is audible. Thus the computational goal of auditory segmentation is to identify each contiguous T-F region of an auditory event as a segment. We refer to such regions as ideal segments. This segmentation goal is consistent with our computational goal of segregation (see Section 1.3), both assigning a T-F unit only to a single sound source. This consistency is important since the purpose of auditory segmentation is to provide a solid foundation for subsequent grouping.

As a working definition, here we treat a phoneme, a basic phonetic unit of speech, as an acoustic event of speech. Based on this definition, we obtain the ideal segments of target utterances for evaluation. There are two issues for treating individual phonemes as events. First, two types of phonemes, stops and affricates, have clear boundaries between a closure and a subsequent burst in the middle of these phonemes. Therefore, we treat a closure in a stop or an affricate as an event on its own. This way, the acoustic signal within each event is generally stable. The second issue is that neighboring phonemes can be coarticulated, which may lead to unnatural boundaries between some consecutive ideal segments. In practice, these boundaries may not be detectable and the ideal segments separated by these boundaries are likely to be put together by a real segmentation system, creating a case of under-segmentation. Alternatively, one may define a syllable, a word, or even a whole utterance from the same speaker as an acoustic event. However, in such a definition many valid acoustic boundaries between phonemes are not taken into account. Consequently, some ideal segments are likely to be divided by a real segmentation system into smaller segments, creating a case of over-segmentation. We will come back to this issue in the evaluation section.

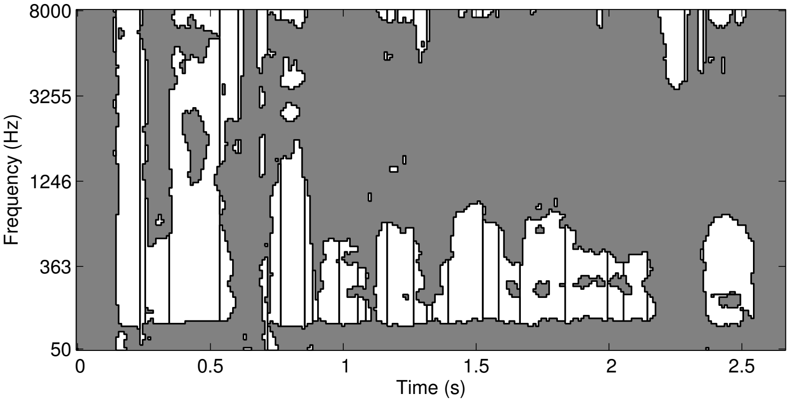

Figure 4.1. Bounding contours (solid lines) of ideal segments for the mixture M1 shown in Figure 1.2(d). The background is represented by gray.

As an example, Figure 4.1 shows the bounding contours of the ideal segments of the utterance (black line) for the mixture M1 shown in Figure 1.2(d). Note that the target utterance is from the TIMIT database, which gives all the phoneme boundaries. In the figure, gray regions form the background corresponding to the entire interference**.** Because the passbands of gammatone filters are relatively wide, particularly in the high-frequency range, adjacent harmonics may activate a number of adjacent filters. As a result, an ideal segment can combine several harmonics.

4.2 Overview of the segmentation procedure

As discussed in Section 2.4B, we have proposed to perform auditory segmentation via a multiscale analysis of event onset and offset. There are several advantages of using

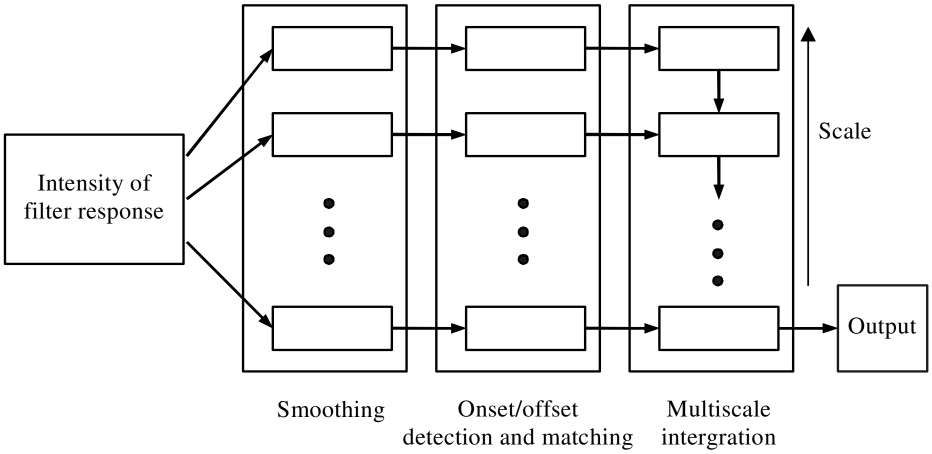
Intensity of filter response
Scale
Output
Smoothing
Onset/offset detection and matching
Multiscale intergration

Figure 4.2. Diagram of the proposed segmentation stage. In each stage, a rectangle represents the processing on the smoothed intensity at a particular scale. The scale increases from bottom to top.

onset and offset for segmentation. First, onsets and offsets, corresponding to sudden intensity changes, tend to delineate auditory events. Second, onset/offset times of a segment, which is a part of an event, usually vary smoothly across frequency. Such smooth variation is partly due to the fact that certain speech events, such as stops and fricatives, exhibit smooth-varying onset and offset boundaries in certain ranges of frequency. Also, the passbands of neighboring frequency channels have significant overlap. Hence, temporal alignment is an effective cue to group neighboring frequency channels.

Figure 4.2 shows the diagram of the segmentation stage. It has three steps: Smoothing, onset/offset detection, and multiscale integration. In the first step, our system smoothes the intensity of filter response as described in Section 3.4. In the second step, our system detects onsets and offsets in each filter channel and then merges simultaneous onsets and offsets in adjacent channels into onset and offset fronts. Onset and offset fronts are vertical contours connecting onset and offset candidates across frequency. Segments are obtained by matching individual onset and offset fronts. As a result of smoothing, event onsets and offsets of small T-F regions may be blurred at a larger (coarser) scale. Consequently, we may miss some true onsets and offset. On the other hand, at a smaller (finer) scale, the detection may be sensitive to insignificant intensity fluctuations within individual events. Consequently, false onsets and offsets may be generated and some ideal segments may be over-segmented. We find it difficult to obtain satisfactory segmentation with a single scale (see Section 4.6C for the detailed results of segmentation with a single scale). Our system handles this issue by integrating onset/offset information across different scales in an orderly manner in the last step, multiscale integration, which yields the final set of estimated segments. For details of the first step, see Section 3.4. A detailed description of the last two steps is given in the following sections.

4.3 Onset/offset detection and matching

A. Onset and offset detection

At a certain scale of smoothing (see Section 3.4), onset and offset candidates are detected by marking peaks and valleys of the time derivative of the smoothed log-intensity, as described in Section 3.4.

The offset time of each onset candidate is indicated by the following candidate. Similarly, the onset time of an offset candidate is indicated by the preceding candidate. To make this situation clear, we illustrate a simple case in Figure 4.3. In this figure, the horizontal dimension represents time and the vertical dimension represents frequency. Each square represents a T-F unit. Figure 4.3(a) shows the bounding contours of two events of a target sound source U and Figure 4.3(b) three events of an interfering sound source V. The ideal segments of target U when mixing U and V are shown in Figure 4.3(c). In this example, we assume that the onsets and offsets of event U1, U2, V1, and V3 are all perfectly detected, whereas onsets and offsets of event V2 are masked by event U1 and U2 and therefore not detected. Detected onset and offset candidates are shown in Figure 4.3(d). As shown in the figure, interfering event V1 starts and ends in the middle of target event U1. The T-F region in the top 6 channels of U1 has two parts. One part starts from the onset of U1 to the onset of V1 and the other from the offset of V1 to the offset of U1. Therefore, in these channels the offset time of an onset candidate of U1 is an onset candidate of V1, and the onset time of an offset candidate of U1 is an offset candidate of V1. On the other hand, the T-F region in the bottom 4 channels of event U1 starts at the onset and ends at the offset of this very event. The above example illustrates that because of interaction between different sound sources, the offset time of an onset candidate can be either an offset candidate from the same event, which suggests the ending of the current event, or an onset candidate from a different event, which suggests the beginning of a new event. Similarly, the onset time of an offset candidate can be either an onset candidate of the same event, or an offset candidate from a different event.

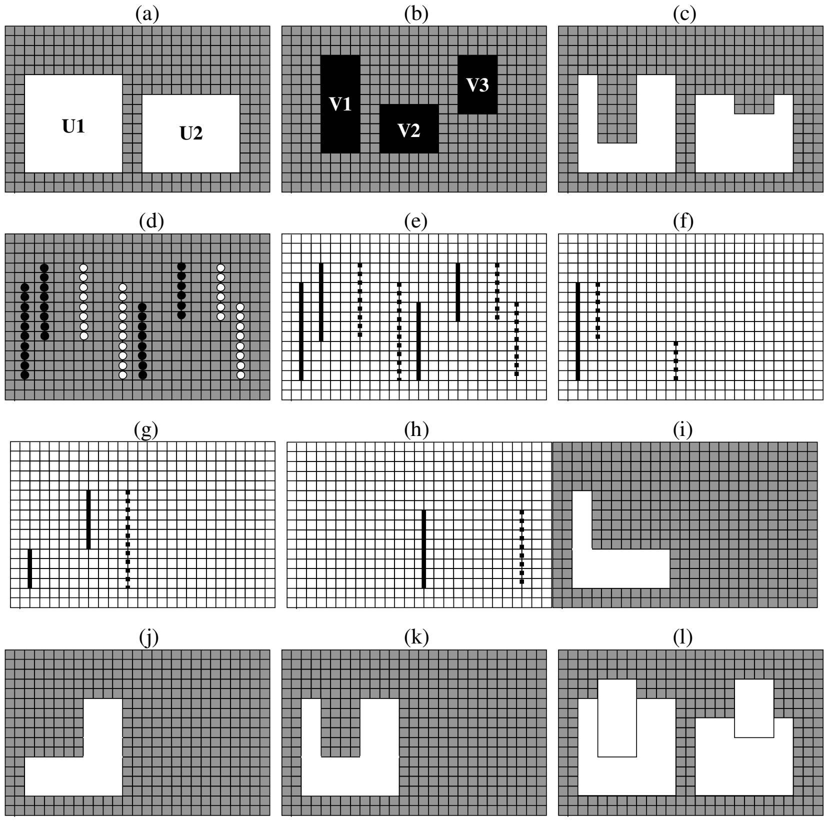

Figure 4.3. Illustration of the procedure of onset/offset detection and matching. In this figure, x-axis indicates time frames and y-axis indicates frequency channels. Each lattice point is a T-F unit. (a) Target source U with two events (white regions). (b) Interfering source V with three events (black regions). (c) Ideal segments for target U in the mixture of U and V. (d) Detected onset candidates (black dots) and offset candidates (white dots). (e) Obtained onset fronts (solid lines) and offset fronts (dotted lines). (f) Onset front for event U1 after onset and offset matching. The solid line is the onset front and dotted lines indicate the corresponding offset times. (g) Offset front for event U1 after onset and offset matching. The dotted line is the offset front and solid lines indicate the corresponding onset times. (h) Onset and offset fronts for event U2. The solid line is the onset front and the dotted line is the offset front. Note that the offset times of the onset front lie on the offset front and the onset times of the offset front lie on the onset front. (i) T-F region between the onset front of event U1 and the corresponding offset times. (j) T-F region between the offset front of event U1 and the corresponding onset times. (k) Obtained segment of event U1. (l) All obtained segments.

B. *Onset and offset front*

Since frequency components with close onset or offset times are likely to arise from the same source, our system connects simultaneous onsets and offsets, which likely correspond to the same event, into onset and offset fronts. There are usually some onset time shifts in adjacent channels in response to the same event. This is because the onset times of the components of an acoustic event may vary across frequency. Masking by interference may further shift detected onset and offset times. Therefore, we allow a tolerance interval when connecting onset/offset candidates in neighboring frequency channels. Specifically, we consider onset candidates in adjacent channels to be simultaneous if the distance between their onset times is shorter than the tolerance interval. The same is for offset candidates. This interval should not be too short; otherwise onsets (or offsets) from the same event will be prevented from joining together. On the other hand, an interval that is too long will connect some onsets from different events together. As found by Darwin (1984) and Turgeon et al. (2002), human listeners start to segregate two sounds when their onset times differ by 20~30 ms. Therefore, we select 20 ms as the tolerance interval.

Since events from different sources may either start or end at the same time, two simultaneous onset candidates may not correspond to the same event. Additional constraints are needed for connecting simultaneous onset candidates into onset fronts. Because it is unlikely that events from different sources *both* start and end at the same time, simultaneous onsets with simultaneous offset times likely correspond to the same event and are connected into onset fronts. Furthermore, signals in adjacent channels from the same source likely have similar responses or response envelopes. The similarity between the responses or response envelopes is measured by their correlations. For a pair of simultaneous onsets in channel *c* and *c*+1, let (t_1, t_2) be the overlapping duration from these onsets to their corresponding offset times. We use 3 correlation measures, $C(c, t_1, t_2)$, the similarity of responses, $C_E(c, t_1, t_2)$, the similarity of envelope fluctuations within the plausible pitch range, and $C_V(c, t_1, t_2, s_c, s_t)$, the similarity of smooth intensities, i.e., the low rate fluctuations of response envelopes. $C_V(c, t_1, t_2)$ is computed as:

$$C_V(c, t_1, t_2, s_c, s_t) = \sum\nolimits_{t=t_1}^{t_2} \hat{v}(c, t, s_c, s_t)\hat{v}(c+1, t, s_c, s_t) \tag{4.1}$$

where $\hat{v}$ indicates the normalized *v* with zero mean and unity variance within (t_1, t_2) and (s_c, s_t) is the scale. $C(c, t_1, t_2)$ and $C_E(c, t_1, t_2)$ can be computed in a similar manner. Since in the previous stage, we have calculated the cross-channel correlation of the ACFs of filter responses and response envelopes within individual T-F units, here we simply use the average of these cross-channel correlations within the duration (t_1, t_2). These two simultaneous onsets are connected if one of the measures is higher than a threshold, that is, $C(c, t_1, t_2) > 0.95$, $C_E(c, t_1, t_2) > 0.95$, or $C_V(c, t_1, t_2, s_c, s_t) > \theta_V(s_c, s_t)$, a scale-dependent threshold. Similarly, simultaneous offsets are connected into offset fronts if they have simultaneous onset times or the corresponding responses or response envelopes in adjacent channels are highly correlated. Among thus formed fronts, we discard those occupying fewer than 5 channels because they usually correspond to interfering noise. Figure 4.3(e) illustrates the onset and offset fronts of the mixture of source U and V.

C. *Onset and offset matching*

As discussed in Section 3.4, to remove peaks and valleys corresponding to insignificant intensity fluctuations, we use thresholding in our onset and offset detection. The advantage of thresholding is that most onset and offset candidates thus obtained do correspond to true onsets and offsets. However, some true onsets and offsets may not be detected, which causes two problems. First, some T-F regions corresponding to target events may not be included into any segment. Second, the offset time of an onset candidate may not be the correct one, and the same for the onset time of an offset candidate. As a result, some segment will contain T-F regions from different events. The first problem is addressed in the next step, multiscale integration. The second one is caused by mismatching between onset candidates and their offset times. Therefore, we re-estimate the offset times of each onset front and the onset times of each offset front by matching individual onset and offset fronts.

The basic observation is that since offset times of an onset front tend to be smooth across frequency channels, they likely match one or more fronts. These fronts may be an offset front from the same event, or an onset front from a different event. Therefore, we can correct some mismatching errors between onset candidates and their offset times by matching the offset times of an onset front with other fronts and update the offset times according to the matching fronts. In order to recover missing offsets due to thresholding, we apply a two-threshold scheme, as proposed by Canny (1986). We first use stringent thresholds to obtain onset and offset fronts, referred to as *reliable* fronts. Then in the same manner we use loose thresholds to obtain onset and offset fronts, referred to as *possible* fronts. These possible fronts are used to re-estimate the offset times of reliable onset fronts. Specifically, let $(t_{OFF}(c), t_{OFF}(c+1), \ldots, t_{OFF}(c+m-1))$ denote the offset times of a reliable onset front occupying m consecutive channels. Among all the possible fronts, we find the one that crosses the most of these offset times. Then the channels from c to $c+m-1$ occupied by this possible front are labeled as "matched", and the corresponding offset times of the reliable front are updated to those of the matching possible front. If all the channels from c to $c+m-1$ are labeled as matched, the matching procedure is finished. Otherwise, the process repeats for the remaining unmatched channels. Similarly, each reliable offset front is matched with possible fronts and the corresponding onset times are updated.

As an illustration, Figure 4.3(f) shows the onset front of target event U1 and the matching fronts. For each matching offset front, only the part in the "matched" channels is shown. This onset front matches two fronts; the corresponding offset front of the same event and the onset front of interfering event V1. Similarly, Figure 4.3(g) shows the offset front of U1 and the matching fronts. This offset front also matches two fronts, the corresponding onset front of the same event and the offset front of interfering event V1. In comparison, Figure 4.3(h) shows the onset front of target event U2, which is well matched by a single front, the offset front of the same event. Note the difference between event U1 and U2: Since interfering event V1 masks a significant part of target U1, the best matching front of the onset front of U1 is the onset front of V1; whereas since only a small part of target U2 is masked by interfering event V2, the best matching front of the onset front of U2 is the offset front of the same event.

4.4 Multiscale integration

Our system integrates onset and offset analysis at different scales. It starts at a coarse scale, i.e., generating reliable onset and offset fronts as described in Section 4.3. Then, at a finer (smaller) scale, it locates more accurate positions for each front and detects new fronts from the current background.

Specifically, let $t_{ON}(c)$ be the time position for an onset candidate in channel c on an onset front. Among the onset candidates in channel c at the current scale, our system first finds the one that is nearest to $t_{ON}(c)$. If the distance between this candidate and $t_{ON}(c)$ is smaller than 20 ms, we consider that this candidate corresponds to $t_{ON}(c)$ and update $t_{ON}(c)$ to the position of this candidate. In this manner, our system updates the position of the offset times of each onset front. The same is for each offset front. Note it is possible that some onset candidates in an onset front do not have corresponding candidates at the current scale. If such a candidate is at an edge channel of an onset front, i.e., it is in the first channel or the last channel of the onset front, this candidate is likely to be a spurious one. Therefore, our system iteratively removes such onset candidates until both candidates in the edge channels of any onset front have corresponding candidates at the current scale. Similarly, some spurious offset candidates are removed from offset fronts. Then our system re-estimates the onset times of each onset front by matching it with the current possible fronts, and the same for each offset front. In addition, our system generates new fronts within the current background, i.e., all the T-F regions that are not covered by the T-F regions between the onset times and the offset times of any front. A newly obtained onset front is connected with an existing onset front into one front if they occupy adjacent channels and the corresponding onset candidates and offset times in these channels are simultaneous. The same applies to offset fronts.

Finally, our system generates segments from the obtained fronts as follows. For each onset front, the T-F region between its onset candidates and the corresponding offset times form a segment. Similarly, a segment is formed for each offset front. Such formed segments may overlap with each other. There are two types of overlaps. First, segments corresponding to a same event overlap and they need to be merged into one segment. Figures 4.3(i) and 4.4(j) illustrate such a case. Second, segments correspond to different events overlap and the overlapping region needs to be assigned to one segment. The second situation occurs when a later event happens in the middle of an earlier event and generally the overlapping T-F region corresponds to the later event. These two types of overlapping can be well distinguished since in the first case, two overlapping segments tend to have some identical onsets or offsets in the overlapping region, whereas in the second case, two overlapping segments are unlikely to have a common onset or offset in any frequency channel. Therefore, our system deals with overlapping segments as follows. Two overlapping segments are merged into one segment if they have the same onset or offset in at least one channel; otherwise our system assigns the overlapping region to the segment starting later.

As an illustration, Figure 4.3(i) shows the obtained segment corresponding to the onset front of event U1, and Figure 4.3(j) that corresponding to the offset front of event U1. These two segments are not identical but have a significant overlap. Figure 4.3(k) shows the obtained segment after merging the segment in Figure 4.3(i) and Figure 4.3(j). It matches the ideal segment of U1 shown in Figure 4.3(c). In addition, the segment corresponding to the onset front of target event U2 has a small overlap with the segment corresponding to interfering event V3. This overlapping region is assigned to event V3 since V3 starts later than U2. Figure 4.3(l) illustrates the bounding contours of all the obtained segments for the mixture of source U and source V. These segments correspond to event U1, U2, V1, and V3. An interfering event V2 is buried in the background since its onset is masked by target event U1 and its offset is masked by target event U2 and therefore there is no onset and offset information that can be used to segment this event.

In this study, we are interested in estimating T-F segments of speech. Since temporal envelope variations down to 4 Hz are essential for speech intelligibility (Drullman et al., 1994a; Drullman et al., 1994b), our system starts segmentation at the time scale s_t = 1/4. In addition, our system starts at the frequency scale $s_c = 6$. We have also considered starting at $s_c = 8$ and $s_c = 4$. In both situations, the system performs slightly worse. In the results reported here, our system forms segments with four scales from coarse to fine: (s_c, s_t) = (6, 1/4), (6, 1/8), (6, 1/14), and (2, 1/14). At each scale, the threshold of onset and offset detection is determined by the smoothed intensity. Specifically, we use $\lambda\sigma$ as the onset threshold and -$\lambda\sigma$ as the offset threshold, where λ is a parameter and σ is the standard deviation of the derivative of $v(c, t, s_c, s_t)$, the smoothed intensity at the current scale. We set $\lambda = 1$ for the stringent thresholds and $\lambda = 0$ for the loose thresholds. The threshold θ_V is 0.999, 0.999, 0.999, and 0.99, respectively; a larger θ_V is used in the first three scales because smoothing over frequency increases the similarity of temporal envelopes in adjacent channels. We set θ_V values close to 1 so that the connected onsets and offsets candidates in adjacent channels are likely from the same event. We have also considered segmentation using more scales and with different types and parameters for the lowpass filter, but did not obtain better results.

As mentioned earlier, our multiscale integration starts at a coarse scale and move to finer scales. One could also start at a fine scale and then move to coarser scales. However, in this case, the chances of over-segmenting an input mixture are higher, which is less desirable than under-segmentation since in subsequent grouping larger segments are preferred (see Section 4.5).

Figure 4.4 shows the bounding contours of segments at different scales for the mixture M1 shown in Figure 1.2(d). Figure 4.4(a) shows the segments formed with obtained onset and offset fronts at the beginning scale (6, 1/4), and Figures 4.4(b), 4.4(c), and 4.4(d) those from the multiscale integration of 2, 3, and 4 scales, respectively. The background is represented by gray. Compared with the ideal segments in Figure 4.1, our system already captures a majority of speech events at the largest scale, but misses some small segments. As the system integrates analysis at smaller scales, more speech segments are formed; at the same time, more segments from interference also appear. Note that the system does not specify the sound source for each segment, which is the task of sequential grouping and will be addressed in Chapter 6.

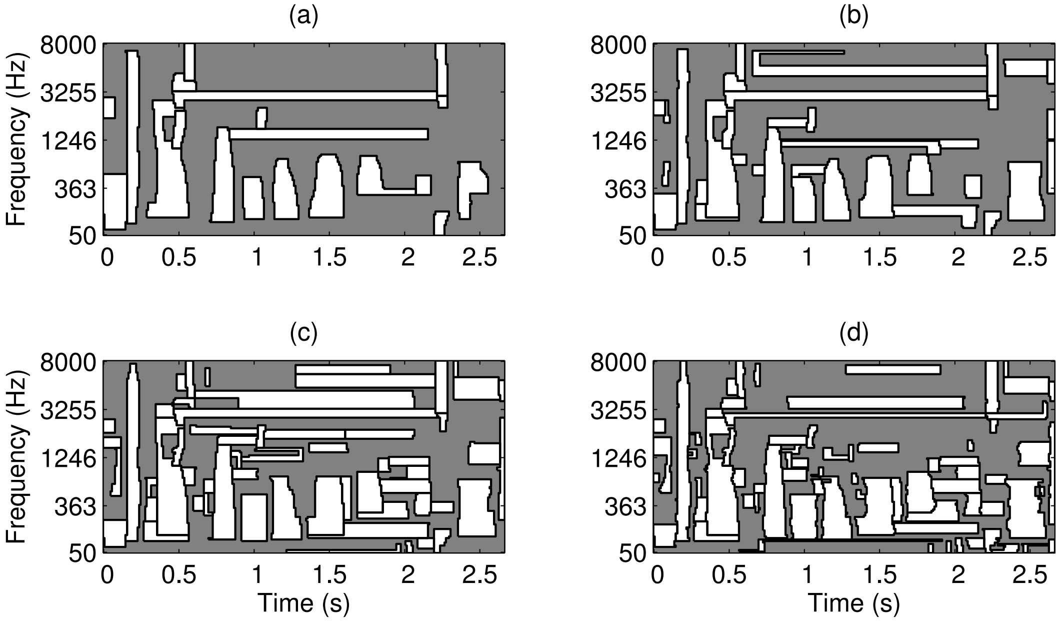

Figure 4.4. Bounding contours of estimated segments from multiscale analysis. (a) One scale analysis at the scale of (6, 1/4). (b) Two-scale analysis at the scales of (6, 1/4) and (6, 1/8). (c) Three-scale analysis at the scales of (6, 1/4), (6, 1/8), and (6, 1/14). (d) Four-scale analysis at the scales of (6, 1/4), (6, 1/8), (6, 1/14), and (2, 1/14). The input is the mixture M1 shown in Figure 1.2(d). The background is represented by gray.

4.5 Evaluation metrics

We compare estimated segments with ideal segments to evaluate the performance of our segmentation algorithm. Note that ultimately the performance of segmentation shall be measured according to its contribution to overall segregation performance. However, it is beneficial to evaluate the performance of segmentation separately.

Only a few previous models have explicitly addressed the problem of auditory segmentation (Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004a) but none have separately evaluated the segmentation performance. How to quantitatively evaluate segmentation results is a complex issue, since one has to consider various types of mismatch between a collection of ideal segments and that of estimated segments. On the other hand, similar issues occur also in image segmentation, which has been extensively studied in computer vision and image analysis. So in this evaluation we adapt region-based metrics by Hoover et al. (1996), which have been widely used for evaluating image segmentation systems.

Our region-based evaluation compares estimated segments with ideal segments of a target source since in many situations one is interested in only target extraction. In other words, how a system segments interference will not be considered in evaluation. Hence, we treat all the T-F regions dominated by interference as the ideal background. Note that this evaluation can be extended to situations where one is interested in evaluating segmentation of multiple sources, say, when interference is a competing talker, by evaluating how a system segments each source separately.

The general idea is to examine the overlap between ideal segments and estimated segments. Based on the degree of overlapping, we label a T-F region as correct, under-segmented, over-segmented, missing, or mismatch. Figure 4.5 illustrates these cases, where ovals represent ideal target segments (numbered with Arabic numerals) and rectangles estimated segments (numbered with Roman numerals). As shown in Figure 4.5, estimated segment I well covers ideal segment 1, and we label the overlapping region

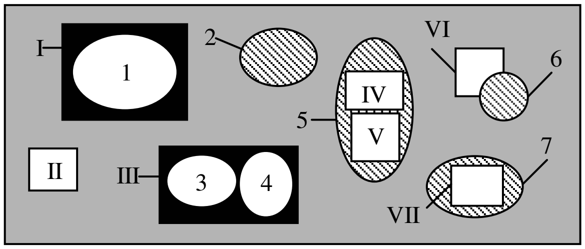

Figure 4.5. Illustration of different matching situations between ideal and estimated segments: Correct segmentation (white regions within segments 1 and 7), under-segmentation (white regions within segments 3 and 4), over-segmentation (white regions within segment 5), missing (regions marked by diagonal lines), and mismatch (black regions). Here an oval indicates an ideal segment and a rectangle an estimated one. The background is represented by gray.

as correct. So is the overlap between segment 7 and VII. Segment III well covers two ideal segments, 3 and 4, and the overlapping regions are labeled as under-segmented. Segment IV and V are both well covered by an ideal segment 5, and the overlapping regions are labeled as over-segmented. All the remaining regions from ideal segments — segments 2 and 6 and the parts of segments 5 and 7 marked by diagonal lines — are labeled as missing. The black region in segment I belongs to the ideal background, but since it is merged with ideal segment 1 into an estimated segment we label this black region as mismatch, as well as the black region in segment III. Note the major differences among under-segmentation, missing, and mismatch. Under-segmentation denotes the error of combining multiple T-F regions belonging to different ideal segments of the same source, whereas missing and mismatch denote the error of mixing T-F regions from different sources. Therefore, if an estimated segment combines T-F regions belonging to different speakers, it is not under-segmentation, but missing or mismatch depending on the degree of overlapping. Segment II is well covered by the ideal background, which is not considered in the evaluation. Much of segment VI is covered by the ideal background and therefore we treat the white region of the segment the same as segment II (Note the difference between segment I and VI). This evaluation provides a comprehensive comparison between the ideal segments and estimated segments, though further research is needed to tell whether or not it has a good correlation with human perception.

