Perceptual Binaural Speech Enhancement in Noisy Environments

Perceptual Binaural Speech Enhancement in Noisy Environments

By

Rong Dong, M.Sc., B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Applied Science

McMaster University

MASTER OF APPLIED SCIENCE (2005) McMaster University

(Department of Electrical and Computer Engineering) Hamilton, Ontario

TITLE: Perceptual Binaural Speech Enhancement in Noisy Environments

AUTHOR: Rong Dong, M.Sc., B.Sc.

SUPERVISOR: Dr. Simon Haykin, Dr. Ian C. Bruce

NUMBER OF PAGES: xiii, 99

Abstract

Speech enhancement in multi-speaker babble remains an enormous challenge. In this study, we developed a binaural speech enhancement system to extract information pertaining to a target speech signal embedded in a noisy background for use in future hearing-aid systems. The principle underlying the proposed system is to simulate the perceptual auditory segregation process carried out in the normal human auditory system. Based on the spatial location, pitch and onset cues, the system can identify and enhance those time-frequency regions which constitute the target speech.

The proposed system is capable of dealing with a wide variety of noise intrusions, including competing speech signals and multi-speaker babble. It also works under mild reverberation conditions. Systematic evaluation shows that the system achieves substantial improvement on the intelligibility of target signal, while it largely suppresses the unwanted background signal.

Acknowledgements

I would like to express my most sincere gratitude to my supervisors Dr. Simon Haykin and Dr. Ian C. Bruce, for their invaluable guidance throughout my pursuit of this degree, and for the freedom they gave me to conduct my research. Without their insightful discussions and constructive feedback, this work would not have been possible.

Special thanks go to my lab mates, Karl Wiklund for providing the testing data. Jeff Bondy and Kevin Kan, for many fruitful discussions and advice regarding the work and willingness to be my listening subjects.

Last but by no means not least, I want to thank my parents for their constant caring and support, Steve for bringing me much joy and inspiration, and my friends for taking me back to the real world once in a while.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Hearing Loss

Hearing loss is one of the most prevalent chronic health conditions, affecting about 500 million people world-wide. According to many surveys, one out of ten people suffers from hearing loss and would benefit from using hearing aids (Hear-it 2004). The most common type of hearing impairment is sensorineural hearing loss, which is typically associated with a dysfunction of the cochlea. People with this kind of hearing loss not only suffer from an increased hearing threshold but also from the reduction of speech intelligibility in noisy environments, which is mainly caused by the loss of temporal and spectral resolution in the processing of the impaired auditory system. To achieve the same intelligibility in a noisy listening condition, hearing-impaired people require an approximate 5-10 dB higher signal-to-noise ratio (SNR) than people with normal hearing (Moore 2003).

1.1.2 Noise Reduction for Hearing Aid Applications

Conventional hearing aids include an amplification stage to compensate for the lifted hearing threshold and optional dynamic compression to compensate for a re-

duced dynamic range in one or more frequency channels. They provide almost complete restoration of speech intelligibility in quiet conditions to the level of normal hearing, but they are not able to restore speech intelligibility in noise (Marzinzik and Kollmeier 1999). It is one of the most common complaints made by hearing-aid users that speech in noise is particularly difficult to understand. This can be explained by the fact that the hearing aids currently in use amplify noise as well as speech and thus do not compensate for any kind of distortion process due to hearing loss. Advanced signal processing techniques for noise reduction can increase the SNR and thereby increase the speech intelligibility, lowering the listening effort and improving the perceived quality of the acoustic environment. Because of the particular damaging effects of background noise on speech intelligibility for people with hearing loss, it is of critical importance to integrate efficient noise reduction techniques into digital hearing aids.

So far, many single-microphone as well as multi-microphone noise reduction algorithms have been proposed in the literature for the application of hearing aid products.

In the last decades, the majority of noise reduction systems proposed for hearing aids have been algorithms for single-microphone input based on spectral subtraction (Ghoreishi and Sheikhzadeh 2000, Wolfe and Godsill 2000). With spectral subtraction, the power spectral density of the clean speech signal is estimated by subtracting the estimated power spectral density of the noise signal from the power spectral density of the corrupted signal. However, single-microphone noise reduction techniques can only differentiate between signals that have different temporal and spectral characteristics. It works well when noise signal is reasonably stationary (e.g., white noise). However, for most of the everyday noises, the frequency spectrum is identical to the spectrum of speech, which makes it difficult for single-channel noise reduction schemes

to effectively eliminate the noise without reducing speech intelligibility at the same time.