Quantitatively, let $\{r_I[k]\}$, $k=0,1,\ldots,K$, be the set of ideal segments, where $r_I[0]$ indicates the ideal background and others the ideal segments of target. Let $\{r_S[l]\}$, $l=0,1,\ldots,L$, be the estimated segments produced by our system, where $r_S[l]$, $l>0$, corresponds to an estimated segment and $r_S[0]$ the estimated background. Let $r[k, l]$ be the overlapping region between $r_I[k]$ and $r_S[l]$. Furthermore, let $E[k, l]$, $E_I[k]$, and $E_S[l]$ denote the corresponding energy in these regions. Given a threshold, we define that an ideal segment $r_I[k]$ is well-covered by an estimated segment $r_S[l]$ if $r[k, l]$ includes most of the energy of $r_I[k]$. That is,

$$E[k,l] > \theta_E \cdot E_I[k] \tag{4.2}$$

Similarly, $r_S[l]$ is well-covered by $r_I[k]$ if

$$E[k,l] > \theta_E \cdot E_S[l] \tag{4.3}$$

For any $\theta_E \in [0.5, 1)$, the above definition of well-coveredness ensures that an ideal segment is well covered by at most one estimated segment, and vice versa.

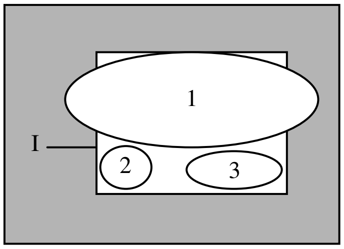

Figure 4.6. Illustration of multiple labels for one overlapping region. Here an oval indicates an ideal segment and a rectangle an estimated one. The background is represented by gray.

Then we label a non-empty overlapping region as follows:

- A region $r[k, l]$, $k>0$ and $l>0$, is labeled as correct if $r_I[k]$ and $r_S[l]$ are mutually well-covered.
- Let $\{r_I[k']\}$, $k'=k_1, k_2, \ldots, k_{K'}$, and $K'>1$, be all the ideal target segments that are well-covered by one estimated segment, $r_S[l]$, $l>0$. The corresponding overlapping regions, $\{r[k', l]\}$, $k'=k_1, k_2, \ldots, k_{K'}$, are labeled as under-segmented if these regions combined include most of the energy of $r_S[l]$, that is:

$$\sum_{k'} E[k',l] > \theta_E \cdot E_S[l], \quad k' = k_1, k_2, \ldots, k_{K'} \tag{4.4}$$

- Let $\{r_S[l']\}$, $l'=l_1, l_2, \ldots, l_{L'}$, and $L'>1$ be all the estimated segments that are well-covered by one ideal segment, $r_I[k]$, $k>0$. The corresponding overlapping regions, $\{r[k, l']\}$, $l'=l_1, l_2, \ldots, l_{L'}$, are labeled as over-segmented if these regions combined include most of the energy of $r_I[k]$, that is:

$$\sum_{l'} E[k,l'] > \theta_E \cdot E_I[k], \quad l' = l_1, l_2, \cdots, l_{L'} \tag{4.5}$$

- If a region $r[k, l]$ is part of an ideal segment of target speech, i.e., $k>0$, but cannot be labeled as correct, under-segmented, or over-segmented, it is labeled as missing.
- Region $r[0, l]$, the overlap between the ideal background $r_I[0]$ and an estimated segment $r_S[l]$, is labeled as mismatch if $r_S[l]$ is not well-covered by the ideal background.

According to the above definitions, some regions may be labeled as either correct or under-segmented. Figure 4.6 illustrates this situation, where estimated segment I and ideal segment 1 are mutually well-covered. Hence, r[1, I] is labeled as correct. On the other hand, segment I also well covers ideal segments 2 and 3, and obviously ideal segments 1-3 together well cover segment I. According to the definition of under-segmentation, r[1, I], r[2, I], and r[3, I] should all be labeled as under-segmented. Therefore, r[1, I] can be labeled as either correct or under-segmented. Similarly, some regions may be labeled as either correct or over-segmented. To avoid labeling a region more than once, we consider a region to be correctly labeled as long as it satisfies the definition of correctness.

Let E_C, E_U, E_O, E_M, and E_N be the summated energy in all the regions labeled as correct, under-segmented, over-segmented, missing, and mismatch, respectively. Further let E_I be the total energy of all ideal segments of target, and E_S that of all estimated segments, except for the estimated background. We use the following metrics for evaluation:

- The correct percentage: $P_C = E_C/E_I \times 100\%$.
- The percentage of under-segmentation: $P_U = E_U/E_I \times 100\%$.
- The percentage of over-segmentation: $P_O = E_O/E_I \times 100\%$.
- The percentage of missing: $P_M = E_M/E_I \times 100\%$.
- The percentage of mismatch: $P_N = E_N/E_S \times 100\%$.

Since $E_C + E_U + E_O + E_M = E_I$, or $P_C + P_U + P_O + P_M = 100\%$, only three out of these four percentages need to be measured.

The advantage of evaluation according to each category is that it clearly shows different types of error. In the context of speech segregation, under-segmentation is not really an error since it basically produces larger segments for target speech, which is good for subsequent grouping. In image segmentation, the region size corresponding to each segment is used for evaluation literally. Here, we use the energy of each segment because for acoustic signals, T-F regions with strong energy are much more important to segment than those with weak energy.

4.6 Evaluation results

Our proposed segmentation process was evaluated with a test corpus containing 20 target utterances from the test part of the TIMIT database (Garofolo et al., 1993) mixed with 20 intrusions. Target utterances and intrusions are listed in Tables 4.1 and 4.2 accordingly. This set of intrusions represents a broad range of real sounds encountered in typical acoustic environments. Each target utterance is mixed with individual intrusions

Target	Content
S1	Put the butcher block table in the garage
S2	Alice's ability to work without supervision is noteworthy
S3	Barb burned paper and leaves in a big bonfire
S4	Swing your arm as high as you can
S5	Shaving cream is a popular item on Halloween
S6	He then offered his own estimate of the weather, which was unenthusiastic
S7	The morning dew on the spider web glistened in the sun
S8	Her right hand aches whenever the barometric pressure changes
S9	Why yell or worry over silly items
S10	Aluminum silverware can often be flimsy
S11	Guess the question from the answer
S12	Medieval society was based on hierarchies
S13	That noise problem grows more annoying each day
S14	Don't ask me to carry an oily rag like that
S15	Each untimely income loss coincided with the breakdown of a heating system part
S16	Combine all the ingredients in a large bowl
S17	Fuss, fuss, old man
S18	Don't ask me to carry an oily rag like that
S19	The fish began to leap frantically on the surface of the small lake
S20	The redcoats ran like rabbits

Table 4.1. Target utterances in the test corpus

Target	Content
N1	White noise
N2	Rock Music
N3	Siren
N4	Telephone
N5	Electric fan
N6	Clock alarm
N7	Traffic noise
N8	Bird chirp with water flowing
N9	Wind
N10	Rain
N11	Cocktail party noise
N12	Crowd noise at a playground
N13	Crowd noise with music
N14	Crowd noise with clap
N15	Babble noise (16 speakers)
N16	Don't ask me to carry an oily rag like that
N17	She had your dark suit in greasy wash water all year
N18	Why were keen to use human w...
N19	The local drugstore was charged with illegally dispensing tranquilizers
N20	There are many such competently anonymous performances among the earlier poems

Table 4.2. Intrusions in the test corpus

at -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB SNR. This test corpus has 400 mixtures at each SNR level and 2000 mixtures altogether.

To determine the ideal segments of a target utterance, we need to decide what constitutes the acoustic events of a speech utterance. Here we treat a phoneme as an acoustic event. As we discussed in Section 4.1, coarticulation between neighboring phonemes may create unnatural boundaries in ideal segments, a case of under-segmentation. This problem is partly taken care of in our evaluation which does not consider under-segmentation as an error. To avoid the problem of coarticulation, one could define a larger unit (e.g. a syllable or a word) as an acoustic event. As discussed earlier, over-segmentation becomes an issue in such a definition. Because it is not clear whether an instance of over-segmentation is caused by a true boundary between two phonemes or a genuine error, over-segmentation is a more thorny issue. This consideration has led us to choose phonemes as event units.

A. Overall performance

Figure 4.7 shows the average P_C, P_U, P_O, and P_N for different θ_E values at different SNR levels. The evaluation is more stringent for higher θ_E. Note that we limit θ_E to be no smaller than 0.5 so that an ideal segment is well covered by at most one estimated segment, and vice versa (see Section 4.5). As shown in the figure, our system performs better as SNR increases. When θ_E is 0.5, the correct percentage is 42.7% at -5 dB SNR and it increases to 67.0% as SNR increases to 15 dB. On the other hand, as θ_E increases, the correct percentage gradually decreases to 0. A significant amount of speech is

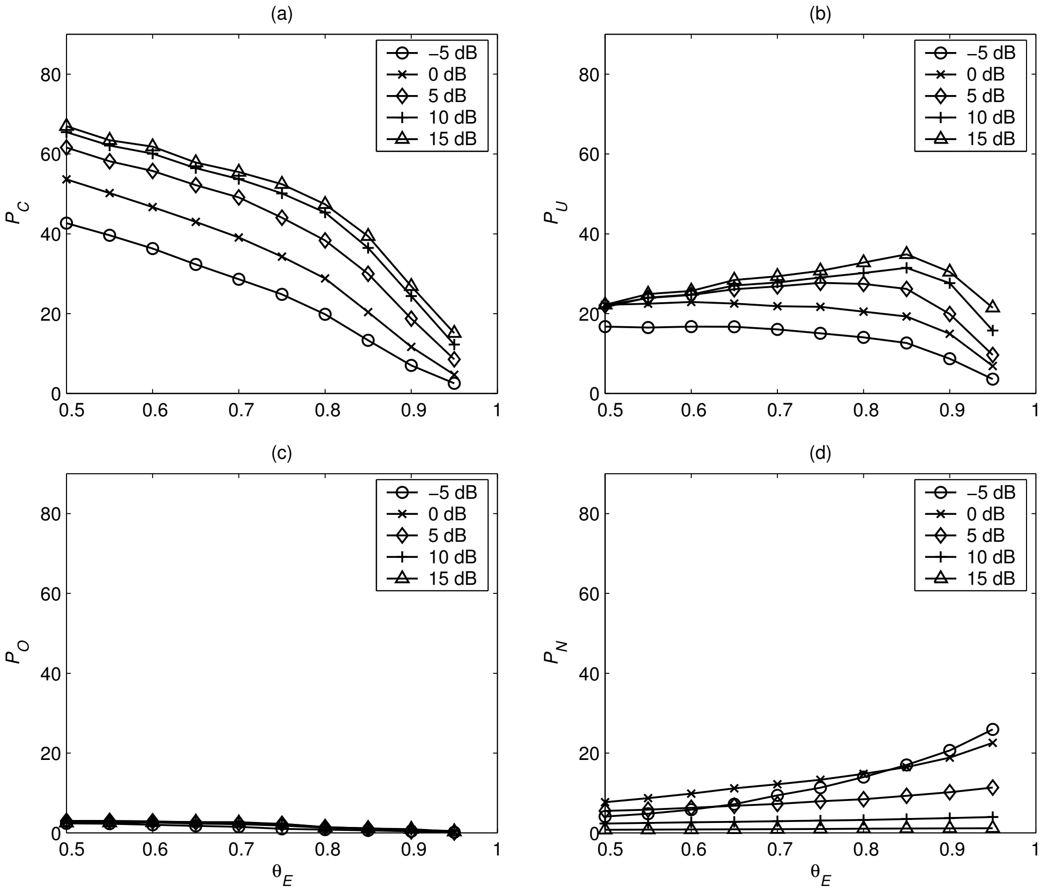

Figure 4.7. Results of auditory segmentation at different SNR levels. (a) Average correct percentage. (b) Average percentage of under-segmentation. (c) Average percentage of over-segmentation. (d) Average percentage of mismatch.

under-segmented, which is due mainly to coarticulation of phonemes. As discussed in Section 4.5, under-segmentation is not really an error. By combining P_C and P_U together, when θ_E is 0.5, our system correctly segments 59.4% of target speech at -5 dB SNR. This number increases to 89.2% at 15 dB SNR. In addition, we can see from the figure that over-segmentation is negligible. The main error comes from missing, which indicates that portions of target speech are buried in the background. The percentage of mismatch averaged over different θ_E values shown in the figure is 12.0% at -5 dB SNR, and it drops to 0.96% when the SNR increases to 15 dB. Compared with the SNRs of mixtures, the percentage of mismatch is not high. This shows that most target and interference are well separated in the estimated segments.

Since voiced speech is generally much stronger than unvoiced speech, the above result mainly reflects the performance of our system on voiced speech. To see how it performs on unvoiced speech, Figure 4.8 shows the average P_C, P_U, and P_O for expanded obstruents, which include the majority of phonemes that contain unvoiced speech energy (see Section 2.2). As shown in the figure, much energy of these phonemes is under-segmented. As expected, the overall performance on these phonemes is not as good as that for other phonemes since unvoiced speech is generally softer and more prone to interference. The average P_C+P_U is 54.1% at -5 dB SNR when θ_E is 0.5, and it increases to 77.0% when SNR increases to 15 dB.

The proposed system is similar to our segmentation system described in Hu and Wang (2004b) and Hu and Wang (2006b), except for one major difference. In this proposed system we allow multiple onset and offset fronts to form one segment, whereas in the

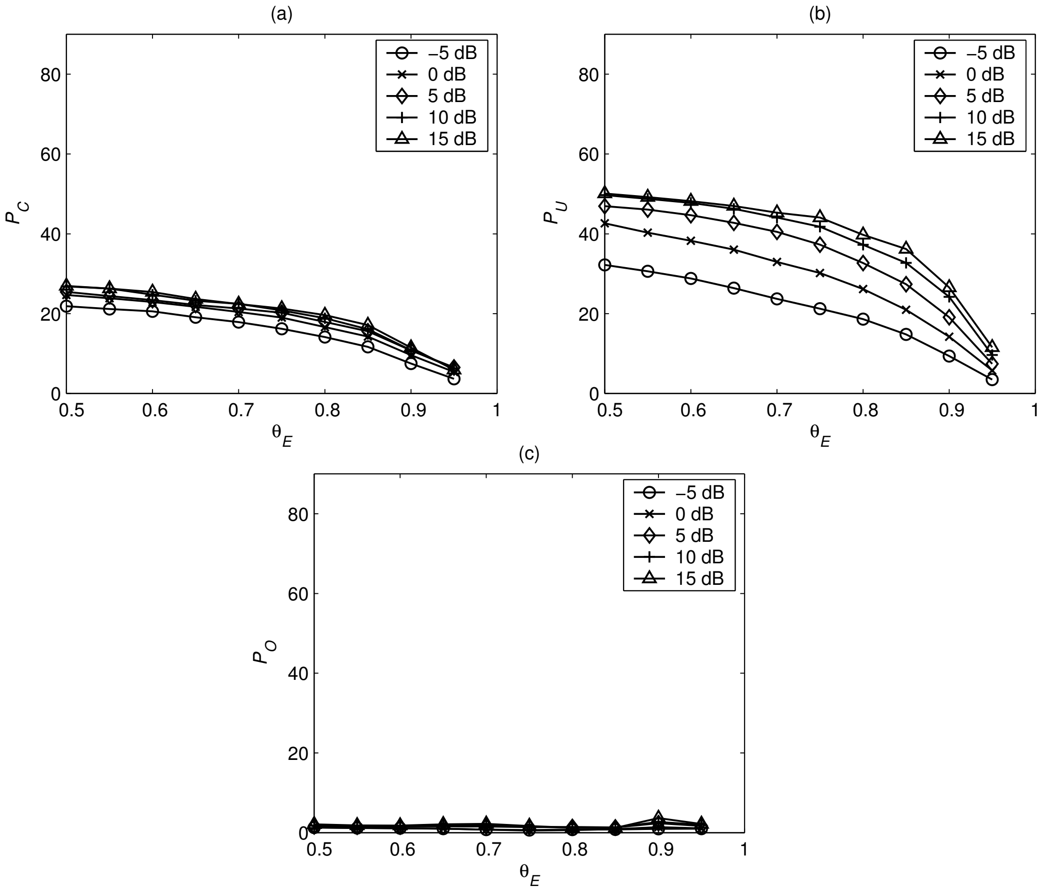

Figure 4.8. Results of auditory segmentation for expanded obstruents at different SNR levels. (a) Average correct percentage. (b) Average percentage of under-segmentation. (c) Average percentage of over-segmentation.

previous system we allowed a segment to have only one onset front. Compared with the previous system, the proposed system performs better on the test corpus. In particular, it yields a correct percentage similar to that of the previous system, but achieves an approximate 10% relative reduction in the percentage of mismatch.

To put the performance of the proposed segmentation process in perspective, we now compare it with the segmentation algorithm described by Brown and Cooke (1994). Their algorithm first produces spectral peak tracks on a frequency transition map and then extends each track in frequency by clustering cross-channel correlation values. Figure 4.9 shows the comparative results for mixtures at 0 dB SNR. Figure 4.9(a) shows the average P_C+P_U scores for all the phonemes. The Brown and Cooke algorithm yields much lower P_C+P_U scores compared with the proposed system. The primary reason is that their algorithm is based on cross-channel correlation of filter responses, which often fails to merge target speech across frequency because target speech may yield different responses in neighboring filter channels. Since their algorithm was mainly intended for segmenting voiced sound, a further comparison for only voiced speech in terms of P_C+P_U is given in Figure 4.9(b). In this case, the voiced portions of each utterance are determined using *Praat*, which has a standard pitch determination algorithm for clean speech (Boersma and Weenink, 2004). The performance gap in Figure 4.9(b) is not much different from that in Figure 4.9(a). Figure 4.9(c) shows the average P_N. Their algorithm produces lower P_N errors compared with the proposed process, because segmentation exploits harmonic structure and most intrusions in the evaluation corpus are noise-like. Taken together, the proposed algorithm performs much better than their algorithm for auditory segmentation.

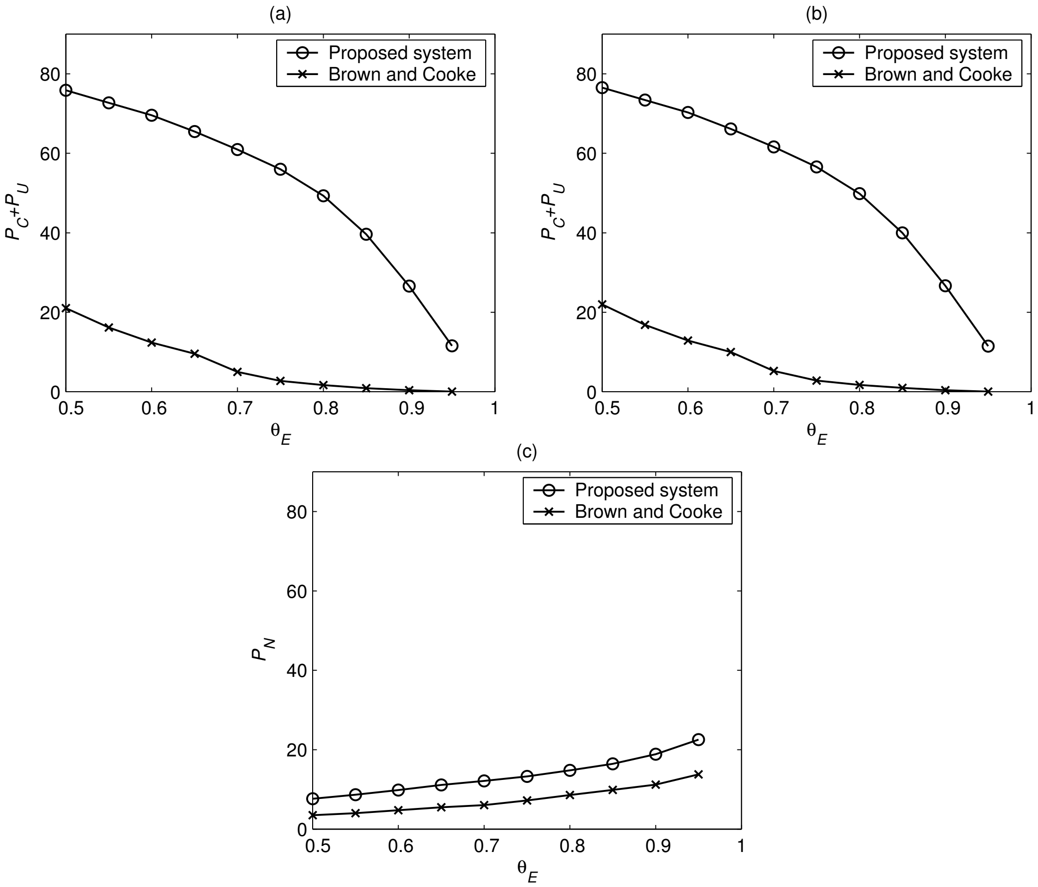

Figure 4.9. Results of auditory segmentation for the proposed system and those from the Brown and Cooke algorithm. Target and interference are mixed at 0 dB SNR. (a) Average correct percentage plus that of under-segmentation for all the phonemes. (b) Average correct percentage plus that of under-segmentation for the voiced portions of utterance. (c) Average percentage of mismatch.

B. Pre-emphasis

Because the low-frequency portion of speech is usually more intense than the high-frequency portion, the above energy-based evaluation may be dominated by the speech energy in the low-frequency range. To present a more balanced picture, we apply a first-order highpass filter with the coefficient 0.95 to the input mixture to pre-emphasize its high-frequency portion, which approximately equalizes the average energy of speech in each filter channel. The energy of each segment after pre-emphasis is used for evaluation.

Figures 4.10 and 4.11 present the results with pre-emphasis for all the phonemes and for expanded obstruents, respectively. As shown in the figures, the P_C scores for all the phonemes with pre-emphasis are about 5% higher than those without pre-emphasis, whereas the P_U scores are about 10% lower. This suggests that more voiced speech is under-segmented in the low-frequency range. The P_C scores for expanded obstruents with pre-emphasis are much higher than those without pre-emphasis, whereas the P_U scores are much lower. The P_C+P_U scores together with pre-emphasis are about 8% higher than those without pre-emphasis. This suggests that our system under-segments most expanded obstruents in the low-frequency range, which is mainly voiced. On the other hand, it correctly separates most expanded obstruents in the high-frequency range, where the energy of unvoiced speech is more distributed, from neighboring phonemes as well as from interference. The average P_N with pre-emphasis is relatively more constant across different SNR levels than that without pre-emphasis. This is because more interference energy is distributed in the high-frequency range and pre-emphasis reduces the SNR variation.

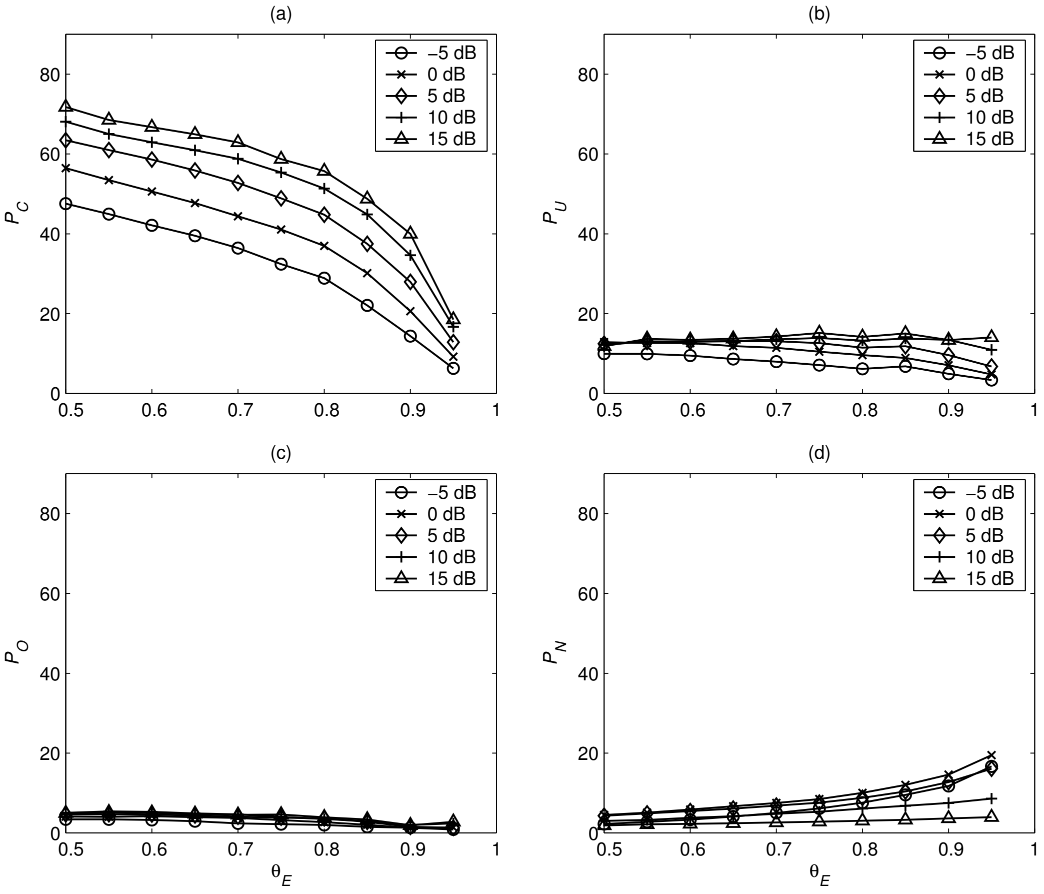

Figure 4.10. Results of auditory segmentation with pre-emphasis at different SNR levels for all the phonemes. (a) Average correct percentage. (b) Average percentage of under-segmentation. (c) Average percentage of over-segmentation. (d) Average percentage of mismatch.