To overcome the limitations of spectral subtraction, spatial information can be exploited by use of microphone arrays combined with a beamforming processing algorithm. This technique aims to preserve a target arriving from a known direction while minimizing jammers, which are independent of the target and emitted from other directions (see an overview by Zurek et al. 1999). Multi-microphone arrays are reported to produce considerable directivity, but large microphone arrays are generally required to achieve a good performance. The physical size and the required additional head-worn devices make such an application, as everyday-life hearing aids, very difficult.

As an alternative, recently developed blind source separation (BSS) algorithms can drastically reduce the number of microphones (Bell and Sejnowski 1995, Parra and Spence 2000). BSS relies on the availability of several differing source mixtures and attempts to invert the mixing process in order to recover each individual stream. But the application of BSS is limited by its relatively strict assumptions on the properties of the sources, such as, statistically independent, linear non-singular mixing, known and fixed number of sources, etc.

All these enhancement techniques have difficulty in dealing with the unpredictable nature of general environments such as a "cocktail-party" environment, where a target sound is mixed with a number of acoustic interferences. The interferences could be competing speech sounds or a variety of nonstationary noises. It remains a challenge for a machine (i.e. hearing aid) to extract the desired sound from a multi-source acoustic environment. In contrast, human beings can communicate effectively by sounds in noisy and reverberant environments. This ability stems from the remarkable capacity of human auditory system to separate the sound source of interest from

the complex, composite signal that is received at the ears. This process is called sound stream segregation. We believe that by modelling the neural computational mechanisms involved in sound stream segregation, we will be able to produce a more flexible and more stable speech enhancement algorithm. The model must be built on our understanding of the sound stream segregation carried out by the human auditory system. In the next section, we will briefly introduce the auditory foundation of this process.

1.1.3 Auditory Scene Analysis

One way of explaining auditory sound segregation is to consider the auditory environment as a complex scene containing multiple objects and to hypothesize that the normal auditory system is capable of grouping these objects into separate perceptual streams based on distinctive perceptual cues. The process is often called "auditory scene analysis". A great variety of research relating to auditory scene analysis has been reviewed by Bregman (1990). It can be summarized as follows.

The peripheral auditory system acts as a frequency analyzer, separating the different frequency components in a complex sound. Somewhere in the brain, the internal representations of these frequency components have to be assigned to their appropriate sources. If the input comes from two sources, A and B, then the frequency components must be split into two groups; the components emanating from source A should be assigned to one stream and components emanating from source B should be assigned to another.

The formation and segregation of auditory objects are governed by Gestalt grouping principles of perceptual organization (Bregman 1990):

1. *Proximity*: Elements are more likely to be grouped into a single perceptual stream, if they are in close proximity of time and frequency.

2. *Similarity*: Elements tend to be grouped together, if they are similar in terms of the intensity, pitch, source localization and other properties of grouping cues.

3. *Continuity*: We should take into account the fact that natural speech articulation is a continuous process. Speech signals of a single stream tend to appear smoothly and continuously evolving properties in intensity, pitch, location and other perceptual cues. Any abrupt change along time or frequency indicates segregation.

4. *Closure*: The percept of streams can be completed even when some parts are actually missing. For example, when the sound from one source is masked by another sound, it still can be perceived as a continuous stream, if there is evidence showing the continuity surrounding the masked segment. The masked partials are filled out according to the continuity principle.

5. *Common fate*: Elements tend to be grouped if they undergo coherent variation along time or frequency, e.g., having the same timing event or modulated at the same rate. On the other hand, both onset asynchrony and difference in the pattern of modulation (AM or FM) lead to segregation.

Bregman distinguishes two types of mechanism that can be used to determine which components belong to a particular source: *Primitive grouping mechanism* (bottom-up) partitioning of the input on the basis of simple perceptual cues, whereas *schema governed mechanisms* (top-down) can exploit prior knowledge with the source and patterns of language to recover the masked or distorted signal. The attraction of primitive mechanism is that it can exploit the general properties of sound sources

without knowing what is going to be heard. Its context-independent nature makes this kind of mechanism particularly very straightforward, hence well suited to be implemented in a machine.

Over the last decade, several researchers have attempted to build computational frameworks that perform auditory scene analysis; the resulting field has been called computational auditory scene analysis (Rosenthal and Okuno 1998). Typically, these computational auditory scene analysis models involve implementation of some small subset of the strategies suggested by Bregman, often in a manner functionally consistent with the early stages of human auditory periphery (as they are currently understood).

1.2 Purpose of The Study

The objective of this work is to develop an adaptive hearing system that extracts a target voice of interest from other interferences. The system is primarily targeted for application in hearing aids with two microphones. As the front-end processor for hearing aids, it helps make up for the perceptual grouping process missing from the auditory system of hearing-impaired person.

As a hearing system for practical application, we aim to design a system which is capable of solving the following three challenges: 1. It must be effective over a wide variety of conditions of interference. For example the interference may be one competing voice, or multiple competing voices; it may be environment noise or a mixture of noise and competing voices under normal reverberation conditions. Usually such a condition is unknown in advance. 2. The total processing time of the algorithm must be very short to avoid the perception of asynchronies between processed sound and bone-conducted sound when monitoring one's own speech, or

between vision and hearing when monitoring the speech of another talker. 3. It must be computationally efficient and ultimately suitable to be implemented on a digital signal processing (DSP) chip for real-time processing.

1.3 Approach

When the input to the system consists of a signal embedded in a complex background of interesting sounds, we can make the reasonable assumption that the spectrum of the foreground and background sounds are different, in which case, some of our channels will be dominated by the foreground and some by the background sound source. If we can determine which channels are dominated by the target sound source, we may seek to enhance our signal-to-noise ratio by selecting (or merely emphasizing) those channels which are dominated by the target signal (foreground sounds source). Conversely, we may attenuate the output of those channels that are dominated by the background (non-target sound source).

In a "cocktail party" environment, people can selectively focus on a primary stream at a time. Consequently, the acoustic properties of the attended speech stream will be enhanced and appear to be more prominent than the background. By contrast, a computational system must rely on some prior knowledge about the target stream in order to distinguish the target stream from the background. Since the hearing aid user can flexibly steer his/her head to the desired source direction (actually, even normal hearing people need to take advantage of directional hearing in a noisy listening environment), it is reasonable to assume that the desired signal comes from the frontal centre direction, while the interference comes from off-centre. To ensure that the extracted target signal is intelligible, we also assume the interference will be lower in energy than target over a significant portion of the time-frequency plane.

The architecture of the system, illustrated in Figure 1.1, is psychophysically motivated by the primitive segregation mechanism used in human auditory scene analysis. Specifically, the model performs bottom-up segregation of an incoming signal as follows:

1. The function of the cochlea is approximated to generate a time-frequency representation of the incoming signal. The output from the cochlear filterbank is processed by a simple model of the inner hair cells, which simulates the nonlinear neural transduction in inner hair cells.

2. For each elementary time-frequency unit, a set of perceptual cues is extracted to reveal its particular acoustic properties. The cues used in the current model are interaural intensity difference (IID), interaural time difference (ITD), onset and pitch (F0). The multiple cue extractions are preformed in a parallel fashion.

3. Information from multiple cues are integrated and time-frequency units that correspond to coherent auditory objects are grouped by exploiting the correlation of the cues. Knowing the direction of the target signal of interest, the target auditory objects can be identified and enhanced. The other elements belonging to interference stream are suppressed.

4. Once a time-frequency representation of the target sound is obtained, it can be inverted in order to reconstruct a time waveform for the enhanced target.

1.4 Thesis Organization

Chapter 1 starts with an introduction of the motivation and objectives driving the current research, then presents an overview of the proposed model along with a summary of the component parts.

The second part of the thesis, consisting of Chapters 2 to 4, presents the individual sections of the model. In Chapter 2, implementation of the auditory peripheral model, including the cochlear filterbank and the inner hair cell model, is described in detail. For the purpose of real-time resynthesis at the end of processing, we also present a low-delay filterbank inversion method in this chapter. Chapter 3 discusses the property and estimation algorithm for each auditory perceptual cue. Chapter 4 describes the strategy to combine the evidence from different kinds of cues and make a grouping decision.