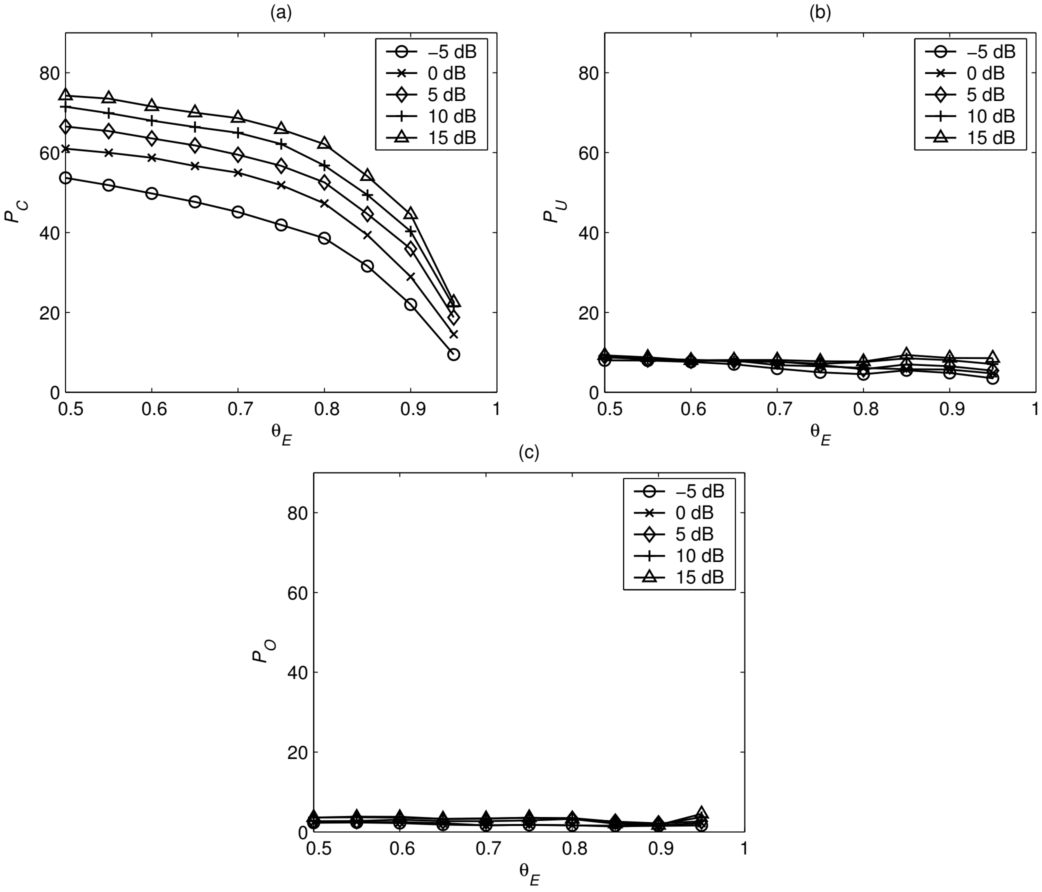

Figure 4.11. Results of auditory segmentation with pre-emphasis for expanded obstruents at different SNR levels. (a) Average correct percentage. (b) Average percentage of under-segmentation. (c) Average percentage of over-segmentation.

C. Advantages of multiscale analysis

To gauge the advantages of the multiscale analysis over single-scale analysis, we segment with several single-scale analyses and compare with the proposed multiscale analysis. Figure 4.12 shows the average P_C+P_U, P_O, and P_N scores obtained with single-scale analyses by varying the parameter λ for the stringent threshold (see Section 4.4). These single scales are (6, 1/4), (6, 1/8), (6, 1/14), and (2, 1/14), corresponding to scales 1, 2, 3, and 4 shown in the figure. Each value in the figure is the average score over 10 different θ_E values ranging from 0.5 to 0.95 with increment 0.05. Figure 4.12 also shows the P_C+P_U, P_O, and P_N scores for the multiscale analysis of these 4 scales. As shown in the figure, the outcome of a single-scale analysis is strongly dependent on the scale value. When a larger scale is used, more target is correctly segmented and less target is over-segmented. However, more target and interference are merged into the same segments and the percentage of mismatch is higher. On the other hand, when a smaller scale is used, target and interference are better separated into the estimated segments, but less target is correctly segmented and more target is over-segmented. The multiscale analysis correctly segments more target than any single scale analysis for λ values between [0.75, 1] and yields a moderate percentage of mismatch. In fact, the percentage of mismatch for the multiscale analysis is better than those of single scale analyses of scales 1 and 2. Furthermore, a single-scale analysis is more sensitive to the threshold of onset and offset detection than the multiscale analysis. As shown in the figure, when the onset and offset thresholds deviate from the best values, the performance of a single scale analysis, especially for small scales, drops more significantly than that of the multiscale

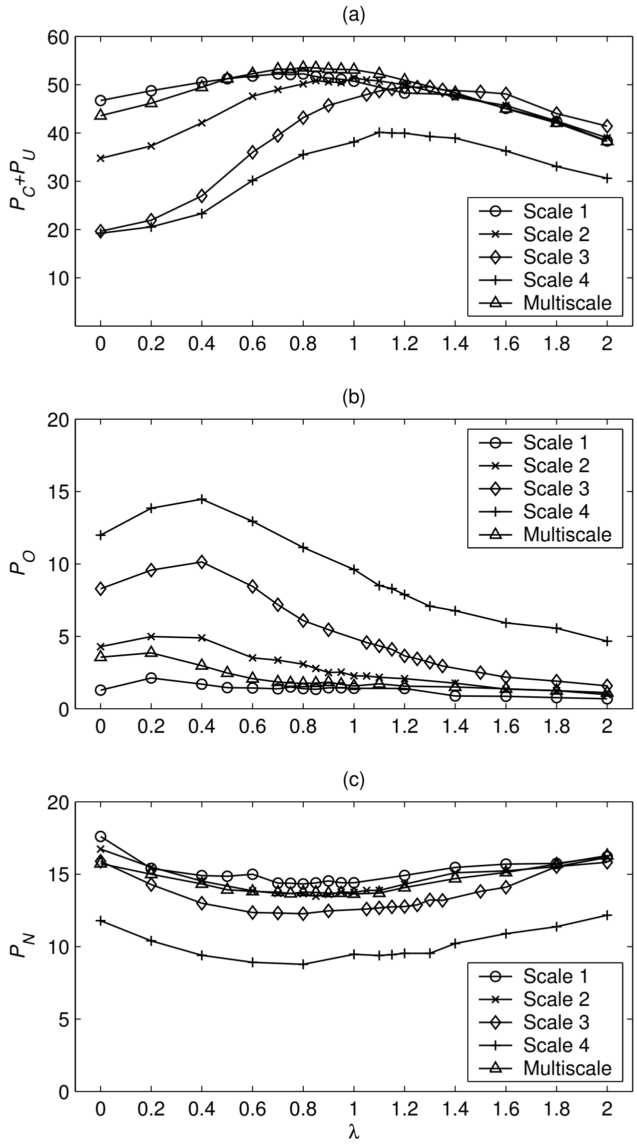

Figure 4.12. A comparison between multiscale analysis and single scale analysis with different values of parameter λ. Scale 1, 2, 3, and 4 are (6, 1/4), (6, 1/8), (6, 1/14), and (2, 1/14), respectively. (a) Average percentage of correct plus that of under-segmentation. (b) Average percentage of over-segmentation. (c) Average percentage of mismatch.

analysis. Therefore, using a multiscale analysis provides an additional advantage of simplifying parameter tuning.

D. Error analysis

There are two types of error in segmentation. The first comes from segmentation within individual channels, i.e., missing true onsets and offsets, falsely detected onsets and offsets, or inaccurate estimation of onset and offset position. The second comes from segmentation across frequency channels when we merge adjacent channels that are dominated by different sources, or when we divide adjacent channels that are dominated by the same source.

To gain further insight into the performance of segmentation, we separate these two types of error by comparing estimated segments and ideal segments channel by channel, which only documents the first type of error. For convenience, we refer to a contiguous T-F region of a segment in a channel as a time-segment (T-segment). Figure 4.13 shows the P_C, P_U, P_O, and P_N scores for estimated T-segments compared with ideal T-segments. Each value is an average across all the channels weighted by the total energy of all ideal segments of target after pre-emphasis. Comparing this figure with Figure 4.10, we can see that the correct percentage of T-segments is much higher than that of entire estimated segments at low SNR levels and the percentage of mismatch for T-segments is much lower. This shows that a significant amount of the mismatch error stems from merging channels across frequency, i.e., from connecting onset and offset candidates into onset and offset fronts (see Section 4.3).

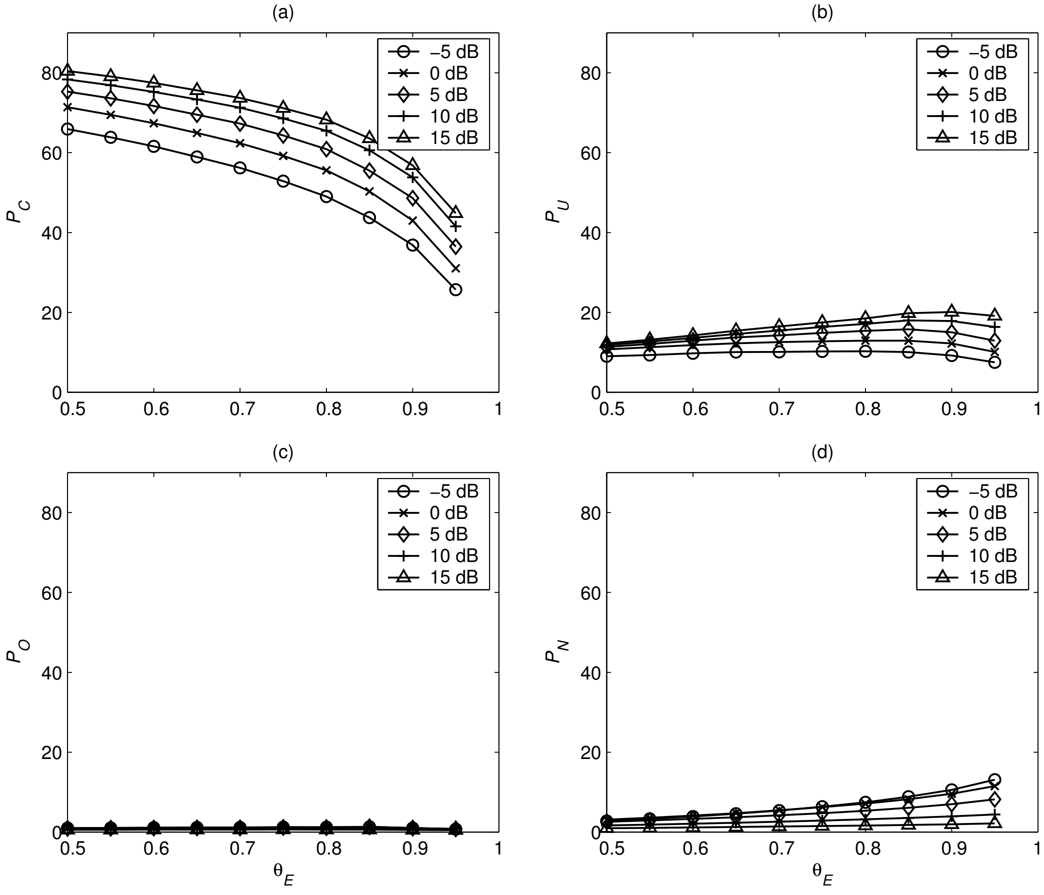

Figure 4.13. Results for estimated T-segments in individual channels at different SNR levels. (a) Average correct percentage. (b) Average percentage of under-segmentation. (c) Average percentage of over-segmentation. (d) Average percentage of mismatch.

CHAPTER 5

ITERATIVE PITCH DETERMINATION AND VOICED SPEECH SEGREGATION

Our computational goal of speech segregation is to obtain the ideal binary mask of a target utterance (see Section 1.3). In such a mask, T-F units dominated by the target are labeled 1 and others are labeled 0. In this chapter, we describe the process that estimates the ideal binary mask for the voiced portions of a target utterance.

As discussed in Section 2.4A, accurate pitch estimation of a target utterance is crucial for segregating the voiced target. However, robust pitch estimation is a very challenging task because interference corrupts target pitch information. To deal with this problem, we propose an algorithm that estimates target pitch and segregates voiced target in an iterative manner. In particular, this algorithm first roughly estimates the pitch of a target utterance in an acoustic mixture and uses this estimate to segregate the voiced target, i.e., to estimate the ideal binary mask of the voiced target. From the estimated mask, our algorithm generates a better estimate of the target pitch and then a better estimate of the mask using the newly estimated pitch, and so on. Our algorithm stops when the iterative process converges. The output of the algorithm is several estimated pitch contours and a binary mask associated with each pitch contour. Note that the algorithm does not specify whether an estimated pitch contour is from a target utterance or not. Determining which pitch contours are from a target utterance is a task of sequential grouping, which will be addressed in the next chapter.

Our algorithm has two key steps: Estimating the ideal binary mask of a target utterance given an estimate of target pitch and estimating the target pitch given an estimate of the ideal binary mask. In the first two sections of this chapter, we first explore the solution to each of the two steps. The iterative algorithm is then presented in detail in Section 5.3. Evaluation results of this algorithm on pitch estimation and voiced speech segregation are given in Section 5.4.

5.1 Ideal binary mask estimation given estimated target pitch

In this section, we discuss the problem of estimating the ideal binary mask of a target utterance given an estimate of the target pitch.

A. Labeling T-F units with information within individual T-F units

We first consider a simple approach: a T-F unit is labeled 1 if and only if the corresponding response or response envelope has a periodicity similar to that of the target.

As discussed in Section 2.3, in the low-frequency range, harmonics are resolved. A T-F unit corresponding to a resolved harmonic can be labeled by comparing target pitch period with the period of the filter response within this unit. In the high-frequency range, harmonics are generally unresolved. A T-F unit corresponding to several unresolved harmonics can be labeled by comparing the target pitch with the AM rate of the filter response within this unit.

We applied the above method in our previous study on voiced speech segregation (Hu and Wang, 2004a). There, we first distinguished T-F units responding to resolved harmonics from other units based on the cross-channel correlation of filter responses. As pointed out in this study, a resolved harmonic usually activates several adjacent channels and the corresponding responses are highly correlated with each other (see Figure 3.2 for an illustration), whereas channels responding to multiple harmonics are usually not as highly correlated. Therefore, we considered T-F units with sufficiently high cross-channel correlations, i.e., $C(c, m) > 0.985$, as responding to individual resolved harmonics and others to multiple harmonics.

We labeled a T-F unit corresponding to a resolved harmonic as follows. Since the periodicity of the filter response is indicated by the peaks in the corresponding ACF, a unit u_{cm} is labeled as target if the corresponding ACF at the estimated pitch period is close to the maximum of the ACF within the plausible pitch range, i.e.:

$$\frac{A(c,m,\tau_S(m))}{\max_{\tau\in\Gamma} A(c,m,\tau)} > \theta_T \tag{5.1}$$

where $\tau_S(m)$ is the estimated pitch period at frame m, Γ is [2.5 ms, 14.29 ms], the plausible pitch range (see Section 3.1), and θ_T is a threshold. In addition, we labeled a T-F unit corresponding to multiple harmonics by comparing the AM fluctuation of the response within this unit with a sinusoidal function that has the period of target pitch. If the sinusoidal function fits the AM fluctuation well, this T-F unit is labeled as target.

In a later study (Hu and Wang, 2006a), we used a method similar to Equation (5.1) to label a T-F unit responding to multiple harmonics since the AM rate of the response is indicated by the peaks in the ACF of the response envelope. In particular, a T-F unit is labeled as target if:

$$\frac{A_E(c,m,\tau_S(m))}{\max_{\tau \in \Gamma} A_E(c,m,\tau)} > \theta_A \tag{5.2}$$

where θ_A is a threshold. This method is much simpler than that used in Hu and Wang (2004a), but yields a comparable result.

We test the above labeling scheme, i.e., Equations (5.1) and (5.2), with the test corpus described in Section 4.6. This corpus contains 20 target utterances from the TIMIT database (Garofolo et al., 1993) and 20 intrusions (see Tables 4.1 and 4.2). The target pitch is obtained by applying *Praat* (Boersma and Weenink, 2004) to the clean target utterance. We use two error measures to evaluate the performance of labeling individual T-F units, the percentage of false acceptance, i.e., labeling an interference-dominant T-F unit as target, and the percentage of false rejection, i.e., labeling a target-dominant T-F unit as interference. Figure 5.1(a) shows the average percentage of false rejection versus the average percentage of false acceptance with different θ_T values for T-F units corresponding to resolved harmonics, and Figure 5.1(b) that with different θ_A values for other units. Each value is the average of 2000 mixtures in the test corpus (see Section 4.6). As shown in the figure, T-F units with high cross-channel correlations are labeled more accurately than other units. In particular, in Figure 5.1(a) an equal error rate of 19.3% for both measures is obtained when $\theta_T = 0.55$, while in Figure 5.1(b) an equal error rate of 28.1% is obtained when $\theta_A = 0.13$.

Equation (5.1) is not a direct comparison of the period of a filter response and the estimated pitch period, and therefore may be misleading sometime. For example, it is possible that $A(c, m, \tau)$ is quite flat within the plausible pitch range. In such a situation, although $A(c, m, \tau)$ does not have a period close to $\tau_S(m)$, $A(c, m, \tau_S(m)) / \max_{\tau \in \Gamma} A(c, m, \tau)$ may be high, and based on Equation (5.1) we will make a mistake. A more direct comparison is to compare the instantaneous frequency of the filter response with the estimated target pitch, which has been suggested by Cooke (1993). Similarly, we can directly compare the instantaneous frequency of the response envelope with the estimated target pitch instead of using Equation (5.2), which was implemented in our previous studies (Hu and Wang, 2002; Hu and Wang, 2004a). However, in practice, it is extremely difficult to accurately estimate the instantaneous frequency of a signal (Boashash, 1992a; Boashash, 1992b), and in most situations, we can only have a good approximation. We found that labeling T-F units by comparing the estimated instantaneous frequency with the target pitch period does not perform better than using the measures in Equation (5.1) and (5.2). Nevertheless, it is better to combine these two measures with instantaneous

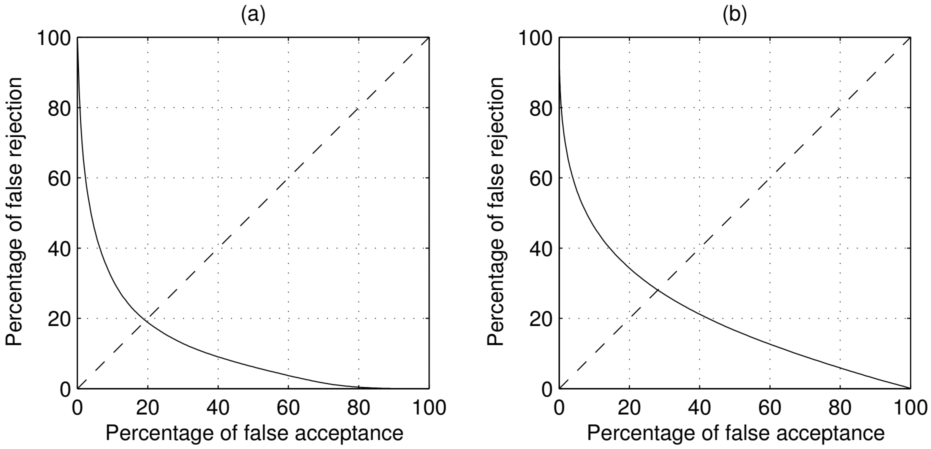

Figure 5.1. Results of labeling T-F units with the method of Hu and Wang (2006a). (a) Percentage of false rejection versus percentage of false acceptance by varying θ_T for T-F units responding to resolved harmonics. (b) Percentage of false rejection versus percentage of false acceptance by varying θ_A for T-F units responding to unresolved harmonics.

frequencies to label T-F units. This observation will be clearer when we present some test results in the later part of this section.

We construct a classifier to label T-F units using the corresponding ACFs at pitch points and instantaneous frequencies as features. Let $\bar{f}(c,m)$ be the estimated average instantaneous frequency of the filter response within a T-F unit u_{cm}. If the filter response has a period close to $\tau_S(m)$, then $\bar{f}(c,m)\cdot\tau_S(m)$ is close to an integer larger than or equal to 1. Similarly, let $\bar{f}_E(c,m)$ be the estimated average instantaneous frequency of the response envelope within u_{cm}. If the response envelope fluctuates at the period of $\tau_S(m)$, then $\bar{f}_E(c,m) \cdot \tau_S(m)$ is close to 1. Let

$$\begin{aligned} r_{cm}(\tau) = (A(c,m,\tau),\ \ \bar{f}(c,m)\tau - \text{int}(\bar{f}(c,m)\tau),\ \ \text{int}(\bar{f}(c,m)\tau), \\ A_E(c,m,\tau),\ \ \bar{f}_E(c,m)\tau - \text{int}(\bar{f}_E(c,m)\tau),\ \ \text{int}(\bar{f}_E(c,m)\tau)) \end{aligned} \quad (5.3)$$

be a set of 6 features, where

$$\text{int}(x) = \begin{cases} \lceil x \rceil & if\ \ (\lceil x \rceil - x) \le 0.5 \\ \lfloor x \rfloor & else \end{cases}$$

Here $\lceil x \rceil$ is the ceiling function that returns the smallest integer greater than or equal to x and $\lfloor x \rfloor$ the floor function. Let H_0 be the hypothesis that a T-F unit is target dominant and H_1 otherwise. u_{cm} is labeled as target if and only if

$$P(H_0 \mid r_{cm}(\tau_S(m))) > P(H_1 \mid r_{cm}(\tau_S(m))) \quad (5.4)$$

Since

$$P(H_0 \mid r_{cm}(\tau_S(m))) = 1 - P(H_1 \mid r_{cm}(\tau_S(m))), \quad (5.5)$$

Equation (5.4) becomes

$$P(H_0 \mid r_{cm}(\tau_S(m))) > 0.5 \quad (5.6)$$

In the feature set we use $A(c, m, \tau_S(m))$ instead of $A(c, m, \tau_S(m))/\max_{\tau\in\Gamma} A(c,m,\tau)$ since we found that these two yield similar performance in labeling T-F units and the first one is easier to compute. The same is for $A_E(c, m, \tau_S(m))$. In addition, we use $\text{int}(\bar{f}(c,m) \cdot \tau_S(m))$ and $\bar{f}(c,m) \cdot \tau_S(m) - \text{int}(\bar{f}(c,m) \cdot \tau_S(m))$ instead of $\bar{f}(c,m) \cdot \tau_S(m)$ since the first two indicate the similarity between the response period and the estimated pitch period more directly, hence simplifying the classifier. The same is for $\bar{f}_E(c,m) \cdot \tau_S(m)$.

In this study, we estimate the instantaneous frequency of the response within a T-F unit simply as half the inverse of the interval between zero-crossings of the response (Boashash, 1992b), assuming that the response is approximately a sinusoidal function. Note that a sinusoidal function crosses zero twice within a period. We have also considered calculating the instantaneous frequency with a more complex method (Kumaresan and Rao, 1999) and found that it yields similar performance in labeling T-F units but entails a much higher cost of computation.

We construct a multilayer perceptron (MLP) (Principe et al., 2000) with one hidden layer to compute $P(H_0|r_{cm}(\tau))$ for each filter channel. The desired output of the MLP is 1 if the corresponding T-F unit is target dominant and 0 otherwise. When there are sufficient training samples, the trained MLP yields a good estimate of $P(H_0|r_{cm}(\tau))$ (Bridle, 1989). In this study, the MLP for each channel is trained with a corpus that includes all the utterances from the training part of the TIMIT database and 100 intrusions. These intrusions include crowd noise and environmental sounds, such as wind, bird chirp, and ambulance alarm. Utterances and intrusions are mixed at 0 dB SNR to generate training samples; target is one utterance and interference is either non-speech intrusion or another utterance. The ideal binary mask of each mixture is obtained from the corresponding premixing target and interference. We use *Praat* to estimate target pitch. The number of units in the hidden layer is determined using cross-validation. Specifically, we divide the training samples equally into two sets, one for training and the

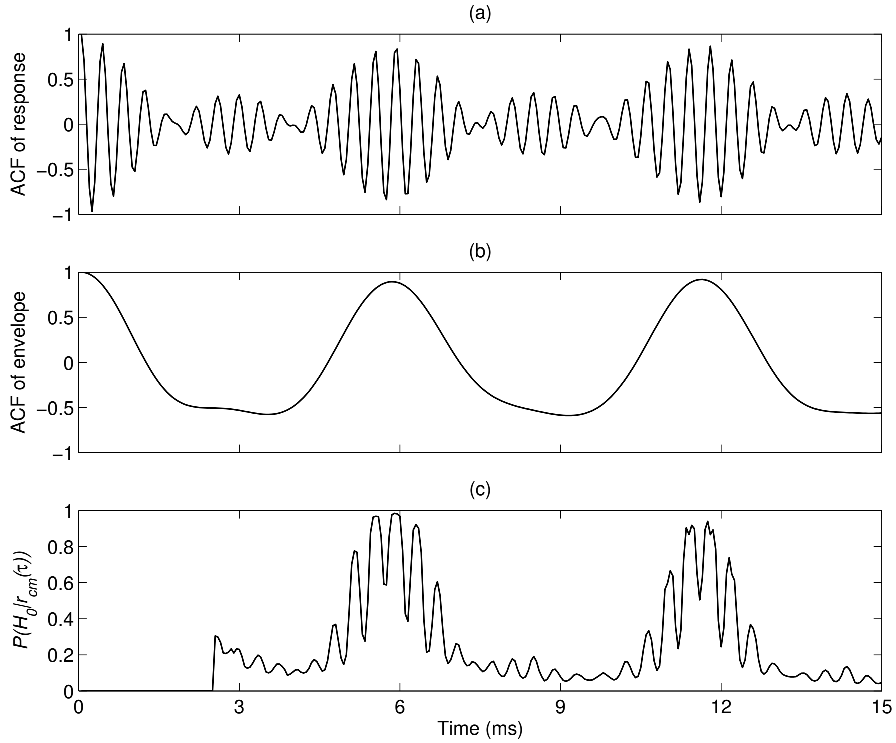

Figure 5.2. (a) ACF of the filter response within a T-F unit in a channel centered at 2.5 kHz. (b) Corresponding ACF of the response envelope. (c) Probability of the unit being target dominant given target pitch period τ. The input is the female utterance shown in Figure 1.2(b).

other for validation. The number of units in the hidden layer is chosen to be the minimum number such that adding more units in the hidden layer will not yield any significant performance improvement on the validation set. Since most obtained MLPs have 5 units in their hidden layers, we let all the MLPs contain 5 units in their hidden layers and train them accordingly.

As an example, Figure 5.2(c) shows the obtained $P(H_0|r_{cm}(\tau))$ for different τ values of a T-F unit in a filter channel centered at 2.5 kHz and at a time frame (from 790 ms to 810 ms). The input is the female utterance shown in Figure 1.2(b). The corresponding ACFs of the filter response and the response envelope are shown in Figures 5.2(a) and 5.2(b), respectively. As shown in the figure, the maximum of $P(H_0|r_{cm}(\tau))$ is located at 5.87 ms, the pitch period of the utterance at this frame.

The obtained MLPs are used to label individual T-F units according to Equation (5.6). Figure 5.3(a) shows the resulting error rate by channel for all the mixtures in the test corpus (see Section 4.6). The error rate is the average of false acceptance and false rejection. As shown in the figure, with features derived from individual T-F units, we could label about $70\% - 90\%$ of the units correctly across the whole frequency range. In general, T-F units in the low-frequency range are labeled more accurately than those in the high-frequency range. Figure 5.3 also shows the error rate by using only subsets of the features from the feature set, $r_{cm}(\tau)$. As shown in Figures 5.3(b) and 5.3(c), the ACF values at the pitch point and instantaneous frequencies provide complementary information. The response envelope is more indicative than the response itself in the high-frequency range. Best results are obtained when all the 6 features are used, and they are much better than the previous method that labels T-F units based on Equations (5.1) and (5.2). In particular, for T-F units corresponding to resolved harmonics, the current method obtains a 9.1% false rejection and an 18.1% false acceptance. The average error rate is 13.6%, which is 6.1% lower than the equal error rate from the previous method (see Figure 5.1). For T-F units corresponding to multiple harmonics, the present method

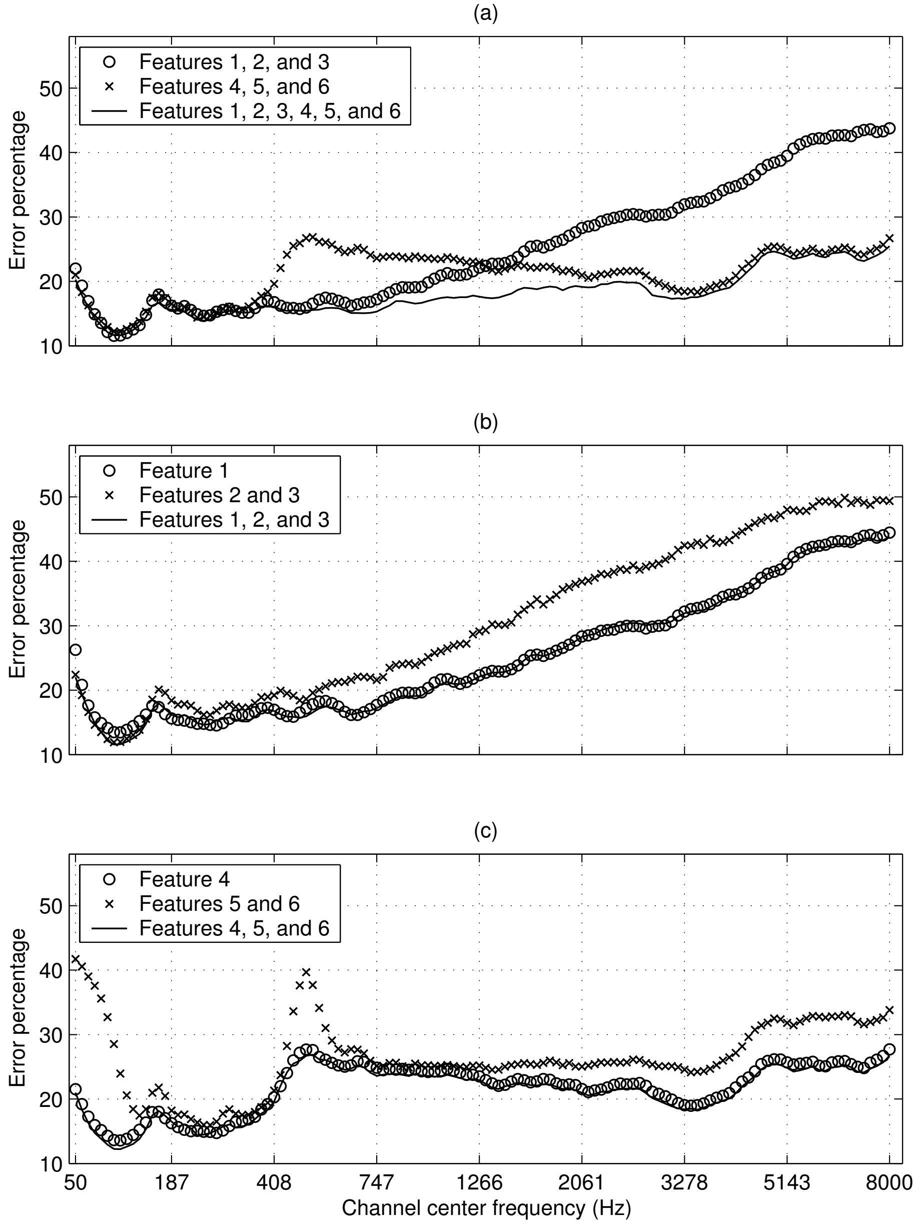

Figure 5.3. Error percentage in labeling T-F units using different features given target pitch. Features 1, 2, 3, 4, 5, and 6 are $A(c, m, \tau_S(m))$, $\mathrm{int}(\bar{f}(c,m)\cdot\tau_S(m))$, $\bar{f}(c,m)\cdot\tau_S(m)-\mathrm{int}(\bar{f}(c,m)\cdot\tau_S(m))$, $A_E(c, m, \tau_S(m))$, $\mathrm{int}(\bar{f}_E(c,m)\cdot\tau_S(m))$, and $\bar{f}_E(c,m)\cdot\tau_S(m)-\mathrm{int}(\bar{f}_E(c,m)\cdot\tau_S(m))$, respectively.

obtains a 37.7% false rejection and a 13.9% false acceptance. The average error rate is 25.8%, which is 2.3% lower than the equal error rate from the previous method. These MLP classifiers perform better than the previous method because they are trained to optimally integrate information from all the features.

Besides using MLPs, we have considered modeling the distribution of the feature set $r_{cm}(\tau)$ for T-F units that are target dominant, i.e., $p(r_{cm}(\tau)|H_0)$, with a Gaussian mixture model (GMM) (Huang et al., 2001) and the same for T-F units that are interference dominant. We then use these models to label T-F units based on Equation (5.4). However, the obtained result is not as good as that from using MLPs since MLPs are trained to distinguish the situation when target is dominant from when interference is dominant and therefore have more discriminative power. We have also considered building a classifier using a support vector machine (SVM) (Vapnik, 1995). In this study, we train the SVM with the software SVM^{light} (Joachims, 1999) with the following kernels: Linear, polynomial, radial, and sigmoid. Among all these kernels, the sigmoid kernel yields the best performance, which is similar to that with MLPs. However, SVM requires much more computation in labeling a T-F unit than MLP.

B. Multiple harmonic sources

When interference contains one or several harmonic signals, there are time frames where both target and interference are pitched. In such a situation, it is more reliable to label a T-F unit by comparing the period of the signal within the unit with both the target pitch period and the interference pitch period. In particular, a unit u_{cm} should be labeled

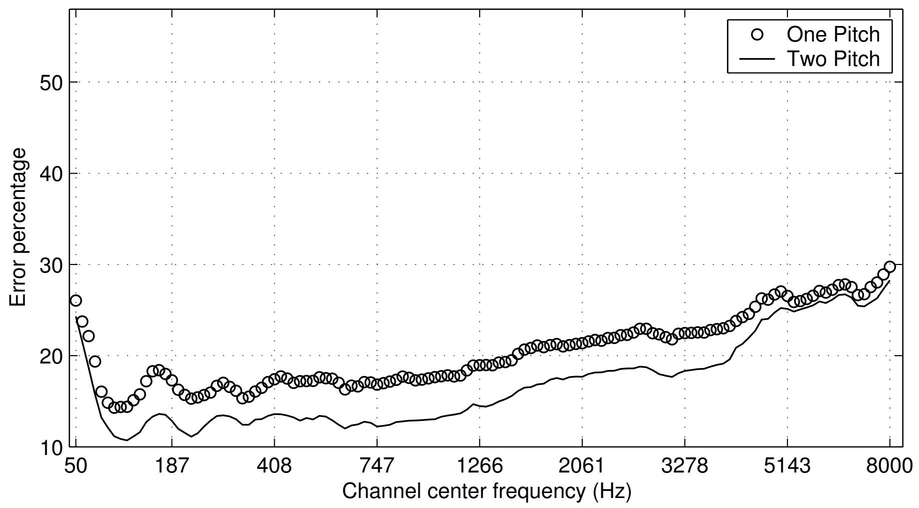

Figure 5.4. Percentage of error in lableing T-F units for the mixtures of two utterances in the test corpus. Circle – labeling T-F units with target pitch. Line – labeling T-F units with both target pitch and interference pitch.

as target if the target pitch period not only matches the period of the signal within this unit but also matches it better than the interference pitch period does, i.e.,

$$\begin{cases} P(H_0 \mid r_{cm}(\tau_S(m))) > P(H_1 \mid r_{cm}(\tau'_S(m))) \\ P(H_0 \mid r_{cm}(\tau_S(m))) > 0.5 \end{cases} \tag{5.7}$$

where $\tau'_S(m)$ is the pitch period of the interfering sound at frame *m*. We use Equation (5.7) to label T-F units for all the mixtures of two utterances in the test corpus (see Section 4.6). Both target pitch and interference pitch are obtained by applying *Praat* to clean utterances. Figure 5.4 shows the corresponding error rate by channel, compared with using only the target pitch to label T-F units. As shown in the figure, better performance is obtained by using pitch information of both speakers.

C. Labeling with information from a neighborhood of T-F units

Labeling a T-F unit using only the local information within the unit still produces a significant amount of error. Since speech signal is wideband and exhibits good temporal continuity, neighboring T-F units potentially provide useful information for T-F unit labeling. For example, a T-F unit surrounded by target-dominant units is more likely target dominant. Therefore, we consider information from a local context. Specifically, we label u_{cm} as target if

$$P(H_0 \,|\, \{P(H_0 \,|\, r_{c'm'}(\tau_S(m')))\}) > 0.5, \ \ |c' - c| \le N_c, |m' - m| \le N_m \qquad (5.8)$$

where N_c and N_m define the size of the neighborhood along frequency and time, respectively, and $\{P(H_0 \,|\, r_{c'm'}(\tau_S(m')))\}$ is the vector that contains the $P(H_0|r_{cm}(\tau_S(m)))$ values of T-F units within the neighborhood. Again, for each filter channel, we train an MLP with one hidden layer to calculate the probability $P(H_0 \,|\, \{P(H_0 \,|\, r_{c'm'}(\tau_S(m')))\})$ using the $P(H_0|r_{cm}(\tau_S(m)))$ values within the neighborhood as features. Since $P(H_0|r_{cm}(\tau_S(m)))$ is derived from $r_{cm}(\tau_S(m))$, we have also considered using $r_{cm}(\tau_S(m))$ directly as features. The resulting MLPs are much more complicated, but yield no performance gain.

The key here is to determine the appropriate size of a neighborhood. Again, we divide the training samples equally into two sets, one for training and the other for validation, and use cross-validation to determine N_c and N_m. Since time and frequency are asymmetric dimensions, we consider them separately. First, we set N_m to 0. An MLP is trained to calculate $P(H_0 \mid \{P(H_0 \mid r_{c'm'}(\tau_S(m')))\})$ with a specific number of N_c for each filter channel. During the training, we set the desired output to 1 if the corresponding T-F unit is target dominant and 0 otherwise. Figures 5.5(a) and 5.5(b) show the average percentages of false rejection and false acceptance for the test corpus, using the obtained MLPs. N_c is from 0 to 10. The values shown in the figure are the average across all the frequency channels. As shown in the figure, the error rate drops significantly when we utilize information from neighboring channels, especially close neighbors. The cross-validation suggests that $N_c = 8$ defines a neighborhood that is sufficient for integrating information across frequency. Therefore, we fix $N_c = 8$ and train an MLP to calculate $P(H_0 \mid \{P(H_0 \mid r_{c'm'}(\tau_S(m')))\})$ with a specific number of N_m for each filter channel. Figures 5.5(c) and 5.5(d) show the average percentages of false rejection and false acceptance for the test corpus, using the obtained MLPs. N_m is from 0 to 5. Again, significant error reduction is obtained when we consider information from neighboring frames. As shown in the figure, by utilizing information from neighboring channels and frames, we reduce the average percentage of false rejection from 20.8% to 16.7% and the average percentage of false acceptance from 13.3% to 8.7% for the test corpus. The cross-validation suggests that $N_c = 8$ and $N_m = 2$ define the appropriate size of the neighborhood. The hidden layer of such a trained MLP has 2 units, which is determined by cross-validation. Note that when both target and interference are pitched, we label a T-F unit according to Equation (5.7) using the probability $P(H_0 \mid \{P(H_0 \mid r_{c'm'}(\tau_S(m')))\})$, and $P(H_1 \mid \{P(H_1 \mid r_{c'm'}(\tau'_S(m')))\})$.

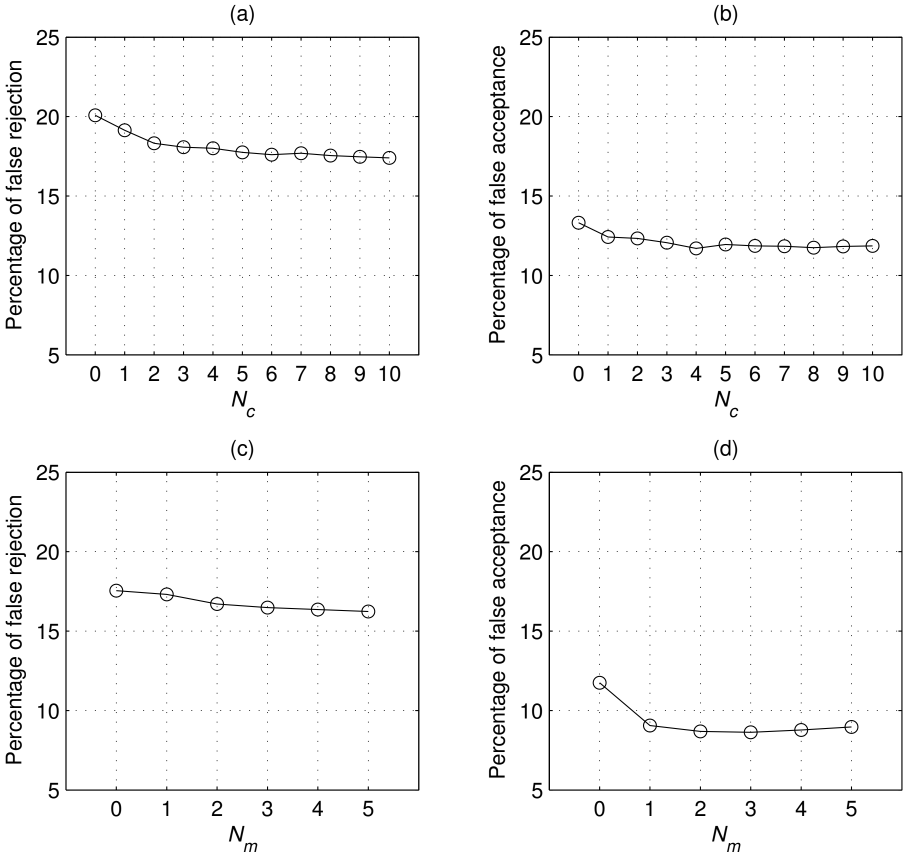

Figure 5.5. Labeling T-F units with information from varied sizes of a local T-F neighborhood. (a) Percentage of false rejection with N_c from 0 to 10 and N_m = 0. (b) Percentage of false acceptance with N_c from 0 to 10 and N_m = 0. (c) Percentage of false rejection with N_c = 8 and N_m from 0 to 5. (d) Percentage of false rejection with N_c = 8 and N_m from 0 to 5.

D. Labeling based on obtained segments

With the segments obtained in the segmentation stage, we have considered labeling T-F units within a segment as a whole instead of labeling them individually, i.e., all the T-F units in a segment are labeled as target if the segment is dominated by voiced target. In particular, we first label a segment as target if

- More than half of its total energy is included in the voiced time frames of target, and
- More than half of its energy in the voiced frames is included in the T-F units labeled as target according to (5.8).

If a segment is labeled as target, all the T-F units within it are labeled as target; otherwise, we keep the labels of individual T-F units. When there are multiple overlapping pitch contours, a segment is attributed to the pitch contour that best matches the period of the signal within the segment.

Figure 5.6 shows the result of labeling T-F units using the estimated segments by channel, compared with labeling T-F units individually with information from a neighborhood. As shown in the figure, using the estimated segments, we recover more target-dominant T-F units. In particular, the average percentage of false rejection is reduced from 16.7% to 12.3%. However, more interference-dominant T-F units are labeled as target at the same time, due to the mismatch error in segmentation. The average percentage of false acceptance increases from 8.7% to 20.0%. As discussed in Section 4.6D, a significant amount of mismatch in segmentation comes from merging

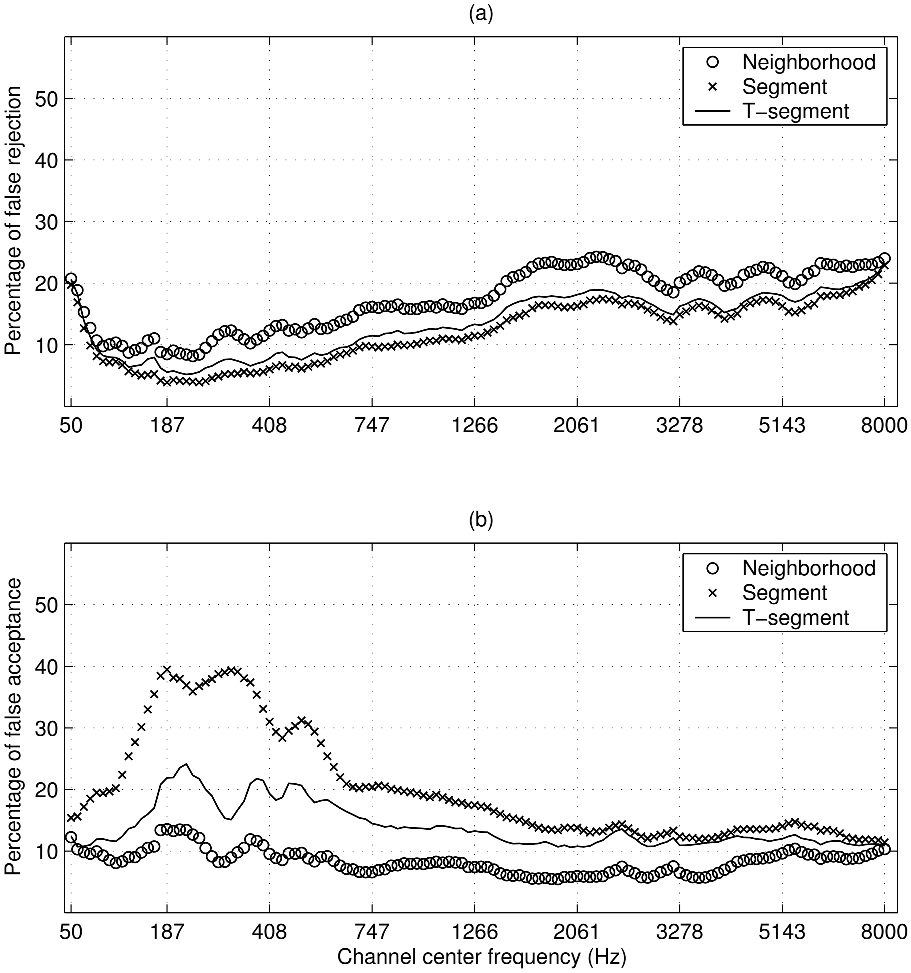

Figure 5.6. Results of labeling T-F units individually using a neighborhood and those using the estimated segments. (a) Percentage of false rejection. (b) Percentage of false acceptance.

channels across frequency. To reduce the false acceptance rate, we use the estimated T-segments, i.e., estimated segments in individual channels (see Section 4.6D), instead of the entire estimated segments to label T-F units. Specifically, if a T-segment is dominated by a voiced target, all the T-F units within the T-segment are labeled as target. The corresponding result of labeling T-F units is shown in Figure 5.6. By using T-segments, we achieve a better balance between accepting target and rejecting interference. In particular, with T-segments, the average percentage of false rejection is 12.1% and the average percentage of false acceptance is 12.8%.

5.2 Pitch determination given voiced target mask

A. Integration across channels

Assume we have an estimated mask of voiced target, i.e., all T-F units considered as target dominant in the voiced region of target are labeled 1 and others are labeled 0. The task here is to estimate target pitch. Let $L(m) = \{L(c, m), \forall c\}$ be the set of mask labels at frame m, where

$$L(c,m) = \begin{cases} 1 & u_{cm} \text{ is labeled as target} \\ 0 & \text{else} \end{cases} \tag{5.9}$$

A frequently-used method for pitch determination is to pool autocorrelations across all the channels and then identify a dominant peak in the summary correlogram (Licklider, 1951; Meddis and Hewitt, 1992). When a harmonic sound is presented, the ACF of the activated filters in a correlogram all exhibit a peak at the delay corresponding to the pitch period (see Figure 3.2). Let $A(m, \tau)$ be the summary correlogram at frame m, that is,

$$A(m,\tau)=\sum_{c}A(c,m,\tau) \tag{5.10}$$

The estimated pitch period at frame m, $\tau_S(m)$, is the lag corresponding to the maximum of $A(m, \tau)$ in the plausible pitch range of the target utterance. As shown in the bottom panel of Figure 3.2(a), there is a significant peak at 5.87 ms, corresponding to the target pitch period at this frame, in the summary correlogram. This method of pitch estimation is not very robust when interference is strong because the autocorrelations in many channels exhibit spurious peaks not corresponding to the target period. One may solve this problem by disregarding interference-dominant T-F units, i.e., calculating the summary correlogram only with T-F units labeled 1:

$$A(m,\tau)=\sum_{c}A(c,m,\tau)L(c,m) \tag{5.11}$$

Again, the estimated pitch period is the lag of the global maximum of $A(m, \tau)$ in the plausible pitch range.

Similar to the ACF of filter response, the profile of the probability that a T-F unit u_{cm} being target dominant given pitch period τ, $P(H_0|r_{cm}(\tau))$, also tends to have a significant peak at the target period when u_{cm} is truly target dominant (see Figure 5.2(c)). One can use the corresponding summation of $P(H_0|r_{cm}(\tau))$,

$$SP_m(\tau)=\sum_{c}P(H_0 \mid r_{cm}(\tau))L(c,m)\,, \tag{5.12}$$

to identify the pitch period at frame m as the maximum of the summation in the plausible pitch range.

Method	Summary ACF		Summary $P(H_0 \vert r_{cm}(\tau))$		Classifier	
	F	M	F	M	F	M
Without temporal continuity	39.6	17.1	18.1	17.2	15.6	17.6
With temporal continuity	31.8	16.3	14.8	15.8	12.7	16.8

Table 5.1. Error rate of pitch estimation with ideal binary mask. Classifier – pitch estimation with a classifier that compares two pitch candidates using their relative locations and the summation of $P(H_0|r_{cm}(\tau))$ as features.

We apply the above two methods for pitch estimation to two utterances from in test corpus listed in Table 4.1, one from a female speaker (S1) and the other from a male speaker (S2). These two utterances are mixed with the 20 intrusions listed in Table 4.2 at 0 dB SNR. In this estimation, we use the ideal binary mask at the voiced frames of the target utterance to estimate a pitch period at each frame. The percentages of estimation error for both methods are shown in the first row of Table 5.1. We use the pitch contours obtained by applying *Praat* to the clean target as the ground truth of the target pitch. An error occurs when the estimated pitch period and the pitch period obtained from *Praat* differ by more than 5%. As shown in the table, using the summation of $P(H_0|r_{cm}(\tau))$ performs much better than using the summary ACF for the female utterance.

Both methods, especially the one using summary ACF, perform better on the male utterance than on the female utterance. This is because the ACF and $P(H_0|r_{cm}(\tau))$ in target-dominant T-F units all exhibit peaks not only at the target pitch period, but also at its double, triple, or other multiples. As a result, their summations have a significant peak not only at the target pitch period, but also at its integer multiples. As shown in the bottom panel of Figure 3.2(a), there is a significant peak at 11.74 ms, corresponding to the double of the target period at this frame, in the summary correlogram. It is very likely, especially for a female utterance, that the plausible pitch range contains several integer multiples of the target pitch period. In this situation, the above methods will make a mistake when the highest peak does not correspond to target period, but some multiple of it.

B. Differentiating true pitch period from its integer multiples

To differentiate the target pitch period from its integer multiples in pitch estimation, we need to take the relative locations of possible pitch candidates into consideration. Let τ_1 and τ_2 be two pitch candidates. We train an MLP-based classifier that selects the better one from these two candidates using their relative locations and the summation of $P(H_0|r_{cm}(\tau))$, $SP_m(\tau)$, as features, i.e., $(\tau_1/\tau_2,\ SP_m(\tau_1), SP_m(\tau_2))$. The training data from the mixtures combining the training part of the TIMIT database and the 100 environmental sounds used in Section 5.1A. In constructing the training data, we obtain $SP_m(\tau)$ at each time frame from all the target-dominant T-F units. In each training sample, the two pitch candidates are the true target pitch period and the lag of another peak of $SP_m(\tau)$ within the plausible pitch range. Without loss of generality, we let $\tau_1 < \tau_2$. The desired output is 1 if τ_1 is the true pitch period and 0 otherwise. The obtained MLP has 3 units in the hidden layer. We use the obtained MLP to select the better one from the two candidates as follows: If the output of the MLP is higher than 0.5, we consider τ_1 as the better candidate; otherwise, we consider τ_2 as the better candidate.

The target pitch is estimated with the classifier as follows:

- Find all the local maxima in $SP_m(\tau)$ within the plausible pitch range as pitch candidates. Sort these candidates according to their time lags from small to large and let the first candidate be the current estimated pitch period, $\tau_S(m)$.
- Compare the current estimated pitch period with the next candidate using the obtained MLP and update the pitch estimate if necessary.

The percentage of error for pitch estimation with the classifier is shown in the first row in Table 5.1. The classifier reduces the error rate on the female utterance but slightly increases the error rate on the male utterance.

C. Pitch estimation using temporal continuity

Speech signals have good temporal continuity, i.e., their structure, such as frequency partials, tends to last for a certain period of time and the signals change smoothly within that period. Consequently, the pitch and the ideal binary mask of a target utterance tend to have good temporal continuity as well. Figure 5.7 shows the histogram of the relative changes of pitch periods in consecutive frames for utterances in the training part of the TIMIT database (Wu et al., 2003). As shown in the figure, the pitch period changes slowly. In fact, less than 0.5% of consecutive frames have more than 20% relative pitch

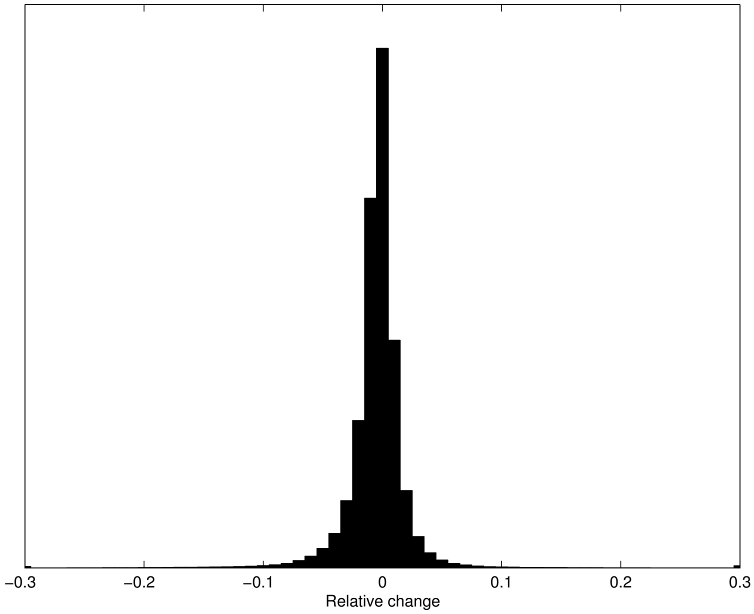

Figure 5.7. Histogram of relative pitch change of speech utterances between consecutive frames.

changes. Thus we utilize pitch continuity to improve the accuracy of pitch estimation as follows:

- First, we check the reliability of the estimated pitch based on temporal continuity. Specifically, for every three consecutive time frames, *m*–1, *m*, and *m*+1, if the pitch changes among these frames are all less than 20%, i.e.,

$$\begin{cases} |\tau_S(m) - \tau_S(m-1)| < 0.2\min(\tau_S(m), \tau_S(m-1)) \\ |\tau_S(m) - \tau_S(m+1)| < 0.2\min(\tau_S(m), \tau_S(m+1)) \end{cases} \quad (5.13)$$

  the estimated pitch periods in these three frames are all considered reliable.

- Second, we re-estimate unreliable pitch points by limiting the plausible pitch range according to neighboring reliable pitch points. Specifically, for two consecutive time frames, $m-1$ and m, if $\tau_S(m)$ is reliable and $\tau_S(m-1)$ is unreliable, we re-estimate $\tau_S(m-1)$ by limiting the plausible pitch range for $\tau_S(m-1)$ to be $[0.8\tau_S(m), 1.2\tau_S(m)]$. On the other hand, if $\tau_S(m-1)$ is reliable and $\tau_S(m)$ is unreliable, we re-estimate $\tau_S(m)$ by limiting the plausible pitch range for $\tau_S(m)$ to be $[0.8\tau_S(m-1), 1.2\tau_S(m-1)]$. Another possible situation is that $\tau_S(m)$ is unreliable while both $\tau_S(m-1)$ and $\tau_S(m+1)$ are reliable. In this case, we use $\tau_S(m-1)$ to limit the plausible pitch range of $\tau_S(m)$ if the mask at frame m is more similar to the mask at frame $m-1$ than the mask at frame $m+1$, i.e.,

$$\sum_c L(c,m)L(c,m-1) > \sum_c L(c,m)L(c,m+1)\,; \tag{5.14}$$

  otherwise, $\tau_S(m+1)$ is used to re-estimate $\tau_S(m)$. Then the re-estimated pitch points are considered as reliable and used to estimate unreliable pitch points in their neighboring frames. This re-estimation process stops when all the unreliable pitch points have been re-estimated.