Chapter 5 presents evaluation results of the system. Finally, Chapter 6 summarizes the conclusions of the current work and provides some suggestions for future research.

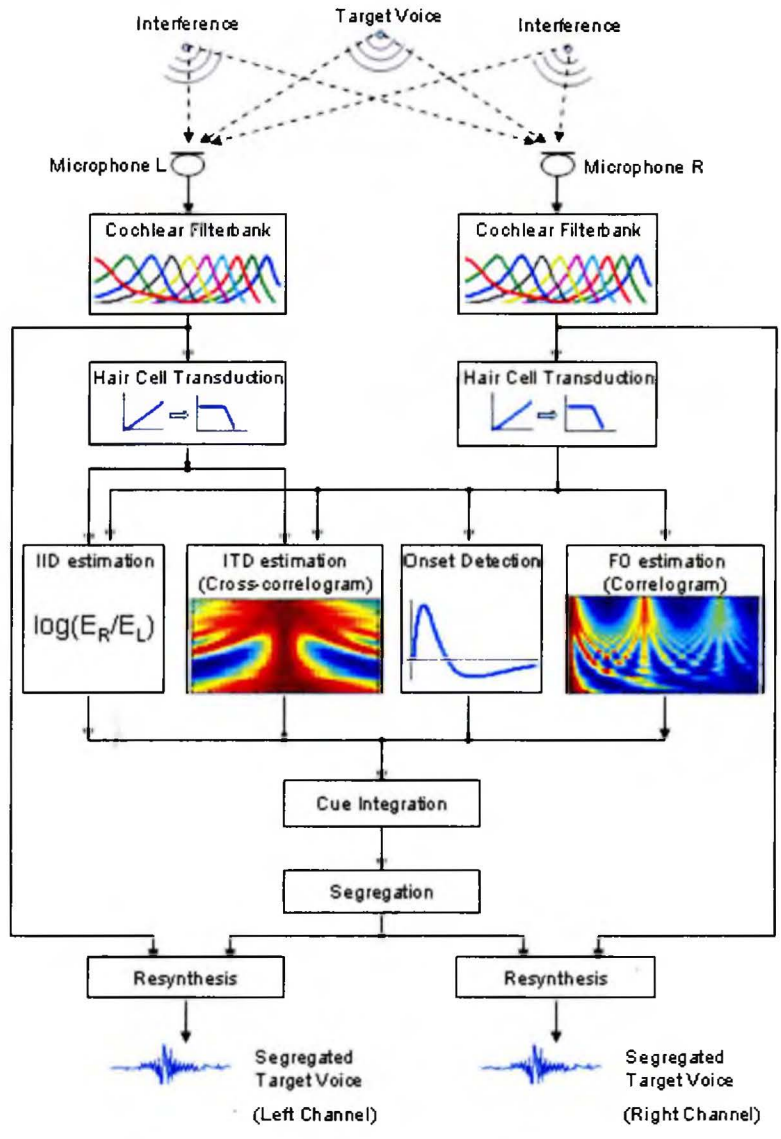


Figure 1.1: Block diagram of the system.

Chapter 2

Auditory Periphery Model and Its Inversion

Sounds from several sources arrive at the ear as a complex mixture. They are largely overlapping in the time domain. In order to organize sounds into their independent sources, it is often more meaningful to transform the signal from the time domain to an internal time-frequency representation. The time-frequency analysis at the first stage of our system is auditory-motivated, which mimics the frequency selectivity of the human cochlea. Specifically, the input signal is passed through a bank of bandpass filters, each of which simulates the frequency response associated with a particular position on the basilar membrane. The output from the cochlear filterbank is processed by a simple model of the inner hair cell, which simulates the nonlinear neural transduction in the inner hair cell. The implementation details of the cochlear and inner hair cell model will be given in Section 2.1. The auditory periphery model will introduce a phase lag into the output signal. In addition, this phase lag is frequency-dependent. To give a synchronous representation of auditory event, an explicit phase alignment is required to compensate for the phase lag, which will be described in Section 2.2. At the end of processing, the enhanced sound need to be resynthesized from its time-frequency representation. In Section 2.3, we describe a low-delay filterbank inversion method to facilitate real-time reconstruction.