The second row in Table 5.1 shows the effect of incorporating temporal continuity in pitch estimation with the algorithms described above. Using temporal continuity yields consistent performance improvement, especially for the female utterance.

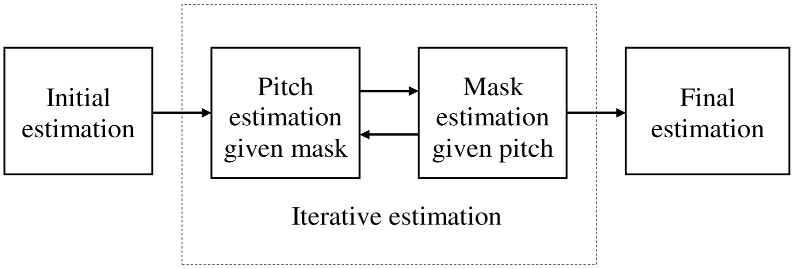

Figure 5.8. Diagram of the proposed iterative algorithm.

5.3 An iterative procedure for pitch estimation and voiced speech segregation

The diagram of our iterative algorithm is shown in Figure 5.8. It contains three steps: Initial estimation, iterative estimation, and final estimation. In the first step, our algorithm generates an initial estimate of pitch contours and masks for up to two sources. In the second step, the algorithm improves the estimation of pitch contours and masks in an iteratively manner. Specifically, we apply the algorithm described in Section 5.2B and the re-estimation process described in Section 5.2C for pitch estimation. Masks are estimated by the algorithm described in Section 5.1C, which labels T-F units individually with a neighborhood. We choose not to use the algorithm for mask estimation with obtained segments described in Section 5.1D in this step since this algorithm is more liberal in labeling T-F units as target. As discussed before, we do not need all the target signals to estimate target pitch. Therefore, we prefer to be conservative in mask generation for pitch estimation, which tends to yield more accurate pitch estimation. Once the iteration stops, we use the obtained T-segments to improve mask estimation (see Section 5.1D) in the final estimation step. The details of the first two steps are given in the following subsections.

A. Initial Estimation

In this step, we first generate up to two estimated pitch periods in each time frame. Since T-F units dominated by a periodic signal tend to have high cross-channel correlations of the filter response or the response envelope, we only consider T-F units with high cross-channel correlations in this estimation. Let $\tau_{S,1}(m)$ and $\tau_{S,2}(m)$ represent the two estimated pitch periods at frame m, and $L_1(m)$ and $L_2(m)$ the corresponding labels of the estimated masks. We first treat all the T-F units with high cross-channel correlations as dominated by a single source. That is:

$$L_1(c,m) = \begin{cases} 1 & C(c,m) > 0.985 \;\; or \;\; C_E(c,m) > 0.985 \\ 0 & else \end{cases} \tag{5.15}$$

We then assign the time delay supported by most T-F units labeled 1 as the first estimated pitch period. A unit u_{cm} is considered supporting a pitch candidate τ if the corresponding $P(H_0|r_{cm}(\tau))$ is higher than a threshold. Accordingly we have:

$$\tau_{S,1}(m) = \arg\max_{\tau} \sum\nolimits_c L_1(c,m) \cdot \mathrm{sgn}(P(H_0 \mid r_{cm}(\tau)) - \theta_P) \tag{5.16}$$

where $\mathrm{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$,

and θ_P is a threshold. Intuitively, we can set θ_P to 0.5. However, such a threshold may not position the estimated pitch period close to the true pitch period because $P(H_0|r_{cm}(\tau))$

tends to be higher than 0.5 in a relatively wide range centered at the true pitch period (see Figure 5.2(c)). In general θ_P needs to be much higher than 0.5 so that we can position $\tau_{S,1}(m)$ more accurately. However, θ_P cannot be too high, otherwise most T-F units labeled 1 cannot contribute to this estimation. We found that 0.75 is a good threshold that allows us to accurately position $\tau_{S,1}(m)$ without ignoring many T-F units labeled 1.

The above process yields an estimated pitch point at many time frames where the target is not pitched. The estimated pitch point at such a frame is usually supported by only a few T-F units unless the interference contains a strong harmonic signal at this frame. On the other hand, estimated pitch points corresponding to target pitch are usually supported by quite a few T-F units. In order to remove spurious pitch points, we discard a detected pitch point if the total number of channels supporting this pitch point is less than a threshold. We found that the appropriate value of this threshold is 7 from analyzing the training data described in Section 5.1A. Most spurious pitch points are thus removed. At the same time, some true pitch points are also removed, but most of them are recovered in the following iterative process.

With the estimate pitch period $\tau_{S,1}(m)$, we re-estimate the mask $L_1(m)$ as:

$$L_1(c,m) = \begin{cases} 1 & P(H_0 \mid r_{cm}(\tau_{S,1}(m))) > 0.5 \\ 0 & else \end{cases} \tag{5.17}$$

Then we use the T-F units that do not support the first pitch period $\tau_{S,1}(m)$ to estimate the second pitch period, $\tau_{S,2}(m)$. Specifically,

$$L_2(c,m)=\begin{cases}1 & P(H_0 \mid r_{cm}(\tau_{S,1}(m))) \le \theta_P \;\; and \;\; (C(c,m)>0.985 \;\; or \;\; C_E(c,m)>0.985) \\ 0 & else\end{cases}$$

(5.18)

We let

$$\tau_{S,2}(m)=\arg\max_{\tau}\sum\nolimits_c L_2(c,m)\cdot \mathrm{sgn}(P(H_0 \mid r_{cm}(\tau))-\theta_P) \tag{5.19}$$

Again, if fewer than 7 T-F units support $\tau_{S,2}(m)$, we set it to 0. Otherwise, we re-estimate $L_2(m)$ as:

$$L_2(c,m)=\begin{cases}1 & P(H_0 \mid r_{cm}(\tau_{S,2}(m))) > 0.5 \\ 0 & else\end{cases} \tag{5.20}$$

Here we estimate up to two pitch points at one frame. Nevertheless, one can easily extend the above algorithm to estimate pitch points of more sources if needed.

After the above estimation, our algorithm combines the estimated pitch periods into pitch contours based on temporal continuity. Specifically, for estimated pitch periods in three consecutive frames, $\tau_{S,k_1}(m-1)$, $\tau_{S,k_2}(m)$, and $\tau_{S,k_3}(m+1)$, where k_1, k_2, and k_3 are either 1 or 2, they are combined into one pitch contour if they have good temporal continuity and their masks also have good temporal continuity. That is,

$$\begin{cases} |\tau_{S,k_2}(m)-\tau_{S,k_1}(m-1)| < 0.2\min(\tau_{S,k_2}(m),\tau_{S,k_1}(m-1)) \\ |\tau_{S,k_2}(m)-\tau_{S,k_3}(m+1)| < 0.2\min(\tau_{S,k_2}(m),\tau_{S,k_3}(m+1)) \\ \sum_c L_{k_2}(c,m)L_{k_1}(c,m-1) > 0.5\max(\sum_c L_{k_2}(c,m),\sum_c L_{k_1}(c,m-1)) \\ \sum_c L_{k_2}(c,m)L_{k_3}(c,m+1) > 0.5\max(\sum_c L_{k_2}(c,m),\sum_c L_{k_3}(c,m+1)) \end{cases} \tag{5.21}$$

The remaining isolated estimated pitch points are considered unreliable and set to 0. Note that requiring only the temporal continuity of pitch periods cannot prevent connecting pitch points from different sources, since target and interference may have similar pitch periods at the same time. However, it is very unlikely that target and interference both have similar pitch periods and occupy the same frequency region at the same time. In most situations, pitch points that are connected according to Equation (5.21) do correspond to a single source. As a result of this step, we obtain multiple pitch contours and each pitch contour has an associated T-F mask.

B. Iterative pitch estimation

In this step, we first re-estimate each pitch contour from the associated mask. A key step in this estimation is to expand the estimated pitch contours based on temporal continuity, i.e., using reliable pitch points to estimate potential pitch points at neighboring frames. Specifically, let τ_k be a pitch contour and $L_k(m)$ be the associated mask. Let m_1 and m_2 be the first and the last frame of this pitch contour. To expand pitch contour τ_k, we first let $L_k(m_1-1) = L_k(m_1)$ and $L_k(m_2+1) = L_k(m_2)$. Then we re-estimate τ_k from this new mask using the algorithm described in Section 5.2B. The re-estimated pitch periods are further verified according to temporal continuity, as described in Section 5.2C except that, here, we use Equation (5.21) instead of Equation (5.13) for continuity verification. If the corresponding source of contour τ_k is pitched at frame m_1-1, our algorithm likely yields an accurate pitch estimate at this frame. Otherwise, the re-estimated pitch period at this frame usually cannot pass the continuity check, and as a result it is discarded and τ_k still starts from frame m_1. The same is for the estimated pitch period at frame m_2+1.

After expansion and re-estimation, two pitch contours may have the same pitch period at the same frame and therefore they are combined into one pitch contour.

Then we re-estimate the mask for each pitch contour as follows. First, we compute the probability of each T-F unit dominated by the corresponding source of a pitch contour *k*, $P(H_0 \mid \{P(H_0 \mid r_{c'm'}(\tau_k(m')))\})$, as described in Section 5.1C. Then we estimate the mask for contour *k* according to the obtained probabilities:

$$L_k(c,m) = \begin{cases} 1 & k = \arg\max_{k'} P(H_0 \mid \{P(H_0 \mid r_{c'm'}(\tau_{k'}(m')))\}) \text{ and} \\ & P(H_0 \mid \{P(H_0 \mid r_{c'm'}(\tau_k(m')))\}) > 0.5 \\ 0 & else \end{cases} \tag{5.22}$$

The iterative algorithm is stopped when the estimation of both pitch and mask converges or runs into a cycle, where there are slight cyclic changes for both estimated pitch and estimated mask after each iteration. Although we cannot guarantee this algorithm of stopping, in our evaluation this algorithm always stops after a small number of iterations, which is usually smaller than 20.

Figure 5.9 shows the detected pitch contours for the mixture M1 shown in Figure 1.2(d). We use the pitch points detected by *Praat* from the clean utterance as the ground truth of the target pitch. As shown in the figure, our algorithm correctly estimates most target pitch points. At the same time, it also yields one pitch contour for interference. Figure 5.10(a) shows the target mask obtained by labeling T-F units with information from a fixed neighborhood, i.e., the mask obtained right before the final estimation step. Comparing this mask with the ideal binary mask of the target shown in Figure 5.10(e), we can see that our system is able to segregate most voiced portions of the target without

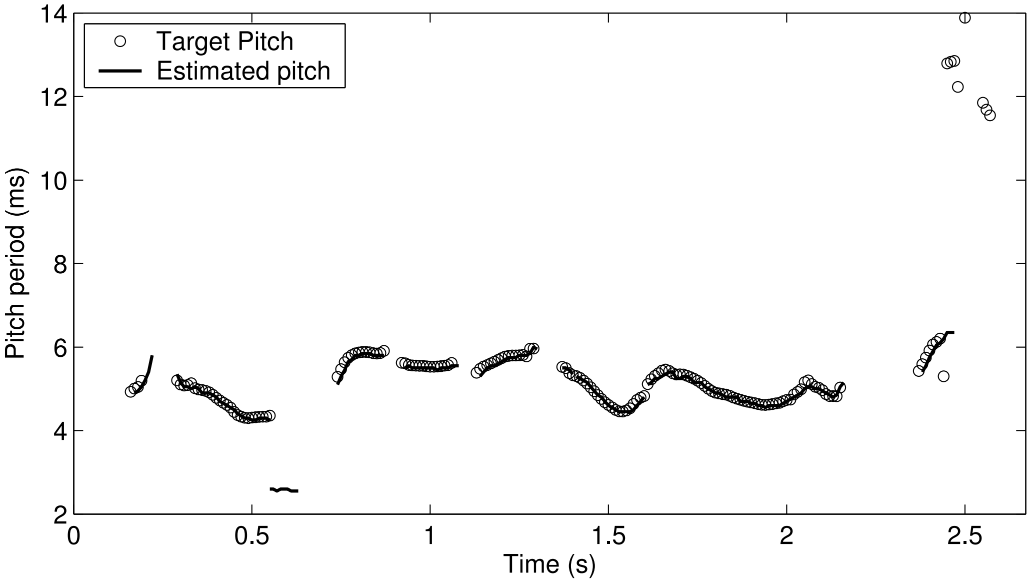

Figure 5.9. Estimated pitch contours for the mixture M1 shown in Figure 1.2(d). Solid lines indicate estimated target pitch contours. True pitch points are marked by circles.

including much interference. These two masks yield similar resynthesized targets, as shown in Figures 5.10(b) and 5.10(f). Figure 5.10(c) shows the target mask obtained by labeling T-F units with the estimated T-segments, which is the mask obtained in the final estimation step. With the estimated T-segments, the system is able to recover even more target energy, but at the expanse of adding a small amount of the interference. Note that the algorithm does not specify whether an estimated contour is from target or from interference. This is a task of sequential grouping to be discussed in the next chapter. The above masks are obtained by assuming perfect sequential grouping, i.e., we group all the masks corresponding to the target utterance to form the segregated target stream.

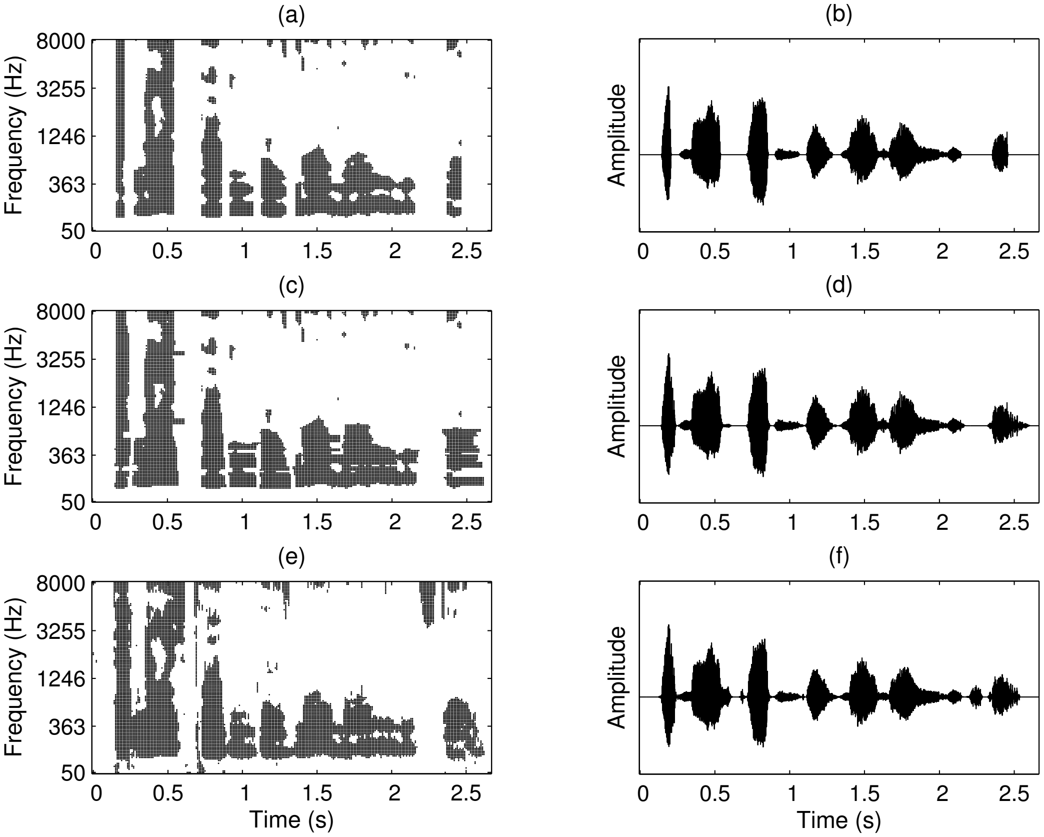

Figure 5.10. Segregated voiced utterance from the mixture M1 shown in Figure 1.2(d). Dark regions indicate T-F units labeled as target. (a) Mask of target segregated by labeling T-F units with information from a neighborhood. (b) Waveform of target resynthesized with the mask in (a). (c) Mask of target segregated using estimated T-segments. (d) Waveform of target resynthesized with the mask in (c). (e) Ideal binary mask of target. (f) Waveform resynthesized from the ideal binary mask.

5.4 Evaluation

A. Pitch Estimation

We first evaluate the above iterative algorithm on pitch determination with utterances from the FDA Evaluation Database (Bagshaw et al., 1993). This database is collected for evaluating pitch determination algorithms and it provides accurate target pitch contours derived from laryngograph data, which are used as the ground truth of target pitch. The database contains utterances from two speakers, one male and one female. We randomly select one sentence that is uttered by both speakers. These two utterances are mixed with the intrusions listed in Table 4.2, at different SNR levels. In order to test the performance of the iterative algorithm when two speakers utter the same sentence simultaneously, we mix the two selected utterances of the same written sentence at different SNR levels and replace the mixture of target and intrusion N20. Figure 5.11 shows the detected pitch contours for the 0 dB mixture of the two target utterances. Most target pitch contours of the utterances are correctly detected by the iterative algorithm.

Figure 5.12(a) shows the average correct percentage of pitch determination with the iterative algorithm on the mixtures of these two target utterances and the 20 intrusions at different SNR levels. Here we compare the estimated pitch contours with only the true pitch contours of the target utterance. In calculating the correct detection percentage, we only consider estimated pitch contours that match the target pitch. An estimated pitch contour matches target pitch if at least half of its pitch points match the target pitch, i.e., the target are pitched at these corresponding frames and the estimated pitch periods differ from the true target pitch periods by less than 5%. As shown in the figure, the iterative

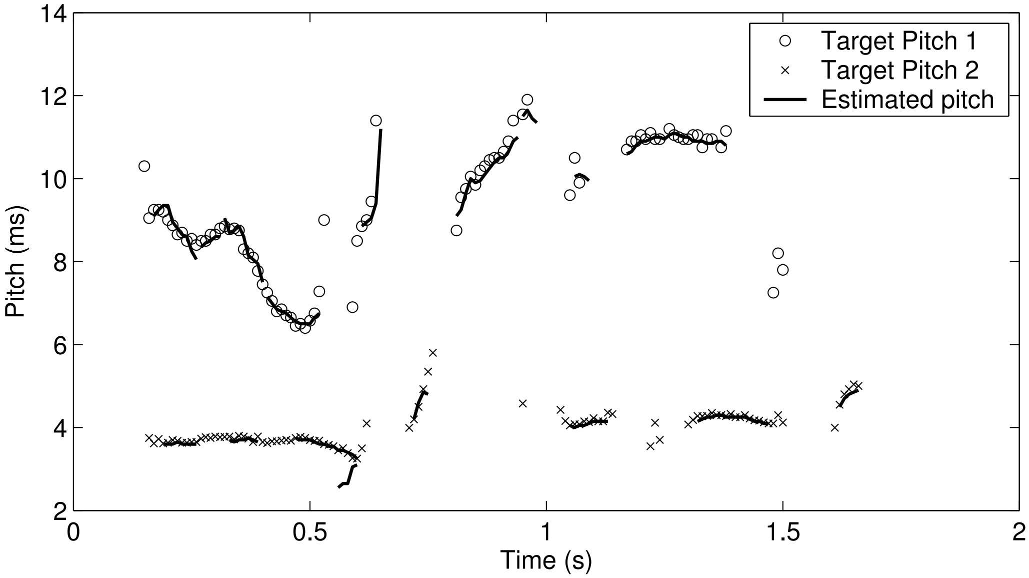

Figure 5.11. Estimated pitch contours for a mixture of one male utterance and one female utterance. Solid lines indicate estimated target pitch contours. True pitch points are marked by "o" and "x".

algorithm is able to detect 69.1% of target pitch even at -5 dB SNR. The correct detection rate increases to about 83.8% as the SNR increases to 15 dB. In comparison, Figure 12(a) also shows the results using *Praat* and that from a multiple pitch tracking algorithm proposed by Wu et al. (2003), which is a state-of-the-art algorithm for robust pitch tracking (Khurshid and Denham, 2004). Note that the Wu et al. algorithm does not yield continuous pitch contours. Therefore, the correct detection rate is computed by comparing estimated pitch with ground truth frame by frame. As shown in the figure, the iterative algorithm performs consistently better than the Wu et al. algorithm in detecting target pitch at all SNR levels. The iterative algorithm is more robust to interference

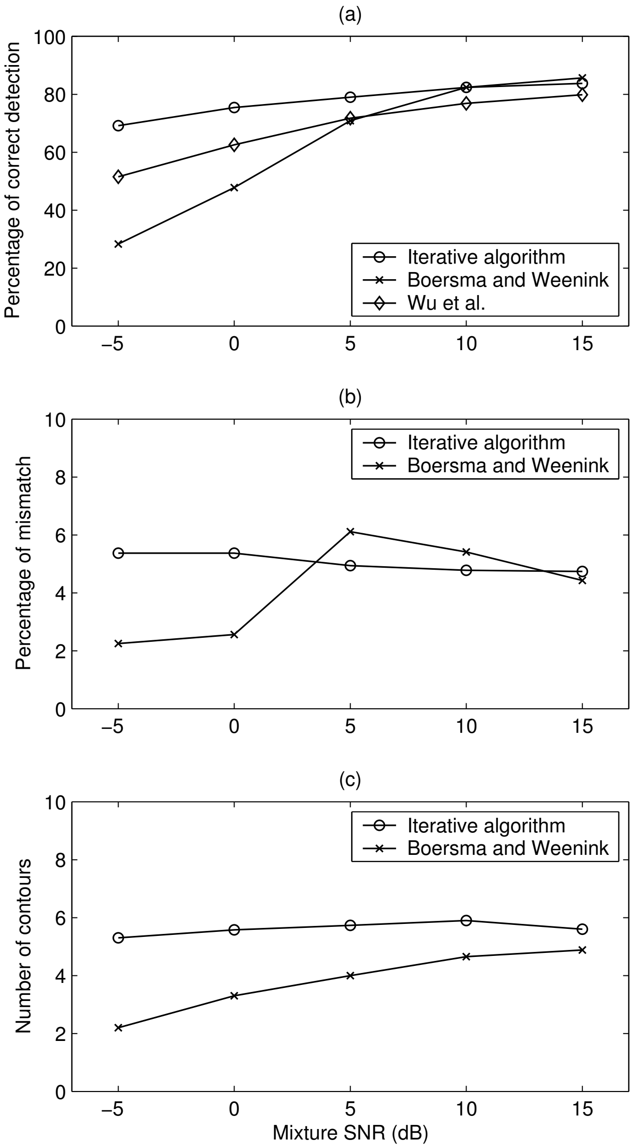

Figure 5.12. Results of pitch determination. (a) Percentage of correct detection. (b) Percentage of mismatch. (c) Number of contours that match the target pitch.

compared to *Praat*, whose performance is good at an SNR level above 10 dB, but drops significantly as SNR decreases.

Besides the detection rate, we also need to measure how well the system separates pitch points of different sources. Figure 5.12(b) shows the percentage of mismatch, which is the percentage of estimated pitch points that do not match target pitch among the contours matching the target pitch. An estimated pitch point is counted as mismatch if either target is not pitched at the corresponding frame or the difference between the estimated pitch period and the true target pitch period is more than 5%. As shown in the figure, the iterative algorithm yields a low percentage of mismatch, which is slightly lower than that of *Praat* when the SNR is above 5 dB SNR. In lower SNR levels, *Praat* has a lower percentage of mismatch, due to the fact that *Praat* detects fewer pitch points. Note that the Wu et al. algorithm does not generate pitch contours, and the mismatch rate is 0. In addition, Figure 5.12(c) shows the average number of estimated pitch contours that match the target pitch. The actual average number of target pitch contours is 5. The iterative algorithm yields an average of 5.6 pitch contours for each mixture. This shows that the iterative algorithm well separates target pitch and interference pitch without dividing target pitch into many short contours. *Praat* yields almost the same numbers of contours as the actual ones at 15 dB SNR. However, it detects fewer pitch contours when the mixture SNR drops. Overall, the iterative algorithm yields better performance than both *Praat* and the Wu et al. algorithm, especially at low SNR levels.

To illustrate the advantage of the iterative process for pitch estimation, Table 5.2 shows the average percentage of correct detection for the above mixtures at -5 dB with

Iteration	0	1	2	3	4	Converge
Percentage of detection	63.0	66.3	67.8	68.8	68.9	69.1

Table 5.2. Average percentage of correct pitch detection with respect to the number of iteration

respect to the number of iteration. In this table, the zero iteration corresponds to the result of initial estimation, and "converge" corresponds to the final output of the algorithm. As shown in the table, the initial estimation already gives a good estimate of target pitch. The iterative algorithm, however, is able to improve the detection rate, especially in the first iteration. Overall, the iterative algorithm improves the detection rate by 6.1% on average. This improvement is not large since the initial estimate is already good. However, the improvement varies considerably among different mixtures. The largest improvement of individual mixtures is 22.1% in our test.

B. Voiced speech segregation

The performance of the system on voiced speech segregation has been evaluated with the test corpus described in Section 4.6. The estimated target masks are obtained by assuming perfect sequential grouping, i.e., all the estimated masks corresponding to the target are grouped together to form the segregated target.

Since our computational goal here is to estimate the ideal binary mask of target, we evaluate the performance of segregation by comparing the estimated mask to the ideal binary mask with two measures (Hu and Wang, 2004a).

- The percentage of energy loss, P_{EL}, which measures the amount of energy in the target-dominant T-F units that are labeled as interference among the total energy in target-dominant T-F units.
- The percentage of noise residue, P_{NR}, which measures the amount of energy in the interference-dominant T-F units that are labeled as target among the total energy in T-F units estimated as target dominant.

P_{EL} and P_{NR} provide complementary error measures of a segregation system and a successful system needs to achieve low errors in both measures.

In addition, to compare waveforms directly we measure the SNR of the segregated voiced target in decibels:

$$SNR = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n [s(n) - \hat{s}_V^2(n)]^2} \tag{5.24}$$

where $s(n)$ is the target signal resynthesized from the ideal binary mask and $\hat{s}_V(n)$ is the segregated voiced target.

The results from our system are shown in Figure 5.13. Each point in the figure represents the average value of 400 mixtures in the test corpus at a particular SNR level (see Section 4.6). Figures 5.13(a) and 5.13(c) show the percentages of energy loss and noise residue. Note than since our goal here is to segregate voiced target, the P_{EL} values shown in Figure 5.13(a) are only for the target energy at the voiced frames of the target.

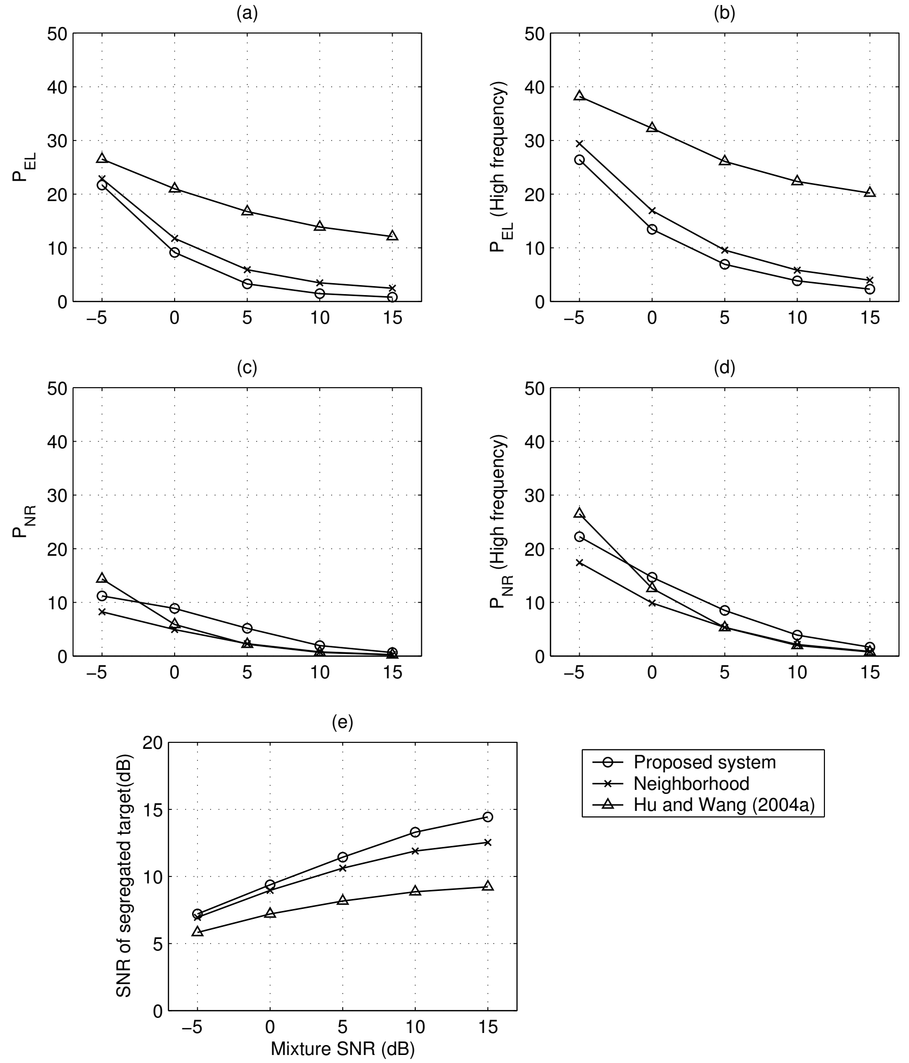

Figure 5.13. Results of voiced speech segregation. (a) Percentage of energy loss on voiced target. (b) Percentage of energy loss on voiced target in the frequency range above 1 kHz. (c) Percentage of noise residue. (d) Percentage of noise residue in the frequency range above 1 kHz. (e) SNR of segregated voiced target.

As shown in the figure, our system segregates 78.3% of voiced target energy at -5 dB SNR and 99.2% at 15 dB SNR. At the same time, 11.2% of the segregated energy belongs to intrusion at -5 dB. This number drops to 0.6% at 15 dB SNR. Figures 5.13(b) and 5.13(d) show the percentage of energy loss and that of noise residue for T-F regions about 1 kHz. As shown in the figure, our system performs better in the low-frequency range than in the high-frequency. Nevertheless, it still captures a majority of target energy and rejects most interfering energy in the high-frequency range.

Figure 5.13(e) shows the SNR of the segregated target. Our system obtains an average 12.2 dB gain in SNR when the mixture SNR is -5 dB. This gain drops to 3.3 dB when the SNR of original mixtures is 10 dB. Note that at 15 dB, our system does not improve the SNR because most unvoiced speech is not segregated. Figure 5.13 also shows the result of the iterative algorithm without the final estimation step ("Neighborhood"), i.e., labeling T-F units individually without using estimated T-segments. As shown in the figure, the corresponding segregated target loses more target, but contains less interference. The SNR performance is better when using estimated T-segments.

In comparison, Figure 5.13 also shows the performance using our previous voiced speech segregation system (Hu and Wang, 2004a; Section 2.3), which is the representative of CASA systems. Because the previous system can only track one pitch contour of the target, in this implementation we provide target pitch estimated by applying *Praat* to clean utterances. As shown in the figure, the previous system yields a lower percentage of noise residue, but has a much higher percentage of energy loss, especially in the high-frequency range. Even with provided target pitch, the previous

Iteration	0	1	2	3	4	Converge
SNR (dB)	6.97	7.44	7.62	7.77	7.89	8.04

Table 5.3. Average SNR of segregated target utterances with different number of iterations

system does not perform as well as the iterative algorithm, especially at high SNR levels. Computationally, the current system is more complicated than our previous system. The additional computation mainly comes from the multiscale segmentation process and the iterative algorithm. Nevertheless, the additional computation is less than 10% of that needed in calculating the correlogram, which consumes the major portion of the overall computation.

To illustrate the effect of the iterative estimation, Table 5.3 shows the average SNR for the mixtures of two utterances, S1 and S2 (see Table 4.1), and all the intrusions at -5 dB SNR (see Table 4.2). Again, in this table, the zero iteration corresponds to the result of the initial estimation, and "converge" corresponds to the final output of the algorithm. On average, the iterative algorithm improves the SNR by 1.07 dB. Again, the SNR improvement varies considerably among different mixtures. The largest improvement of individual mixtures is 7.27 dB.

As an additional benchmark, we have evaluated our algorithm on Cooke's test corpus. The average SNR for each intrusion is shown in Figure 5.14, compared with those of the original mixtures and our previous system (Hu and Wang, 2004a). The proposed system obtains results comparable to our previous system for most intrusions, except for N0, N2,

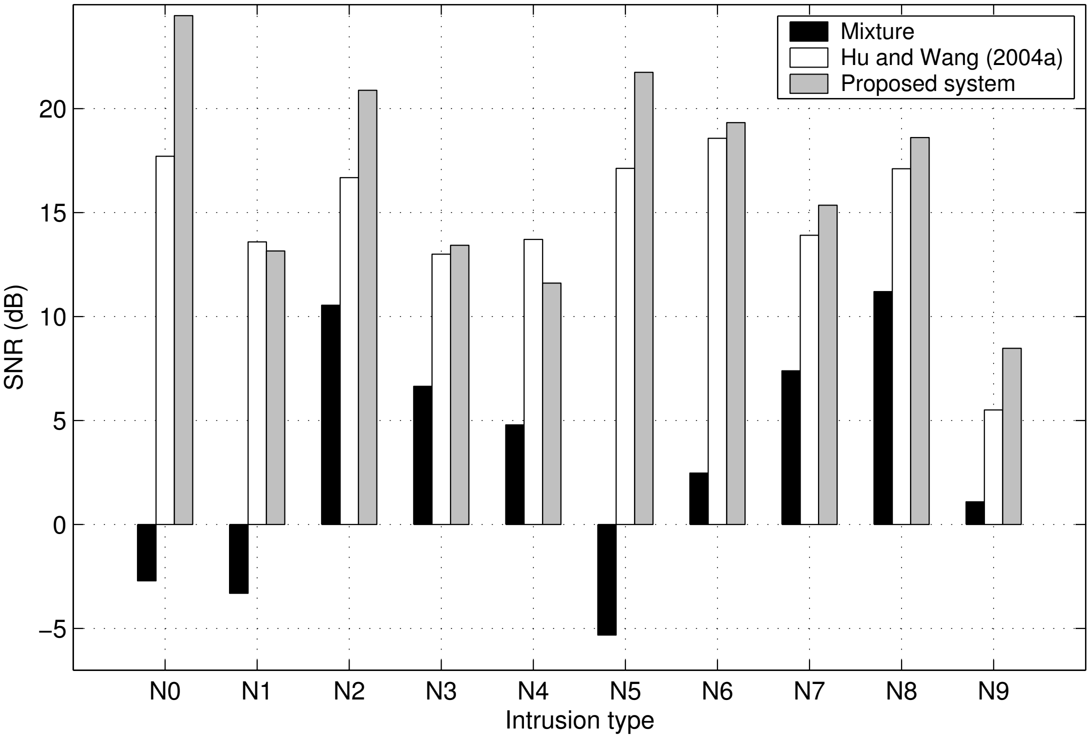

Figure 5.14. SNR results for segregated speech and original mixtures.

and N5, in which the proposed system has at least 4 dB more in SNR improvement than our previous system. On average, the proposed system obtains a 13.4 dB SNR gain, which is about 2.0 dB better than our previous system.

CHAPTER 6

SEQUENTIAL GROUPING USING FEATURE-BASED CLASSIFICATION

In the previous chapter, we presented our iterative algorithm for estimating target pitch and segregating voiced targets. For an acoustic mixture, the outcome of the algorithm is several estimated pitch contours and estimated binary masks associated with individual pitch contours. An obtained mask labels the T-F units within the time interval of the corresponding estimated pitch contour, where a T-F unit is labeled 1 if it is dominated by the source of the corresponding pitch contour and 0 otherwise. This mask is used to resynthesize a segregated voiced sound from the mixture. Since interfering sounds often contain periodic signals, a segregated voiced sound may correspond to interference. The iterative algorithm does not determine whether or not this segregated sound is part of a target utterance, which is a problem of sequential grouping as discussed in Section 2.4C. A systematic study of sequential grouping is beyond the scope of this dissertation. Here we focus on a common situation encountered in practice: Target utterances corrupted by non-speech interference. In this situation, we propose to handle the sequential organization problem by first classifying whether a segregated voiced sound is speech or non-speech and then grouping all the sounds that are classified as speech. In addition, we apply the above procedure for segregating unvoiced speech. In the segmentation stage, we have obtained segments for voiced target, unvoiced target, and interference. The task here is to identify those segments corresponding to unvoiced target and group them together with the segregated voiced target. Again, with the assumption that interference is non-speech, we handle this problem by classifying a segment as dominated by either unvoiced speech or interference and grouping it accordingly.

This chapter gives a detailed description of the above procedures that sequentially group voiced and unvoiced speech. In Section 6.1, we discuss the features used for distinguishing speech from non-speech interference. In the next two sections, we describe the sequential grouping method for voiced speech and that for unvoiced speech. Section 6.4 presents a systematic evaluation of the overall system. Our preliminary results on unvoiced speech segregation have been published in three *ICASSP* papers (Hu and Wang, 2003; Hu and Wang, 2005; Wang and Hu, 2006).

6.1 Features for classification

Our task is to classify a T-F region, which is either a mask yielded by our iterative algorithm or a segment obtained in the segmentation stage, as speech dominant or interference dominant. Since the signal within such a region is mainly from one source, it should have similar acoustic-phonetic properties to those of clean speech if it is dominated by speech, and otherwise if it is dominated by interference. Therefore, we propose to classify a T-F region as either speech or interference using acoustic-phonetic features.

A basic speech sound is characterized by the following acoustic-phonetic properties: Short-term spectrum, formant transition, voicing, and phoneme duration (Stevens, 1998; Ladefoged, 2001). These features have been proven useful in recognizing speech, e.g., to distinguish different phonemes or words (Rabiner and Juang, 1993; Ali and Van der Spiegel, 2001a; Ali and Van der Spiegel, 2001b). These properties may also be useful in distinguishing speech from interference. However, it is important to treat these properties appropriately considering that we are dealing with speech corrupted by interference and our task here is to distinguish speech from interference. In particular, we give the following considerations.

- *Spectrum*. The short-term spectrum of an acoustic mixture at a particular time may be quite different from that of the target utterance or that of the interference in the mixture. Therefore, features representing the overall shape of a short-term spectrum may not be appropriate for this task. Such features include Mel-frequency cepstral coefficients (MFCC) (Rabiner and Juang, 1993) and linear predictive coding (LPC) (Rabiner and Juang, 1993), which are commonly used in ASR. Nevertheless, the short-term spectra in the T-F regions dominated by speech are expected to be similar to those of clean utterances, while the short-term spectra of other T-F regions tend to be different. Therefore, we use the short-term spectrum within a T-F region as a feature to decide whether this region is dominated by speech or interference. More specifically, we use the energy within individual T-F units as the feature to represent the short-term spectrum.

- *Formant transition*. It is difficult to estimate the formant frequency of a target utterance when there is strong interference. In addition, formant transition is embodied in the corresponding short-term spectrum. Therefore, we do not explicitly use formant transition in this study.
- *Voicing*. The pitch contours estimated by our system provide voicing information of a target utterance, which affords the opportunity to deal with voiced speech and unvoiced speech differently. In particular, one may build a single classifier that distinguishes any speech signal from non-speech signal. One may also build two classifiers, one for voiced speech and the other for unvoiced speech. The latter choice appears more advantageous since the type of interference that tends to confuse voiced speech is generally different from that which confuses unvoiced speech. In particular, the first type of interference is periodic or quasi-periodic and the second type aperiodic. This way, we may build a classifier that optimally distinguishes voiced speech from periodic or quasi-periodic interfering sounds instead of all interfering sounds. The same reasoning holds true for unvoiced speech.
- *Duration*. The duration of an interfering sound should be random, whereas for speech, each phoneme has a specific range of durations. However, as discussed in Chapter 4, we may not be able to detect the boundaries of phonemes that are strongly coarticulated. Therefore it is difficult to find the accurate durations of

individual phonemes from an acoustic mixture, and the durations of individual phonemes are not utilized in this study.

In summary, we use the signal energy within individual T-F units to derive the features for distinguishing speech from interference. Voiced speech and unvoiced speech are handled separately.

6.2 Voiced speech classification

Let H_0 be the hypothesis that a T-F region is dominated by speech and H_1 the hypothesis that it is dominated by interference. Let $X(c, m)$ be the energy in a T-F unit u_{cm} and $X(m) = \{X(c, m), \forall c\}$ the vector of the energy in all the T-F units at time frame m. $X(m)$ is referred to as the *cochleagram* at frame m (Wang and Brown, 2006). For an estimated pitch contour k, let $L_k(c, m)$ be the corresponding mask label of u_{cm}, i.e., $L_k(c, m)$ is 1 if u_{cm} is dominated by the source of pitch contour k and 0 otherwise (see Section 5.2 and Section 5.3), and S_k the corresponding segregated source. We use $X_k(m) = \{X_k(c, m), \forall c\}$ to represent the cochleagram of S_k at frame m, where,

$$X_k(c,m) = X(c,m)L_k(c,m) \tag{6.1}$$

Assuming that an estimated pitch contour k lasts from frame m_1 to frame m_2, let $\boldsymbol{X}_k$ be the cochleagram of S_k within these frames, i.e., $\boldsymbol{X}_k = \{X_k(m_1), X_k(m_1+1), \ldots, X_k(m_2)\}$. We label S_k as speech if:

$$P(H_0 \mid \boldsymbol{X}_k) > P(H_1 \mid \boldsymbol{X}_k) \tag{6.2}$$

Because estimated pitch contours have varied durations, directly estimating $P(H_0|\mathbf{X}_k)$ and $P(H_1|\mathbf{X}_k)$ for each possible duration is not computationally feasible. Therefore, we first consider a simple approximation that assumes each time frame being independent. That is,

$$P(H_i \mid \boldsymbol{X}_k) = \prod\nolimits_{m=m_1}^{m_2} P(H_i \mid X_k(m)), \;\; i = 0, 1 \tag{6.3}$$

and Equation (6.2) becomes

$$\prod\nolimits_{m=m_1}^{m_2} P(H_0 \mid X_k(m)) > \prod\nolimits_{m=m_1}^{m_2} P(H_1 \mid X_k(m)) \tag{6.4}$$

Since we do not know *a priori* $P(H_0|X_k(m))$ and $P(H_1|X_k(m))$, we need to estimate $P(H_0|X_k(m))$ from training data. With the estimated $P(H_0|X_k(m))$, $P(H_1|X_k(m))$ can then be calculated since $P(H_1|X_k(m)) = 1 - P(H_0|X_k(m))$. In this study, the training samples are obtained from the mixtures of all the utterances in the training part of the TIMIT database and 100 intrusions used in Section 5.1A. The iterative algorithm presented in Chapter 5 is used to estimate pitch contours and the associated masks for each mixture. We label all the obtained masks by comparing them with the corresponding ideal binary masks. For a particular mask, if more than half of its energy comes from target-dominant T-F units, we label it as target; otherwise, we label it as interference. As in Section 5.1A, we use these labeled samples to train an MLP that yields the desired label given a cochleagram. As discussed in Section 5.1A, the trained MLP yields a good estimate of $P(H_0|X_k(m))$. The number of units in the hidden layer of the MLP is determined to be 20 using cross-validation.

We have tested this classifier on the test corpus used in Section 5.3. In particular, for an acoustic mixture, we label each segregated voiced sound as either target or

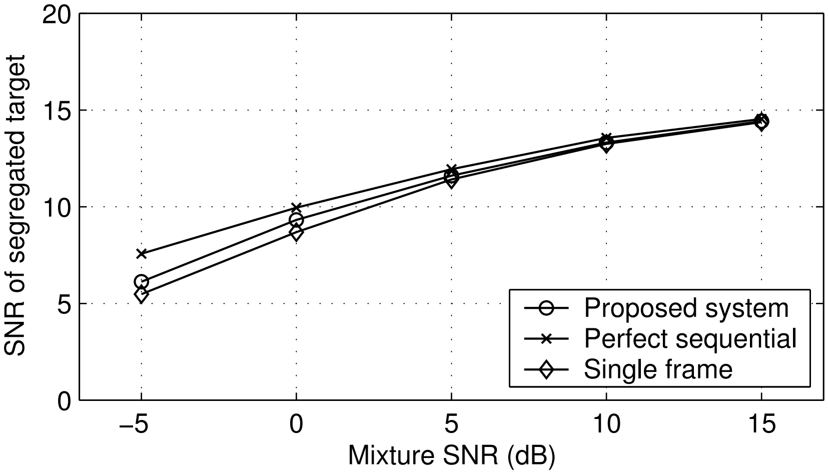

Figure 6.1. SNR of segregated voiced target after sequential grouping.

interference according to Equation (6.4). All the segregated sounds classified as speech are grouped to form the segregated target.

Figure 6.1 shows the average SNR of thus segregated targets (“Proposed system”). In this figure, each value is the average SNR of the segregated target over the 300 mixtures of individual targets (see Table 4.1) and intrusions N1-N15 (see Table 4.2); we do not use intrusions N16-N20 since they are speech utterances. In comparison, Figure 6.1 also shows the SNR of the segregated target with perfect sequential grouping, i.e., all the segregated voiced sounds that correspond to target are correctly grouped to form the segregated target. As shown in the figure, our system performs well when the mixture SNR is high. The average SNR of the segregated target obtained by our system is very close to that obtained with perfect sequential grouping when the mixture SNR is 5 dB or higher. The performance gap increases as the mixture SNR drops. Specifically, there is a

1.5 dB performance gap between our system and perfect sequential grouping when the mixture SNR is -5 dB. This gap mainly comes from classifying interference as target. Note that harmonic sounds within some intrusions, e.g. crowd noise and babble noise, are similar to speech signal.

In our previous study (Hu and Wang, 2005), we considered an alternative method of classification. In that study, we first modeled the cochleagram of clean speech at each frame using a GMM (Huang et al., 2001); the same is done for interference. We then used these models to compute $p(X_k(m)|H_0)$ and $p(X_k(m)|H_1)$ using the marginal distributions in the channels where $L_k(c, m) = 1$. A segment is classified as target if

$$\prod_{m=m_1}^{m_2} p(X_k(m) | H_0)P(H_0) > \prod_{m=m_1}^{m_2} p(X_k(m) | H_1)P(H_1) \qquad (6.5)$$

where $P(H_0)$ is the prior probability of a segment to be target dominant and $P(H_1)$ interference dominant. We found that the performance is not as good as MLP-based classification. The main reason, we believe, is that GMM is trained to represent the distributions of speech and interference accurately, whereas MLP is trained to distinguish speech and interference and therefore has more discriminative power, as we have discussed in Section 5.1A.

Instead of treating individual frames as independent, it is possible to consider the dependence between consecutive frames, as we did in our previous studies (Hu and Wang, 2005; Wang and Hu, 2006). More specifically, we have considered the case of the observation at a particular frame being dependent only on the pervious frame:

$$p(X_k(m) | H_i, X_k(1), X_k(2), \ldots, X_k(m-1)) = p(X_k(m) | H_i, X_k(m-1)), \quad i = 0, 1 \qquad (6.6)$$

Consequently, Equation (6.2) becomes

$$P(H_0 \mid X_k(m_1))\prod_{m=m_1+1}^{m_2}\frac{P(H_0 \mid X_k(m-1), X_k(m))}{P(H_0 \mid X_k(m-1))} >$$

$$P(H_1 \mid X_k(m_1))\prod_{m=m_1+1}^{m_2}\frac{P(H_1 \mid X_k(m-1), X_k(m))}{P(H_1 \mid X_k(m-1))} \tag{6.7}$$

Similar to the training used in the estimation of $P(H_0|X_k(m))$, we train an MLP to estimate $P(H_0|X_k(m-1), X_k(m))$ and use Equation (6.7) to distinguish speech and interference. The obtained result is similar to that obtained by assuming independence among frames. Hence for this task, it appears that considering dependence between consecutive frames does not help in performance, which may be due to the following reason. The advantage of considering the dependence between consecutive frames is the ability to incorporate the dynamics of signal across time. Since in the iterative algorithm described in the pervious chapter, masks are estimated with the constraint of temporal continuity (see Section 5.3), the signal within a mask tends to be stable for either speech or interference. Further considering the dynamics of a signal within a mask does not seem to provide more information for distinguishing speech and interference. In fact, the signals within many masks are so stable across time that by using only one frame for each mask, we can already distinguish speech and interference well. To illustrate this point, we classify each obtained mask by using the cochleagram at a single frame in the middle of the mask. The average SNR of such segregated target is shown in Figure 6.1, which is only slightly worse that those using information from all the frames.

6.3 Unvoiced speech classification

The task here is to identify the segments dominated by unvoiced speech among the segments obtained in the segmentation stage. An obtained segment may be dominated by voiced speech, unvoiced speech, or non-speech interference. Segments dominated by voiced speech have been identified in the previous process of voiced speech segregation (see Section 5.3). Our task here is to distinguish the segments dominated by unvoiced speech from those dominated by non-speech interference. This is performed in two steps: Segment reduction and segment categorization. In segment reduction, we utilize the segregated voiced speech to remove some segments that do not correspond to unvoiced speech. In segment categorization, we identify the segments dominated by unvoiced speech among the remaining ones and group them into the segregated speech.

A. Segment reduction

Since our task here is to group segments for unvoiced speech, segments mainly contain periodic or quasi-periodic signals unlikely originate from unvoiced speech. A segment is removed if more than half of its total energy is included in the T-F units dominated by a periodic signal. We consider a T-F unit u_{cm} dominated by a periodic signal if it is labeled 1 in the masks associated with an estimated pitch contour or has a high cross-channel correlation, i.e., $C(c, m) > 0.985$ or $C_E(c, m) > 0.985$.

Among the remaining segments, a segment dominated by unvoiced target is likely located in the unvoiced time frames of a target utterance, though it may contain some T-F units at the voiced time frames of the target speech since expanded obstruents often contain both voiced and unvoiced signal (see Section 2). This property is, however, not shared by interference-dominant segments that may have significant energy in the voiced frames of the target. Such segments are removed as follows.

We first label the voiced frames of a target utterance that unlikely contain an expanded obstruent, according to the segregated voiced target. Let $L_T(m)$ be the corresponding labels of T-F units of the segregated target at frame m and $X_T(m)$ the corresponding cochleagram, i.e. $X_T(m) = \{X_T(c, m), \forall c\}$ where $X_T(c, m) = X(c, m)L_T(c, m)$. Let $H_{0,a}$ be the hypothesis that the segment is dominated by an expanded obstruent and $H_{0,b}$ that the segment is dominated by any other phoneme. A voiced frame m is labeled as not dominated by an expanded obstruent if

$$P(H_{0,a} \mid X_T(m)) < P(H_{0,b} \mid X_T(m)) \tag{6.8}$$

Again, an MLP with one hidden layer of 20 units is trained to estimate $P(H_{0,a}|X_T(m))$ and $P(H_{0,b}|X_T(m))$.

A segment is removed if its energy in these labeled frames is greater than 50% of its total energy. As a result of this step, many segments dominated by interference are removed. We find that this step increases the robustness of the system and greatly reduces the computational burden for the following segment categorization.

B. Segment categorization

In this step, we classify a remaining segment as being dominated by either unvoiced speech or interference. Let s be a remaining segment lasting from frame m_1 to m_2, and $X_s(m) = \{X_s(c, m), \forall c\}$ be the corresponding cochleagram at frame m. That is,

$$X_s(c,m) = \begin{cases} X(c,m) & if\ u_{cm} \in s \\ 0 & else \end{cases} \tag{6.9}$$

Let $\boldsymbol{X}_s = [X_s(m_1), X_s(m_1+1), \ldots, X_s(m_2)]$. s is classified as dominated by unvoiced speech if:

$$P(H_{0,a} \mid \boldsymbol{X}_s) > P(H_1 \mid \boldsymbol{X}_s) \tag{6.10}$$

Recall that H_1 denotes the hypothesis that a T-F region is interference dominant. As discussed before, to simplify the computation of $P(H_{0,a}|\boldsymbol{X}_s)$ and $P(H_1|\boldsymbol{X}_s)$, we assume that individual frames are statistically independent. Therefore, Equation (6.10) becomes:

$$\prod_{m=m_1}^{m_2} P(H_{0,a} \mid X_s(m)) > \prod_{m=m_1}^{m_2} P(H_1 \mid X_s(m)) \tag{6.11}$$

The prior probabilities $P(H_{0,a})$ and $P(H_1)$ depend on the SNR of acoustic mixtures. Figure 6.2 shows the observed logarithmic ratios between $P(H_{0,a})$ and $P(H_1)$ from the training data at different mixture SNR levels. The relationship shown in the figure can be approximated with a linear function.

$$\log \frac{P(H_{0,a})}{P(H_1)} = 0.1166\,\mathrm{SNR} - 1.8962 \tag{6.12}$$

If we can estimate the mixture SNR by taking advantage of the segregated voiced target, we can estimate the log ratio of $P(H_{0,a})$ and $P(H_1)$ and use it in Equation (6.11). This essentially allows us to be more stringent in identifying a segment as speech when the mixture SNR is low, which is beneficial since at lower SNR levels intrusion is more disruptive.

We propose to estimate the SNR of an acoustic mixture using the voiced target segregated from the mixture, i.e., the target stream obtained after sequentially grouping

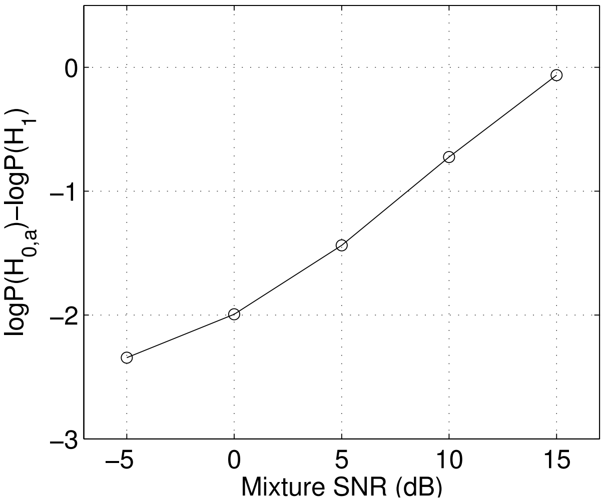

Figure 6.2. Ratio between the prior distribution of target and that of interference as a function of the mixture SNR.

segregated voiced sounds. Let E_1 be the total energy included in the T-F units labeled 1 at the voiced frames of the target. One may use E_1 to approximate the target energy at voiced frames and estimate the total target energy as αE_1. By analyzing the training part of the TIMIT database, we found that parameter α — the ratio between the total energy of a speech utterance and the total energy at the voiced frames of the utterance — varies substantially across individual utterances. In this study, we set α to 1.09, the average value of all the utterances in the training part of the TIMIT database. Let E_2 be the total energy included in the T-F units labeled 0 at the voiced frames of the target, N_1 the total number of these voiced frames, and N_2 the total number of other frames. We use E_2/N_1 to approximate the interference energy per frame and estimate the total interference energy as $E_2(N_1+N_2)/N_1$. Consequently, the estimated mixture SNR is:

$$SNR = 10\log_{10}\frac{\alpha N_1 E_1}{(N_1+N_2)E_2} = 10\log_{10}\frac{E_1}{E_2} + 10\log_{10}\alpha + 10\log_{10}\frac{N_1}{N_1+N_2} \quad (6.13)$$

Given $\alpha = 1.09$, $10\log_{10}\alpha = 0.37$ dB. We have applied this SNR estimation to the test corpus. Figure 6.3(a) shows the mean and the standard deviation of the estimation error at each SNR level of the original mixtures; the estimation error equals to the estimated SNR subtracted by the true SNR. As shown in the figure, the system yields a reasonably good estimate when the mixture SNR is lower than 10 dB. When the mixture SNR is higher than 10 dB, Equation (6.13) tends to yield a value that is lower than the true SNR. As we discussed in Section 2.2, some voiced frames of the target, such as those corresponding to expanded obstruents, may contain some unvoiced target energy, which is not included in E_1 but in E_2. When the mixture SNR is low, this part of unvoiced target energy is much lower than the voiced target energy and the interference energy. Therefore, it is negligible and Equation (6.13) provides a good estimate of the mixture SNR. When the mixture SNR is high, this unvoiced target energy is comparable to interference energy and as a result the estimated SNR is systematically lower than the true SNR.