2.1 Auditory Periphery Model

2.1.1 Cochlear Filterbank

The frequency decomposition performed by the cochlea is simulated using a bank of Gammatone filters (Slaney 1993). The impulse response of a Gammatone auditory filter is

$$h(t) = at^{n-1}e^{-2\pi bERB(f_c)t}\cos(2\pi f_c t + \phi) \tag{2.1}$$

where a, b are constants, ϕ is a phase shift, $n = 4$ is order of the filter, f_c is the centre frequency, and $ERB(f_c)$ is the equivalent rectangular bandwidth (ERB) corresponding to its centre frequency. ERB is determined by the bandwidth of the human auditory filter at different characteristic frequencies along the cochlea. In our implementation, the ERB value at the centre frequency f_c follows the following formula (Slaney 1993)

$$ERB(f_c) = 24.7 + f_c/9.26 \tag{2.2}$$

In the human auditory system, there are around 3000 inner hair cells along the 35mm length of the cochlea. Each hair cell could resonate to a certain frequency within a suitable critical bandwidth. This means that there are approximately 3000 bandpass filters in the human auditory system. This resolution of filters can not be implemented practically using computational modelling techniques. However, we can approximate this density of channels. It can be achieved by specifying the number of filters and a certain frequency range to be covered. Centre frequencies of filters are spaced so that each filter overlaps its neighbors by the same amount. The summation of all filter frequency responses would result in a flat magnitude across frequency. Logarithmic

spacing was used for computational convenience. The Equation 2.2 can then be solved to find the proper centre frequency spacing.

The particular implementation of Gammatone filters is based on the work of Slaney (1993). It is simply a cascade of four second-order IIR filters. Equation 2.3 gives the form of each second-order IIR filter, whose structure is illustrated in Figure 2.1.

$$h(z) = \frac{a_0 + a_1 z^{-1}}{1 + b_1 z^{-1} + b_2 z^{-2}} \tag{2.3}$$

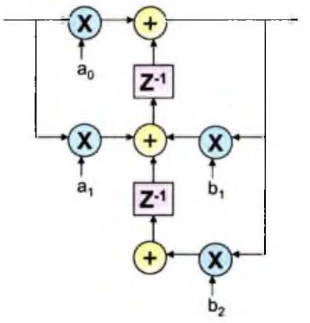


Figure 2.1: Structure of a second-order IIR filter in the implementation of Gammatone filter.

The sampling rate of the original sounds specific for this implementation was fixed at 16 kHz, thus the Nyquist frequency for these filters is at 8 kHz. There are 32 filters used starting from 100 Hz to the highest centre frequency at 7596 Hz, just below the Nyquist frequency. The composite frequency response of the Gammatone filterbank is show in Figure 2.2.

Generally, having a greater number of frequency bands leads to better frequency resolution. In Chapter 5, we will evaluate the effect of increasing the number of frequency bands.

Note that Gammatone filter provides a linear and impulse-invariant transform. Although it does not exactly reflect the nonlinear, dynamic function of the cochlea, it

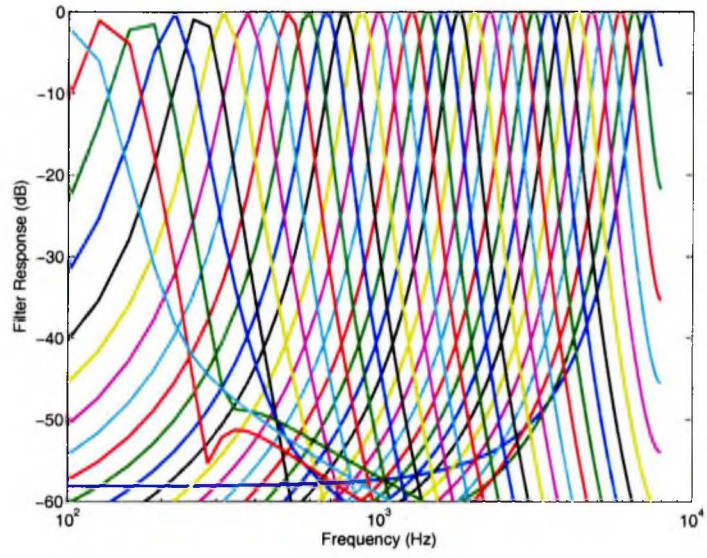


Figure 2.2: Magnitude of frequency responses of a 32-channel Gammatone filterbank.

has the advantage of simple structure and efficient computation. In addition, linearity of analysis filterbank is a very important property to make the inversion (resynthesis) possible at the end of processing.