Alternatively, since our task here is to group segments for unvoiced speech, we have considered estimating the mixture SNR at the unvoiced frames of the target and then using this estimated SNR to determine the corresponding $P(H_{0,a})$ and $P(H_1)$ values in a manner similar to Equation (6.12). In particular, we approximate the target energy at unvoiced frames as $(\alpha-1)E_1$ and the interference energy at these frames as $E_2 \cdot N_2/N_1$. The estimated SNR at unvoiced frames is then:

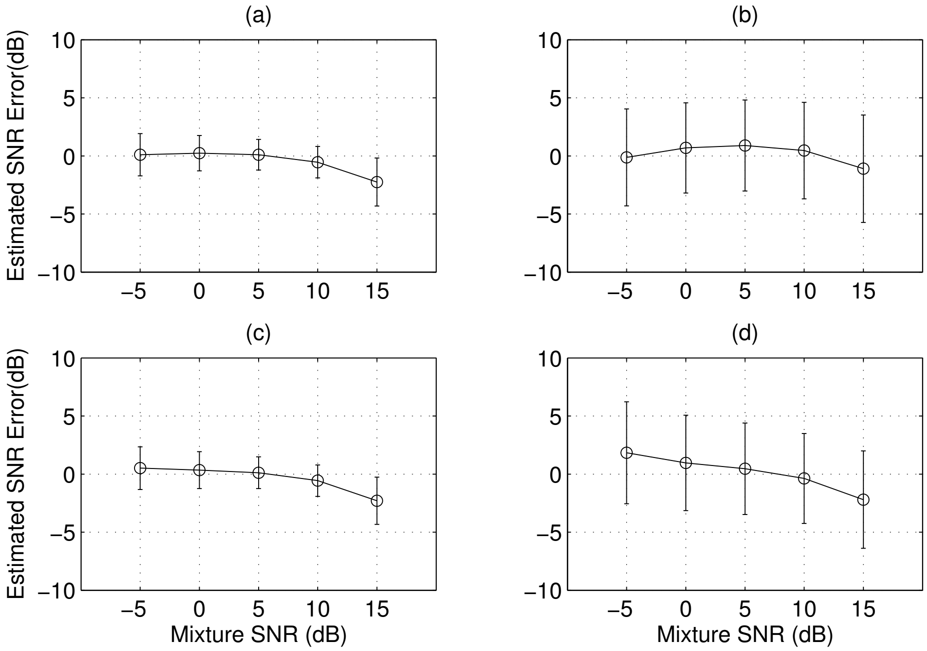

Figure 6.3. Mean and standard deviation of the estimated SNRs of the mixtures in the test corpus. (a) Estimate of the overall mixture SNR using Equation (6.13). (b) Estimate of the mixture SNR at the unvoiced frames of the target using Equation (6.14). (c) Estimate of the overall mixture SNR using Equation (6.15). (d) Estimate of the mixture SNR at the unvoiced frames of the target using Equation (6.16).

$$SNR = 10\log_{10}\frac{(\alpha-1)N_1E_1}{N_2E_2} = 10\log_{10}\frac{E_1}{E_2} + 10\log_{10}(\alpha-1) + 10\log_{10}\frac{N_1}{N_2} \qquad (6.14)$$

Given α= 1.09, 10log$_{10}$(α−1) = -10.45 dB. We have applied this SNR estimate to the test corpus. Figure 6.3(b) shows the mean and the standard deviation of the estimation error at each SNR level of the original mixtures; here the estimation error equals to the estimated SNR at the unvoiced frames of the target subtracted by the true SNR at unvoiced frames. As shown in the figure, this estimate at unvoiced frames, not surprisingly, is not as good as the estimate of the overall mixture SNR shown in Figure 6.3(a). In particular, the standard deviation is much larger. As we have previously discussed, α varies substantially across individual utterances. Since α is close to 1, a small change of the α value causes a more significant change of $10\log_{10}(\alpha-1)$ than that of $10\log_{10}(\alpha)$. As a result, there is much higher variation in the error estimation from Equation (6.14) than Equation (6.13).

The variation of the parameter α is partially from the variation of the relative durations of unvoiced speech across different utterances. To remove the influence of this duration variation in our SNR estimation, we have considered estimating the target energy at the unvoiced frames based on the frame-level ratio of the average energy of speech utterances at unvoiced frames and that at voiced frames. Let β be the ratio. We approximate the total target energy at the unvoiced frames as $\beta E_1 \cdot N_2/N_1$. Therefore, we estimate the overall mixture SNR as:

$$SNR = 10\log_{10} \frac{E_1 + \dfrac{N_2}{N_1}\beta E_1}{\dfrac{(N_1+N_2)}{N_1}E_2}$$

$$= 10\log_{10} \frac{(N_1+\beta N_2)E_1}{(N_1+N_2)E_2} = 10\log_{10}\frac{E_1}{E_2} + 10\log_{10}\frac{N_1+\beta N_2}{N_1+N_2}, \qquad (6.15)$$

and the SNR at the unvoiced frames of the target as:

$$SNR = 10\log_{10}\frac{\beta E_1}{E_2} = 10\log_{10}\frac{E_1}{E_2} + 10\log_{10}\beta \qquad (6.16)$$

Here we set β to 0.12, estimated from the training part of the TIMIT database. We have applied these two SNR estimates to the test corpus. Figures 6.3(c) and 6.3(d) show the means and the standard deviations of the corresponding estimation errors, respectively. From Figures 6.3(a) and 6.3(c), we can see that Equations (6.13) and (6.15) yield similar estimates of the overall mixture SNR on the test corpus. In particular, the standard deviations of both estimates are almost the same. On the other hand, as shown in Figures 6.3(b) and 6.3(d), the SNR at the unvoiced frames estimated using Equation (6.14) is notably better than that using Equation (6.16). However, the standard deviations of both estimates are still quite similar. This suggests that the variation of all the above SNR estimates is caused mainly by the energy variation of phonemes in speech utterances.

We have applied the above estimates to obtain the corresponding $P(H_{0,a})$ and $P(H_1)$ values. These values are combined with the output of the obtained MLP to yield $P(H_{0,a}|X_s(m))$ at the estimated SNR level. Then we label a segment as either speech or interference according to Equation (6.11). All the segments labeled as speech are added to the segregated voiced stream, which yields the final segregated stream. In our test, the above SNR estimates yield similar performance. The result reported here is using the SNR estimate of Equation (6.13).

This method for segregating unvoiced speech is very similar to a previous version (Wang and Hu, 2006), except that here we use the estimated mixture SNR to determine an SNR-dependent prior probabilities while in the previous system we used fixed prior probabilities at all the SNR levels. We find that using SNR-dependent prior probabilities gives better performance, especially when the mixture SNR is high. In another preliminary study (Hu and Wang, 2005), we used GMM to model both speech and interference and then classified a segment using the obtained models. The performance in that study is not as good as the present method for the reasons given in Section 6.2. In addition, we note that considering the dependence between consecutive frames yields similar performance, again for the same reasons as given in Section 6.2.

We have also considered using SNR-dependent prior probabilities in the sequential grouping of segregated voiced sounds (see Section 6.2), but did not obtain a significant performance gain. This is presumably because our classifier already works well without knowing SNR levels, especially when the mixture SNR is high.

6.4 Overall Evaluation

We now evaluate the overall performance of the complete system on the test corpus described in Section 4.6. This test corpus contains 20 utterances randomly selected from the test part of the TIMIT database listed in Table 4.1 and 20 intrusions listed in Table 4.2. These intrusions have a considerable variety. Some of them are noise-like, such as the wind (N9) and the cocktail party noise (N11), and some contain strong harmonic sounds, such as the siren (N3) and the electric fan (N5). They form a good corpus for testing the capacity of a CASA system in dealing with various types of interference. In this evaluation, we only use intrusions N1-N15 since intrusions N16-N20 are speech utterances.

We evaluate our system by comparing the segregated target with the ideal binary mask — the stated computational goal. As in Section 5.4, two error measures are used here:

percentage of energy loss P_{EL} and percentage of noise residue P_{NR} (Hu and Wang, 2004a).

The P_{EL} and P_{NR} values at different input SNR levels are shown in Figures 6.4(a) and 6.4(b). Each value in the figure is the average over the 300 mixtures of individual targets and intrusions N1-N15. As shown in the figure, for the final segregation, our system captures an average of 76.4% of target energy at -5 dB SNR. This value increases to 97.6% when the mixture SNR increases to 15 dB. On average 22.9% of the segregated target belongs to interference at -5 dB. This value decreases to 0.7% when the mixture SNR increases to 15 dB. In summary, our system captures a majority of target without including much interference.

To see the performance of our system on voiced speech and unvoiced speech separately, we measure P_{EL} for the target in the voiced frames and that in the unvoiced frames. The average of these P_{EL} values at different SNR levels are shown in Figures 6.4(c) and 6.4(d). Note that since some voiced frames contain unvoiced target, these are not exactly the P_{EL} values of voiced speech and unvoiced speech. Nevertheless, they are close to the real values. As shown in the figure, our system performs very well on voiced speech. In particular, our system captures 79.0% of the target energy at the voiced frames when the mixture SNR is -5 dB and 99.1% when the mixture SNR is 15 dB. As expected, the system does not perform nearly as well for unvoiced speech. It captures 30.3% of the target energy at the unvoiced frames when the mixture SNR is -5 dB and 78.6% when the

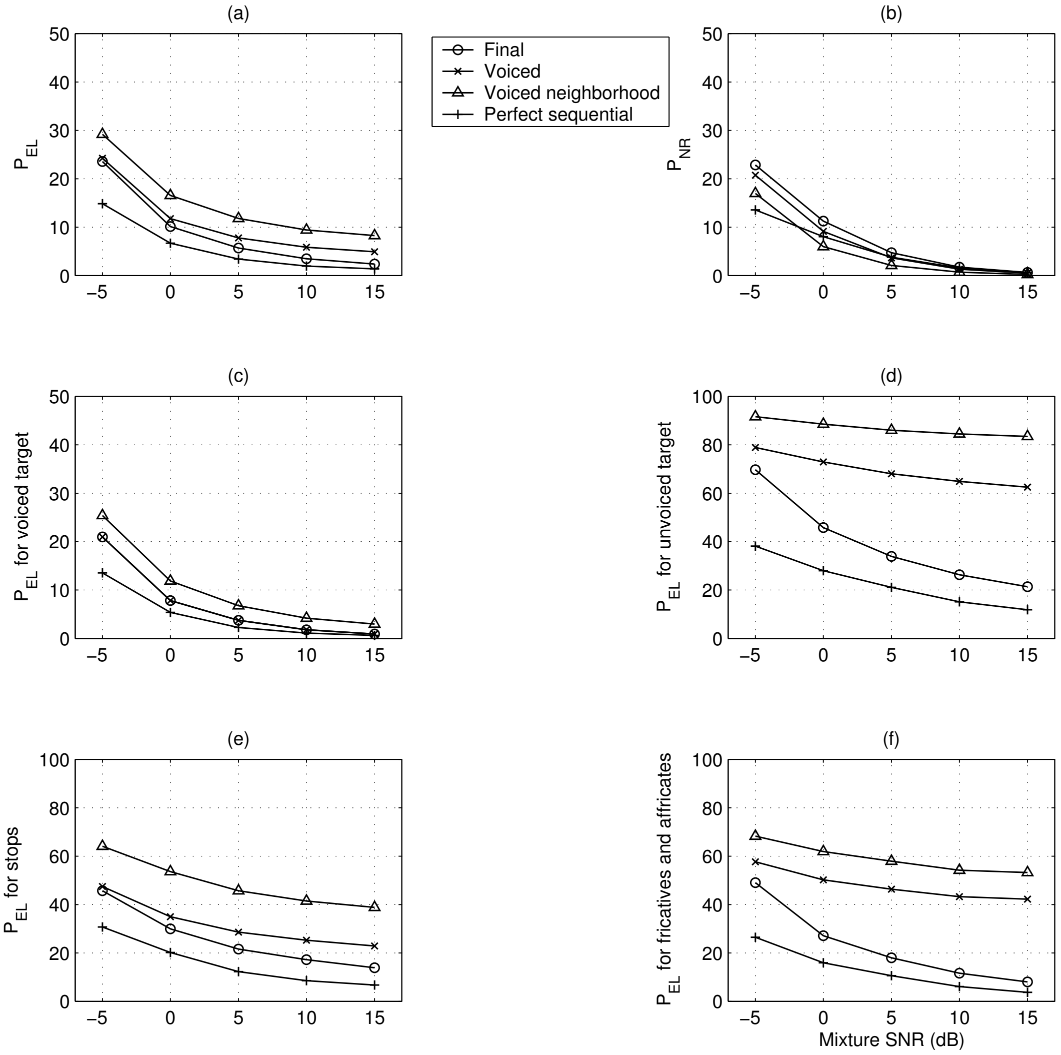

Figure 6.4. System performance. In this figure, “Final” refers to the final segregated target, “Voiced” refers to the segregated voiced target, “Voice neighborhood” refers to the segregated voiced target without using estimated T-segments, and “Perfect sequential” refers to segregated target with perfect sequential grouping of both voiced and unvoiced speech. (a) Average percentage of energy loss. (b) Average percentage of noise residue. (c) Average percentage of energy loss for voiced speech. (d) Average percentage of energy loss for unvoiced speech. (e) Average percentage of energy loss for stop consonants. (f) Average percentage of energy loss for fricatives and affricates.

mixture SNR is 15 dB. Overall, our system is able to capture more than 50% of target energy at the unvoiced frames when the mixture SNR is greater than or equal to 0 dB.

As discussed in Section 2.2, expanded obstruents often contain voiced and unvoiced signals at the same time. Therefore, we measure P_{EL} for these phonemes separately in order to gain more insight into system performance. Because affricates do not occur very often and they are similar to fricatives, we measure the P_{EL} for fricatives and affricates together. The average of these P_{EL} values at different SNR levels are shown in Figures 6.4(e) and 6.4(f). As shown in the figure, our system performs somewhat better for stops when the mixture SNR is lower than 0 dB and somewhat better for fricatives and affricates when the mixture SNR is higher than 5 dB. On average, the system captures about 50% of these phonemes when the mixture SNR is -5 dB and about 90% of them when the mixture SNR is 15 dB.

To see the advantages of segmentation, Figure 6.4 also show the P_{EL} and P_{NR} values for the segregated voiced target obtained by sequentially grouping masks generated by our system with the iterative algorithm described in Chapter 5. As discussed in Chapter 5, we consider two methods in labeling T-F units, one using information from a neighborhood of T-F units ("Voiced neighborhood") and the other further using estimated T-segments ("Voiced"). Figure 6.4 shows the results for these methods. As shown in the figure, the second method captures about 3.2% more of target energy at the voiced time frames and 17.4% more at the unvoiced frames on average. This additional 17.4% of target energy mainly corresponds to unvoiced phonemes that have strong coarticulation with neighboring voiced phonemes. By comparing these P_{EL} and P_{NR}

values with those of the final segregated target, we can see that grouping segments dominated by unvoiced speech helps to recover a significant amount of unvoiced target. It also includes a small amount of additional interference energy, especially when the mixture SNR is low. Overall, the performance clearly demonstrates that segmentation is important for segregating unvoiced speech.

In addition, Figure 6.4 shows the P_{EL} and P_{NR} values for the segregated target obtained from perfect sequential grouping of masks for voiced speech and segments for unvoiced speech. As shown in the figure, there is a performance gap that can be narrowed with better sequential grouping, especially when the mixture SNR is low.

We also measure the system performance in terms of SNR by treating the target resynthesized from the corresponding ideal binary mask as signal (see Section 5.4B). Figures 6.5(a) and 6.5(b) show the overall average SNR values at different levels of mixture SNR and the corresponding SNR gain. Our system improves SNR in all input conditions. In particular, the average SNR gain is 10.7 dB when the mixture SNR is -5 dB and 2.8 dB when the mixture SNR is 15 dB.

As in the evaluation of our previous system (Hu and Wang, 2004a), to see the ability of our system in dealing with various types of intrusions, we show in Figure 6.6 the average SNR improvement obtained by our system for each intrusion when the mixture SNR is 0 dB. These intrusions are listed in Table 4.2. As shown in the figure, our system obtains significant SNR improvement for all the intrusions. The best performance is obtained for three types of intrusions: N1 (white noise), N6 (clock alarm), and N12 (crowd noise from a playground). Our system does not perform as well on three

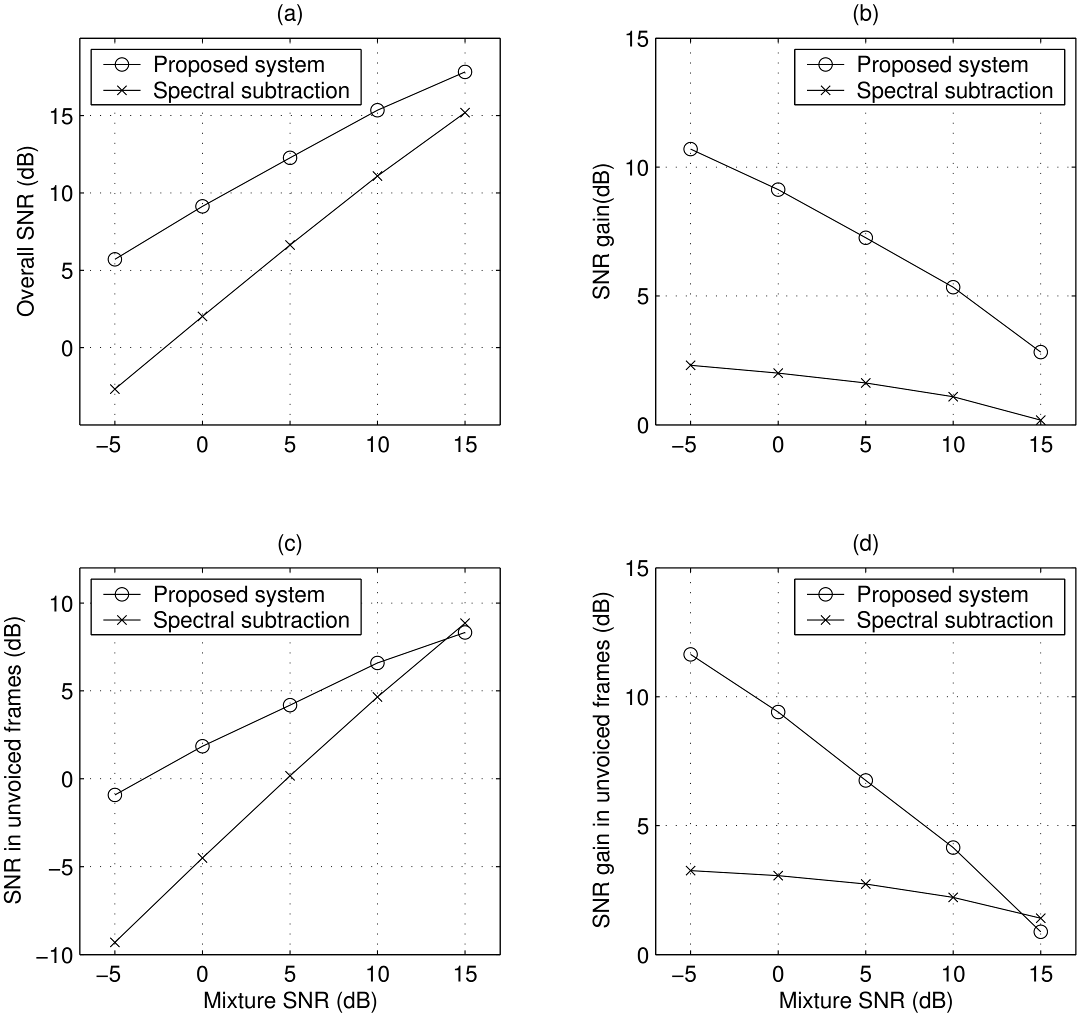

Figure 6.5. (a) SNR of segregated target. (b) SNR gain of segregated target. (c) SNR of segregated target at unvoiced frames. (d) SNR gain of segregated target at unvoiced frames.

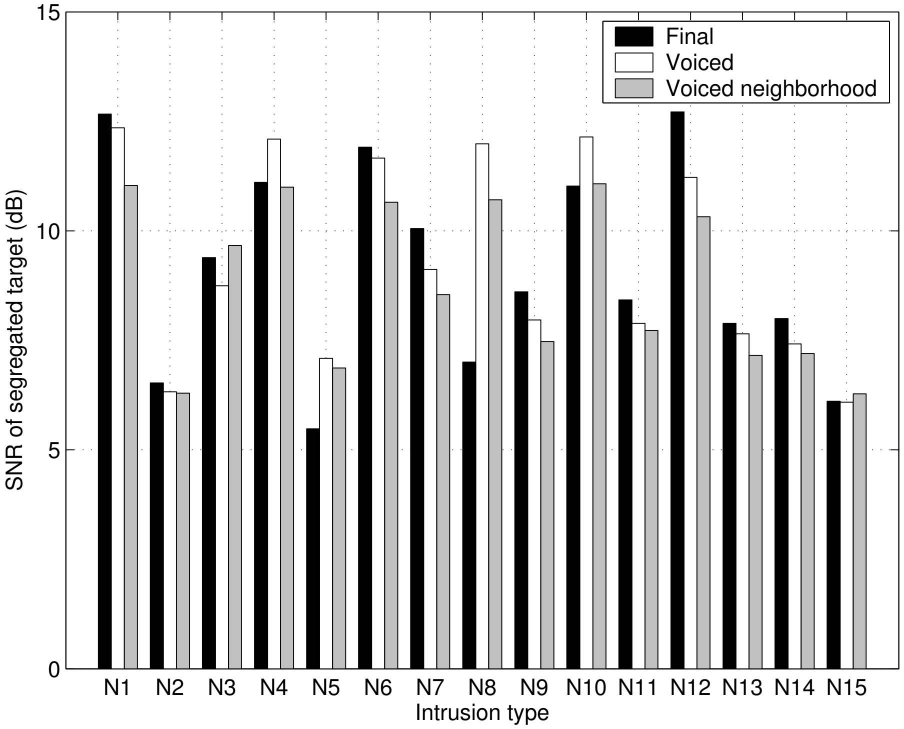

Figure 6.6. SNR of segregated target for each intrusion. The input SNR is 0 dB. In this figure, "Final" refers to the final segregated target, "Voiced" refers to the segregated voiced target, and "Voice neighborhood" refers to the segregated voiced target without using estimated T-segments.

intrusions, N2 (rock music), N5 (electrical fan), and N15 (babble noise), as others. The errors for these three intrusions are mainly caused by erroneous sequential grouping, in which our system misclassifies some sounds from these intrusions as target signal. In addition, Figure 6.6 also shows the SNR improvement of the voiced targets segregated with and without obtained T-segments, referred to as "Voiced" and "Voiced neighborhood", respectively. In most conditions, using obtained T-segments helps to improve the performance, except for intrusions N3 (siren) and N15 (babble noise), where the performance degradation is caused by the mismatch error in segmentation. Further sequential grouping of unvoiced segments improves SNR for most intrusions, but not for intrusions N4 (telephone), N5 (electric fan), N8 (bird chirp and water flowing), and N10 (rain), especially for N8. The performance degradation is caused by labeling interference-dominant segments as target during the sequential grouping of unvoiced segments. On average, we obtain a 9.1 dB gain of SNR for the proposed system ("Final"), a 9.3 dB gain for the voiced targets segregated using obtained T-segments ("Voiced"), and an 8.8 dB gain for those without using obtained T-segments ("Voiced neighborhood"). The final version performs the best overall if the results for N8 are excluded.

To put our performance in perspective, we have compared it with spectral subtraction, a standard method for speech enhancement (Huang et al., 2001). The method is applied as follows. For each acoustic mixture, we assume that the time positions of the silent portions of a target utterance are known and use the short-term spectra of interference in these portions as the estimates of interference. Interference is attenuated by subtracting the most recent interference estimate from the mixture spectrum at every time frame. Figure 6.5(a) shows the SNR obtained by spectral subtraction, and Figure 6.5(b) the SNR at unvoiced frames of the target utterance. As shown in the figure, both methods obtain SNR improvement in all the SNR levels. Our system performs substantially better than the spectral subtraction method for both voiced and unvoiced speech except for unvoiced speech at the input SNR of 15 dB. The improvement is more pronounced when the mixture SNR is low.

As an example, Figures 6.7(e) and 6.7(f) show the mask and the waveform of the target segregated from the mixture M1 shown in Figure 1.2(d). Compared with the ideal mask in Figure 6.7(g) and the corresponding resynthesized waveform in Figure 6.7(h), our system segregates most target energy and rejects most interfering energy. The SNR of this segregated target is 15.1 dB. In addition, Figures 6.7(a) and 6.7(b) show the mask and the waveform of the voiced target segregated by labeling individual T-F units using information from a neighborhood of T-F units. Figures 6.7(c) and 6.7(d) show the mask and the waveform of the voiced target segregated using T-segments. The target utterance, "That noise problem grows more annoying each day," includes 5 stops (/t/ in "that", /p/ and /b/ in "problem", /g/ in "grows", and /d/ in "day"), 3 fricatives (/ð/ in "that", /z/ in "noise", and /z/ in "grows"), and 1 affricate (/tʃ/ in "each"). The unvoiced part of some consonants that have strong coarticulation with the voiced speech, such as /ð/ and /t/ in "that" and /d/ in "day", are segregated by using T-segments. The unvoiced part of /z/ in "noise" and /tʃ/ in "each" are segregated by grouping the corresponding segments. Except for a significant loss of energy of /p/ in "problem" and some energy loss of /g/ in "grows", our system segregates most energy of the above consonants.

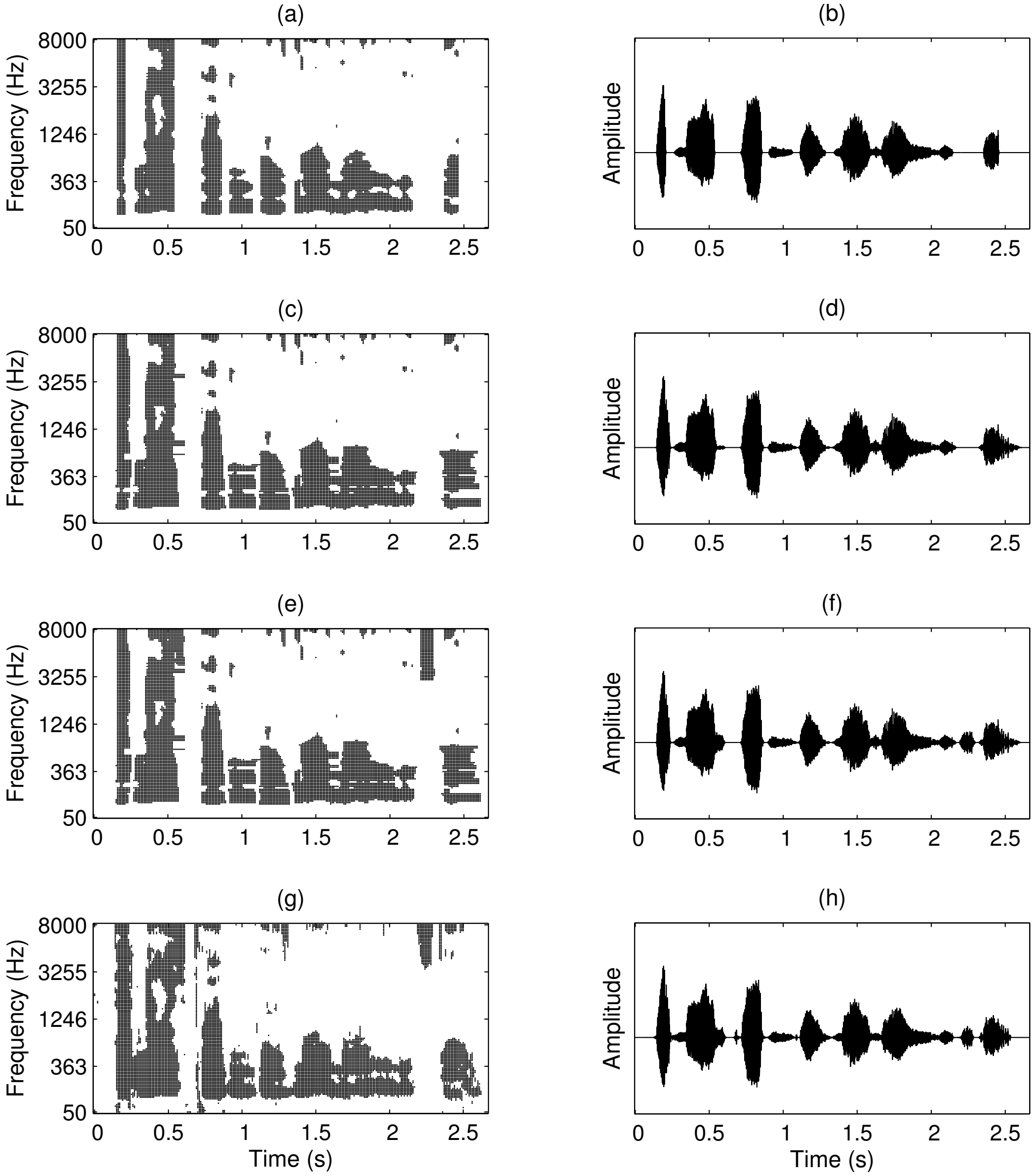

Figure 6.7. Segregated target of the mixture M1 shown in Figure 1.2(d). (a) Mask of the voiced target segregated by labeling individual T-F units using information from a neighborhood of T-F units. (b) Waveform resynthesized from the mask in (a). (c) Mask of the voiced target segregated using estimated T-segments. (d) Waveform resynthesized from the mask in (c). (e) Mask of segregated target, including both voiced and unvoiced portions of target. (f) Waveform resynthesized from to the mask in (e). (g) Ideal binary mask of target. (h) Waveform resynthesized from the ideal binary mask.

CHAPTER 7

CONTRIBUTIONS AND FUTURE WORK

7.1 Summary

In this dissertation, we have proposed a CASA system for monaural speech segregation. The proposed system segregates a target utterance from an acoustic mixture in four stages: Peripheral analysis, feature extraction, segmentation, and grouping. In the first stage, the system decomposes the acoustic mixture into T-F units with bandpass filtering and subsequent time windowing. Each T-F unit corresponds to a small T-F area within a filter channel and a time frame. In the second stage, our system extracts auditory features corresponding to ASA cues, including harmonicity, AM rates, onset and offset. In the third stage, our system segments the input mixture via a multiscale analysis of onset and offset. In particular, segments for both voiced and unvoiced speech are generated by our system. In the last stage, our system groups T-F units dominated by the target utterance. It first estimates pitch contours of the target utterance and segregates sounds corresponding to the estimated pitch contours in an iterative manner. Our system then sequentially groups the segregated voiced sounds from a target utterance. In this study, we consider the situation when interference is non-speech signal and sequentially group voiced speech using a feature-based classifier that distinguishes speech signal from non-speech signal. Similarly, our system segregates unvoiced speech by classifying the segments obtained in the third stage as either target or interference and grouping speech segments with the segregated voiced target.

Our proposed system has been systematically and extensively evaluated with mixtures of speech utterances and various types of intrusions. The evaluation shows that our system captures most energy of a target utterance without including much interference. The performance of our system is substantially better than previous CASA systems and the spectral subtraction method for speech enhancement. Our research has advanced the state-of-the-art in speech segregation by a considerable margin.

7.2 Contributions

Our study on monaural speech segregation makes several novel contributions.

First, we have proposed segregation of voiced speech in the high-frequency range based on the AM of unresolved harmonics. In addition, we use supervised learning to optimally combine the periodicity cue and the AM cue to segregate voiced speech. As a result, our system is able to segregate a majority of voiced speech in the high-frequency range. Recall that voiced speech segregation in the high-frequency range has been a serious problem for previous CASA systems.

Second, we have proposed an algorithm that estimates target pitch and segregates voiced target in tandem. This algorithm iteratively improves the estimation of both target pitch and voiced target. Our algorithm is robust to interference and produces a good estimate of both the pitch and the voiced utterance even in the presence of strong interfering sound.

Third, we have proposed a method of segmenting an acoustic scene via a multiscale analysis of onset and offset of auditory events. This analysis provides a general framework for auditory segmentation since onsets and offsets are common cues to all sounds – voiced speech, unvoiced speech, and non-speech sounds. Consequently, our system is able to segment both voiced and unvoiced speech. Our study shows that event onsets and offsets can play a fundamental role in sound organization (Turgeon et al., 2002). Although it is well known that onset and offset are important ASA cues, few computational studies have previously explored their usage. Brown and Cooke incorporated common onset and common offset as grouping cues in their CASA system but did not find any significant performance improvement (Brown and Cooke, 1994).

Fourth, we have proposed segregating unvoiced speech by onset/offset-based segmentation and subsequent feature-based classification. This is the first systematic study of unvoiced speech segregation. Our system segregates unvoiced speech well. In particular, the system captures more than 50% of unvoiced speech even when the mixture SNR is at 0 dB.

7.3 Insights gained

There are several important insights gained during this dissertation study. Our first observation is that the temporal properties of acoustic signals are very useful for segregation. Our system includes extensive usage of temporal properties. In particular, the system groups target sounds in consecutive frames based on the temporal continuity of speech signal. It uses temporal continuity to improve the estimation of target pitch. Furthermore, our system generates segments based on analyzing sound intensity across time, i.e., onset and offset detection. The importance of temporal properties of speech for human speech recognition has been demonstrated by Shannon et al. (1995; 1998). In addition, several studies in ASR suggest that long-term temporal information helps to improve the recognition rate (Hermansky and Sharma, 1999; Sharma et al., 2000). All these observations show that temporal information plays a critical role in sound organization and recognition.

We also find that it may be advantageous to segregate voiced speech first and then use the segregated voiced speech to aid the segregation of unvoiced speech. As discussed before, unvoiced speech is more vulnerable to interference and more difficult to segregate. Segregation of voiced speech is more reliable and can be used to improve the segregation of unvoiced speech. Our system shows that the unvoiced speech having strong coarticulation with voiced speech can be readily segregated based on segregated voiced speech and estimated T-segments. Segregated voiced speech is also used to specify the possible T-F locations of unvoiced speech. As a result, our system need not search the entire T-F region for segments dominated by unvoiced speech and is less likely to identify an interference-dominant segment as target. In addition, we have proposed a method of using an estimate of the mixture SNR using segregated voiced speech that helps the system to adapt the prior probabilities in identifying segments of unvoiced speech.

In addition, segment formation is important for unvoiced speech segregation. In our system, the segmentation stage provides T-segments that help to segregate unvoiced speech having strong coarticulation with voiced speech. As shown by Cole et al. (Cole et al., 1996), such portions of speech are important for speech intelligibility. Furthermore, most unvoiced speech is segregated by grouping the estimated segments that are dominated by unvoiced speech.

Finally, we have introduced the notion of a time-segment (T-segment) in a filter channel. A T-segment defines a period of time within a narrow passband which contains signal mainly from one source. Our study shows that one can accurately determine T-segments based on onset and offset analysis. In addition, our study includes two methods of grouping T-segments across frequency channels. First, T-segments are merged across frequency based on common onset and offset in the segmentation stage. This method works for both voiced and unvoiced speech. Second, in grouping voiced speech, our system integrates T-segments across frequency based on the estimated pitch. We find that the second method works better for voiced speech than the first one, which suggests that periodicity is a more reliable cue for organizing signal across frequency.

7.4 Remaining challenges and future work

We should point out that our approach is primarily feature-based. Features such as periodicity, AM, and onset, are general properties of all types of sounds. Our system does not employ specific prior knowledge of target or interference, except in sequential grouping, where we have utilized the acoustic and phonetic properties of speech and interference. Prior knowledge helps human ASA in the form of schema-based, or model-based, grouping (Bregman, 1990). As discussed before, model-based sound organization has been studied (Ellis, 1996; Roweis, 2001; Roweis, 2003). In particular, Barker et al. coupled segmentation with explicit speech models (Barker et al., 2005). Srinivasan and Wang used word models to restore phonemes that are masked by interference (Srinivasan and Wang, 2005). In general, feature-based approaches have broader applicability, whereas model-based approaches work better when the models fit the real situation. An important issue for future research in CASA is how to integrate feature-based and model-based approaches.

A natural speech utterance contains silent gaps and other sections masked by interference. In practice, one needs to group the utterance across such time intervals. This is the problem of sequential grouping. In this study, we handle this problem in a limited way by applying feature-based classification assuming non-speech interference. Systematic evaluation shows that although our system yields good performance, it is still far from perfect with regard to sequential grouping. The assumption of non-speech interference is obviously not applicable to mixtures of multiple speakers. As discussed in Section 2.3, sequentially grouping segments or masks may be achieved by using speech recognition in a top-down manner (Barker et al., 2005) or by speaker recognition using trained speaker models (Shao and Wang, 2005). Nevertheless, these studies on sequential grouping are still not mature. Substantial effort is needed to develop a general approach for sequential grouping.

In this study, we use the term *target* to refer to the target utterance we aim to segregate. In practice, which sound source should be considered as target is task-dependent. Even when the target is known in advance, it is not a trivial problem for a CASA system to decide which segregated sounds belong to the target. This problem is closely related to the problem of sequential grouping and in many situations one can deal with them at the same time. For example, if our goal is to segregate utterances from a particular speaker, we could solve both problems by recognizing the speaker of a segregated stream and grouping streams accordingly. In other situations, one may approach these two problems differently.

Room reverberation is another important issue that must be addressed before speech segregation systems can be deployed in real world environments. Reverberation raises several challenges for CASA. First, it smears temporal information, such as onset and particularly offset, within a sound. Second, it corrupts periodic information of a harmonic source. Since both temporal and periodicity cues are important for our system, we expect segregation performance to drop in a reverberant condition. A few previous studies have addressed the problem of speech segregation with room reverberation. Kingsbury et al. suggested using the modulation spectrogram to remove reverberation and extract robust features for speech recognition (Kingsbury et al., 1998). Palomaki et al. proposed to separate reverberant speech from interference based on an onset analysis of target speech using the modulation spectrogram (Palomaki et al., 2004). Roman and Wang recently conducted a study on pitch-based segregation of reverberant speech (Roman and Wang, 2005). Despite these studies, room reverberation remains a considerable challenge for future research.

Finally, we remark that CASA is a highly promising approach to the speech segregation problem. Although there is still a significant gap between the performance of speech segregation systems and that of a human listener with normal hearing, we believe that further advances along the CASA path will eventually bridge this gap.

BIBLIOGRAPHY

M. Abe and S. Ando, "Auditory scene analysis based on time-frequency integration of shared FM and AM," in *Proc. ICASSP*, Vol. 4, pp. 2421-2424, 1998.

A.M.A. Ali and J. Van der Spiegel, "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE Trans. Speech and Audio Process.*, Vol. 9, pp. 833-841, 2001a.

A.M.A. Ali and J. Van der Spiegel, "Acoustic-phonetic features for the automatic classification of fricatives," *J. Acoust. Soc. Am.*, Vol. 109, pp. 2217-2235, 2001b.

P.F. Assmann and Q. Summerfield, "Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, Vol. 88, pp. 680-697, 1990.

P.C. Bagshaw, S. Hiller, and M.A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," in *Proc. Eurospeech*, pp. 1003-1006, 1993.

J.P. Barker, M.P. Cooke, and D.P.W. Ellis, "Decoding speech in the presence of other sources," *Speech Comm.*, Vol. 45, pp. 5-25, 2005.

A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, Vol. 7, pp. 1129-1159, 1995.

J. Benesty, S. Makino, and J. Chen, Ed., *Speech enhancement*, New York: Springer, 2005.

J. Bird and C.J. Darwin, "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing,* A.R. Palmer, A. Rees, A.Q. Summerfield, and R. Meddis, Ed., London, UK: Whurr, pp. 263-269, 1998.

B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals," *Proc. IEEE*, Vol. 80, pp. 520-538, 1992a.

B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications," *Proc. IEEE*, Vol. 80, pp. 540-568, 1992b.

P. Boersma and D. Weenink, *Praat: Doing phonetics by computer*, Version 4.2.31, http://www.fon.hum.uva.nl/praat/, 2004.

M. Brandstein and D. Ward, Ed., *Microphone arrays*, New York: Springer, 2001.

A.S. Bregman, *Auditory scene analysis*, Cambridge, MA: MIT Press, 1990.

J. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing: algorithm, architectures, and applications,* F. Fogelman-Soulie and J. Herault, Ed., New York: Springer, pp. 227-236, 1989.

G.J. Brown and M.P. Cooke, "Computational auditory scene analysis," *Comput. Speech and Language*, Vol. 8, pp. 297-336, 1994.

G.J. Brown and D.L. Wang, "Modelling the perceptual segregation of double vowels with a network of neural oscillators," *Neural Networks*, Vol. 10, pp. 1547-1558, 1997.

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, Ed., New York, NY: Springer, pp. 371-402, 2005.

D.S. Brungart, P.S. Chang, B.D. Simpson, and D.L. Wang, "Isolating the energetic component of speech-on-speech masking with an ideal binary mask," *Submitted for journal publication*, 2005.

J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 8, pp. 679-698, 1986.

R.P. Carlyon and T.M. Shackleton, "Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms?" *J. Acoust. Soc. Am.*, Vol. 95, pp. 3541-3554, 1994.

P.S. Chang, *Exploration of behavioral, physiological, and computational approaches to auditory scene analysis*, M.S. Thesis, The Ohio State University Dept. Comput. Sci. & Eng., 2004 (available at http://www.cse.ohio-state.edu/pnl/theses).

R.A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," in *Proc. ICASSP*, Vol. 2, pp. 853-856, 1996.

M.P. Cooke, *Modelling auditory processing and organisation*, Cambridge, UK: Cambridge University Press, 1993.

M.P. Cooke, "Glimpsing speech," *J. Phonetics*, Vol. 31, pp. 579-584, 2003.

M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, Vol. 34, pp. 267-285, 2001.

C.J. Darwin, "Perceiving vowels in the presence of another sound: Constraints on formant perception," *J. Acoust. Soc. Am.*, Vol. 76, pp. 1636-1647, 1984.

C.J. Darwin, "Auditory grouping," *Trends in Cognitive Science*, Vol. 1, pp. 327-333, 1997.

A. de Cheveigné, "Concurrent vowel identification III: A neural model of harmonic interference cancellation," *J. Acoust. Soc. Am.*, Vol. 101, pp. 2857-2865, 1997.

A. de Cheveigné, "Multiple F0 estimation," in *Computational auditory scene analysis: principles, algorithms, and applications,* D.L. Wang and G.J. Brown, Ed., New York: IEEE Press/Wiley, 2006.

J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-time processing of speech signals*, New York: IEEE Press, 2000.

G. Dewey, *Relative frequency of English speech sounds*, Cambridge, MA: Harvard University Press, 1923.

H. Dillon, *Hearing aids*, New York: Boomerang Press, 2001.

L.A. Drake, *Sound source separation via computational auditory scene analysis (CASA)-enhanced beamforming*, Ph.D. Dissertation, Northwestern University Dept. Elec. Eng., 2001.

R. Drullman, J.M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, Vol. 95, pp. 1053-1064, 1994a.

R. Drullman, J.M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, Vol. 95, pp. 2670-2680, 1994b.

D.P.W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. Dissertation, MIT Dept. Elec. Eng. and Comput. Sci., 1996.

Y. Ephraim, H. Lev-Ari, and W.J.J. Roberts, "A brief survey of speech enhancement," in *The Electronic Handbook,* CRC Press, 2005.

Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech, and Signal Process.*, Vol. 32, pp. 1109-1121, 1984.

Y. Ephraim, D. Malah, and B.H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust. Speech, and Signal Process.*, Vol. 37, pp. 1846-1856, 1989.

Y. Ephraim and H.L. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Process.*, Vol. 3, pp. 251-266, 1995.

H. Fletcher, *Speech and hearing in communication*, New York, NY: Van Nostrand, 1953.

D.A. Forsyth and J. Ponce, *Computer vision: A modern approach*, Upper Saddle River NJ: Prentice Hall, 2002.

S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, Vol. 49, pp. 1614-1626, 2001.

J. Garofolo, L. Lamel, *et al.*, "Darpa TIMIT acoustic-phonetic continuous speech corpus," *NISTIR 4930*, 1993.

O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech and Audio Process.*, Vol. 2, pp. 115-132, 1994.

B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, Vol. 47, pp. 103-138, 1990.

S. Greenberg, J. Hollenback, and D.P.W. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *Proc. ICSLP*, pp. 24-27, 1996.

J.H.L. Hansen and M.A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, Vol. 39, pp. 795-805, 1991.

M.L. Hawleyb, R.Y. Litovskyc, and J.F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Am.*, Vol. 115, pp. 833-843, 2003.

H. Helmholtz, *On the sensation of tone*, 2nd ed., New York, NY: Dover Publishers, 1863.

H. Hermansky and S. Sharma, "Temporal patterns (TRAPs) in ASR of noisy speech," in *Proc. ICASSP*, Vol. 1, pp. 289-292, 1999.

W.J. Hess, *Pitch determination of speech signals*, New York: Springer, 1983.

J. Holdsworth, I. Nimmo-Smith, R.D. Patterson, and P. Rice, "Implementing a gammatone filter bank," *MRC Applied Psych. Unit*, 1988.

A. Hoover, G. Jean-Baptiste*, et al.*, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 18, pp. 673-689, 1996.

G. Hu and D.L. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79-82, 2001.

G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," in *Proc. ICASSP*, Vol. 2, pp. 553-556, 2002.

G. Hu and D.L. Wang, "Separation of stop consonants," in *Proc. ICASSP*, Vol. 2, pp. 749-752, 2003.

G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Net.*, Vol. 15, pp. 1135-1150, 2004a.

G. Hu and D.L. Wang, "Auditory segmentation based on event detection," in *Proc. ISCA Tutorial and Research Workshop on Stat. & Percept. Audio Process.*, 2004b.

G. Hu and D.L. Wang, "Separation of fricatives and affricates," in *Proc. ICASSP*, Vol. 1, pp. 1101-1104, 2005.

G. Hu and D.L. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Selected methods for acousitic echo and noise control,* G. Schmidt and E. Haensler, Ed., Berlin: Springer, 2006a.

G. Hu and D.L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio Speech and Language Process.*, in press, 2006b.

X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: A guide to theory, algorithms, and system development*, Upper Saddle River, NJ: Prentice Hall, 2001.

A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, New York, NY: Wiley, 2001.

G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *J. Machine Learning Research*, Vol. 4, pp. 1365-1392, 2003.

C.R. Jankowski, H.H. Vo, and R.P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech and Audio Process.*, Vol. 3, pp. 286-293, 1995.

J. Jensen and J.H.L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech and Audio Process.*, Vol. 9, pp. 731-740, 2001.

T. Joachims, "Making large-scale SVM learning practical.," in *Advances in Kernel Methods - Support Vector Learning, ,* B. Schölkopf, C. Burges, and A. Smola, Ed., Cambridge, MA: MIT Press, 1999.

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 Trans. Neural Net.*, Vol. 15, pp. 1112-1124, 2004.

M.C. Killion, "Revised estimate of minimal audible pressure: Where is the 'missing 6 dB'?," *J. Acoust. Soc. Am.*, Vol. 63, pp. 1501-1510, 1978.

B.E.D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Comm.*, Vol. 25, pp. 117-132, 1998.

A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. ICASSP*, Vol. 6, pp. 3089-3092, 1999.

H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Magazine*, Vol. 13, pp. 67-94, 1996.

R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *J. Acoust. Soc. Am.*, Vol. 105, pp. 1912-1924, 1999.

P. Ladefoged, *Vowels and consonants*, Oxford, UK: Blackwell, 2001.

T.W. Lee, M.S. Lewicki, M. Girolami, and T.J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Process. Letters*, Vol. 6, pp. 87-90, 1999.

Y. Li and D.L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *OSU-CISRC-9/05-TR61* Department of Computer Science and Engineering, The Ohio State University. 2005.

J.C.R. Licklider, "A duplex theory of pitch perception," *Experientia*, Vol. 7, pp. 128-134, 1951.

J. Lim, Ed., *Speech enhancement*, Englewood Cliffs NJ: Prentice Hall, 1983.

R.P. Lippmann, "Speech recognition by machines and humans," *Speech Comm.*, Vol. 22, pp. 1-16, 1997.

C. Liu, B.C. Wheeler, *et al.*, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interfaces," *J. Acoust. Soc. Am.*, Vol. 110, pp. 3218-3231, 2001.

R.F. Lyon, "Computational models of neural auditory processing," in *Proc. ICASSP*, Vol. 9, pp. 41–44, 1984.

D. Marr, *Vision*, New York, NY: Freeman, 1982.

R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Process.*, Vol. 9, pp. 504-512, 2001.

I. Masuda-Katsuse and H. Kawahara, "Dynamic sound stream formation based on continuity of spectral change," *Speech Comm.*, Vol. 27, pp. 235-259, 1999.

R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Acoust. Speech Signal Process.*, Vol. 28, pp. 137-145, 1980.

R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Am.*, Vol. 83, pp. 1056-1063, 1988.

R. Meddis and M. Hewitt, "Modelling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, Vol. 91, pp. 233-245, 1992.

D.K. Mellinger, *Event formation and separation in musical sound*, Ph.D. Dissertation, Stanford University Department of Computer Science Dept. 1992.

B.C.J. Moore, *An introduction to the psychology of hearing*, 5th ed., San Diego, CA: Academic Press, 2003.

T. Nakatani and H.G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Comm.*, Vol. 27, pp. 209-222, 1999.

S. Nooteboom, "The prosody of speech: Melody and rhythm," in *The handbook of phonetic sciences,* W.J. Hardcastle and J. Laver, Ed., Oxford: UK: Blackwell, pp. 640-673, 1997.

K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. of ICASSP*, Vol. 12, pp. 177-180, 1987.

K.J. Palomaki, G.J. Brown, and J. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Comm.*, Vol. 43, pp. 123-142, 2004.

T.W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, Vol. 60, pp. 911-918, 1976.

R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *MRC Applied Psych. Unit.*, 1988.

R.D. Patterson, K. Robinson, *et al.*, "Complex sounds and auditory images," in *Auditory Physiology and Perception,* Y. Cazals, L. Demany, and K. Horner, Ed., Oxford: UK: Pergamon, pp. 429–446., 1992.

J.O. Pickles, *An introduction to the physiology of hearing*, 2nd ed., London, UK: Academic Press, 1988.

R. Plomp, "The ear as a frequency analyzer," *J. Acoust. Soc. Am.*, Vol. 36, pp. 1628-1636, 1964.

R. Plomp and A.M. Mimpen, "The ear as a frequency analyzer II," *J. Acoust. Soc. Am.*, Vol. 43, pp. 764-767, 1968.

J.C. Principe, N.R. Euliano, and W.C. Lefebvre, *Neural and adaptive systems*, New York: Wiley & Sons, 2000.

L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition.," *Proc. IEEE*, Vol. 77, pp. 257-286, 1989.

L.R. Rabiner and B.H. Juang, *Fundamentals of speech recognition*, Englewood Cliffs NJ: Prentice-Hall, 1993.

A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancment," *IEEE Trans. Speech and Audio Process.*, Vol. 9, pp. 87-95, 2001.

N. Roman and D.L. Wang, "A pitch-based model for separation of reverberant speech," in *Proc. Interspeech*, pp. 2109-2112, 2005.

N. Roman, D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, Vol. 114, pp. 2236-2252, 2003.

B.H. Romeny, L. Florack, J. Koenderink, and M. Viergever, Ed., *Scale-space theory in computer vision*, New York, NY: Springer, 1997.

D.F. Rosenthal and H.G. Okuno, Ed., *Computational auditory scene analysis*, Mahwah, NJ: Lawrence Erlbaum Associates, 1998.

J. Rouat, Y.C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Comm.*, Vol. 21, pp. 191-207, 1997.

S.T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems (NIPS'00)*, 2001.

S.T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech*, pp. 1009-1012, 2003.

H. Sameti, H. Sheikhzadeh, L. Deng, and R.L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech and Audio Process.*, Vol. 6, pp. 445-455, 1998.

A. Shamsoddini and P.N. Denbigh, "A sound segregation algorithm for reverberant conditions," *Speech Comm.*, Vol. 33, pp. 179-196, 2001.

R.V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, Vol. 270, pp. 303-304, 1995.

R.V. Shannon, F.-G. Zeng, and J. Wygonski, "Speech recognition with altered spectral distribution of envelope cues," *J. Acoust. Soc. Am.*, Vol. 104, pp. 2467-2476, 1998.

Y. Shao and D.L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Acoust. Speech and Signal Process.*, Vol. 14, pp. 289-298, 2005.

S. Sharma, D.P.W. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, "Feature extraction using non-linear transformation for robustspeech recognition on the Aurora database," in *Proc. ICASSP*, Vol. 2, pp. 1117-1120, 2000.

M. Slaney and R.F. Lyons, "A perceptual pitch detector," in *Proc. ICASSP*, Vol. 1, pp. 357-360, 1990.

S. Srinivasan and D.L. Wang, "A schema-based model for phonemic restoration," *Speech Comm.*, Vol. 45, pp. 63-87, 2005.

K.N. Stevens, *Acoustic phonetics*, Cambridge, Mass.: MIT Press, 1998.

M. Turgeon, A.S. Bregman, and P.A. Ahad, "Rhythmic masking release: Contribution of cues for perceptual organization to the cross-spectral fusion of concurrent narrow-band noises," *J. Acoust. Soc. Am.*, Vol. 111, pp. 1819-1831, 2002.

M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Comm.*, Vol. 27, pp. 261-279, 1999.

V.N. Vapnik, *The nature of statistical learning theory*, Berlin: Springer, 1995.

A.P. Varga and R.K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, pp. 845-848, 1990.

N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Process.*, Vol. 7, pp. 126-137, 1999.

A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, Vol. IT-13, pp. 260-269., 1967.

D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines,* P. Divenyi, Ed., pp. 181-197, 2005.

D.L. Wang, "Feature-based speech segregation," in *Computational auditory scene analysis: Principles, algorithms, and applications,* D.L. Wang and G.J. Brown, Ed., New York: IEEE Press/Wiley, in press, 2006.

D.L. Wang and G.J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Net.*, Vol. 10, pp. 684-697, 1999.

D.L. Wang and G.J. Brown, Ed., *Computational auditory scene analysis: Principles, algorithms, and applications*, New York: IEEE Press/Wiley, in press, 2006.

D.L. Wang and G. Hu, "Unvoiced speech segregation," in *Proc. ICASSP*, in press, 2006.

M. Weintraub, *A theory and computational model of auditory monaural sound separation*, Ph.D. Dissertation, Stanford University Dept. Elec. Eng., 1985.

M. Wu, D.L. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech and Audio Process.*, Vol. 11, pp. 229-241, 2003.