2.1.2 Inner Hair Cell Transduction Model

The signal at the output of the Gammatone filterbank is half-wave rectified and low-pass filtered at 1 kHz. This processing roughly simulates the transduction in the inner hair cells. Basically, the inner hair cell model performs envelope extraction in the high-frequency band, while passing the signals in the low-frequency band.

As an example, Figure 2.3 shows the output of cochlear filterbank and inner hair cell model in response to an impulse train of repetition rate 200/s. This stimulus has a fundamental frequency of 200 Hz and contains all the harmonics in the series. At the output of the low-frequency Gammatone filters, because the resolution of filters is sufficient to separate the harmonics, the output of each individual filter is approx-

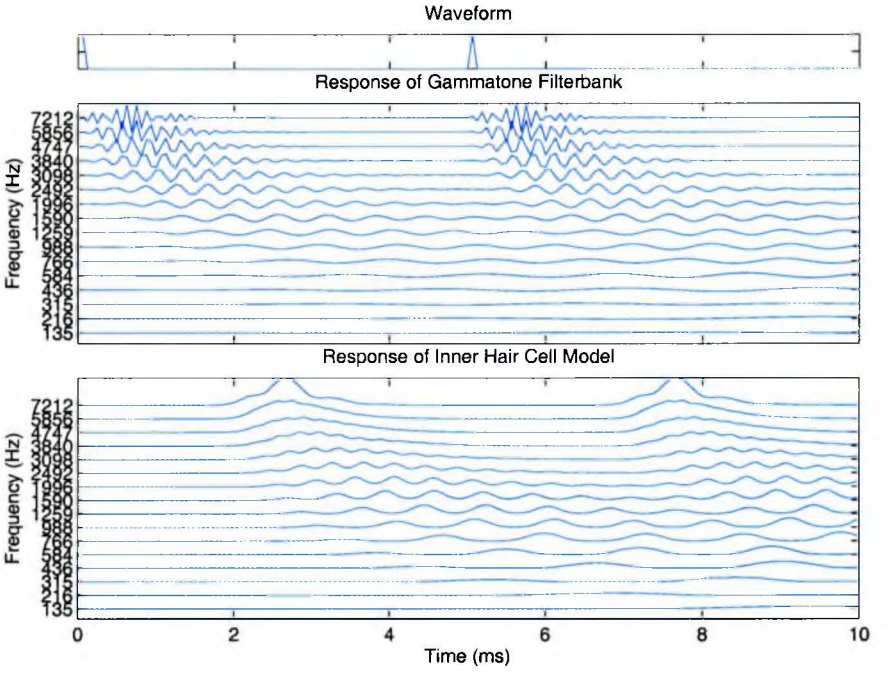


Figure 2.3: Time responses of Gammatone filters and inner hair cell model to an impulse train of repetition rate 200/s.

imately a sinusoidal waveform. These harmonics are so called "resolved harmonics". As the bandwidth of the filters becomes broader at the higher frequencies, several adjacent harmonics will pass through the same filter, called "unresolved harmonics". The output in this kind of frequency channel is not simply a sinusoid. Beating between the adjacent harmonics will cause an amplitude-modulated filter output. The modulation frequency corresponds to the fundamental frequency. Then through the inner hair cell model, the low-pass filtering essentially preserves the envelope of the signal at those high-frequency channels.

2.2 Phase Alignment

From Figure 2.3, we can see there is a strong rightward skew at the output of an auditory peripheral filter. This can be interpreted as a wave that starts at the high-frequency side of cochlea and travels down to the low-frequency side with a finite propagation speed. The low-frequency side shows a 10 ms or even longer phase lag compared to the high-frequency side. Information carrying by natural speech signals are non-stationary, especially during the rapid transition (e.g., onset). A form of phase alignment is thus required to compensate for the phase difference and thereby align the frequency channel responses to give a synchronous representation of auditory events. Normally, this is done by time-shifting the response with the value of a local phase lag, so that the impulse responses of all the frequency channels reflect the moment of maximal excitation at approximately the same time. This approach entails that the response of high-frequency channels at time t is lined up with the response of low-frequency channels at $t + 10$ ms or even later. A real-time system for hearing aid applications obviously cannot afford such a long delay. In our implementation, each channel is only advanced by one cycle of its centre frequency. Given the lowest centre frequency 100 Hz, the maximum phase lag compensation is 10 ms. With this phase compensation scheme, the onset timing is nearly aligned across frequencies.

The low-pass filter of the inner hair cell model produces an additional 2 ms group delay in the auditory peripheral response. Unlike the phase lag in the cochlear filterbank, this delay is constant across frequency channels, and hence it does not cause asynchrony across frequencies. Still we cannot ignore this group delay in resynthesis.

2.3 Resynthesis

In the auditory peripheral model, frequency decomposition is applied on the incoming sound signal. Subsequent grouping and segregation will be performed on the time-frequency plane. As a result, a group of sound elements will be assigned to the target stream. Due to the fundamental requirement of hearing-aid application, at the end of processing the desired waveform must be reconstructed and conducted to the ears.

Because of the linearity of Gammatone filters, the cochlear filterbank used in our system is completely invertible. Hair cell model inversion is hard due to its nonlinear nature. Though the perceptual cue will be estimated from the output of inner hair cell model, segregation is intended to be performed on the output of cochlear model so that the enhanced waveform can be faithfully recovered. In this section, a scheme of low-delay cochlear filterbank resynthesis is described for our particular application.

The framework of Gammatone analysis/synthesis filterbank is illustrated in Figure 2.4.

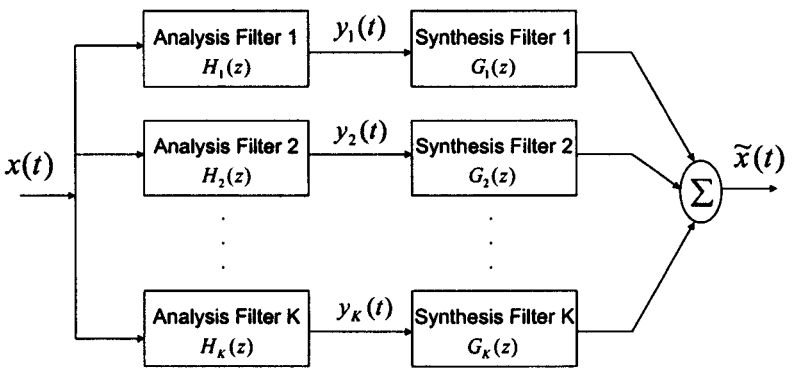


Figure 2.4: Gammatone filterbank analysis/resynthesis.

Let K be the number of channels in auditory peripheral model, and $H_k(z)$, $k = 1, 2, ..., K$, be the transfer function of the individual analysis filter. Equation 2.4 gives an ideal choice for the transfer function of synthesis filter:

$$G_k(z) = \frac{H_k^*(z)}{\sum_{k=1}^{K} H_k(z)H_k^*(z)} \tag{2.4}$$

Using this definition, the overall analysis/synthesis system approximates an all pass filter. The synthesis filterbank produces an exact reconstruction of the original input signal $x(t)$. Since Gammatone filters are distributed on the ERB scale, summation of the transfer functions across all the frequency channels is approximately a constant:

$$\sum_{k=1}^{K} H_k(z)H_k^*(z) = \sum_{k=1}^{K} |H_k(z)|^2 \approx c \tag{2.5}$$

The constant c is a value dependent on the number of frequency bands. The estimated relationship is listed in Table 2.1.

Table 2.1: Scaling factor c in resynthesis

Number of frequency bands K	Scaling factor c
32	0.96
64	0.472
128	0.236

Therefore, in the time domain the impulse response of the synthesis filter is equivalent to the time-reversed impulse response of its corresponding analysis filter. By noting that $H_k(z)$ is implemented as an IIR filter and $H_k^*(z)$ is equal to $H_k(\frac{1}{z})$, $G_k(z)$ must be noncausal and unstable filter. Hence, a direct implementation of this solution is not practical. An alternative approach (Lin, Holmes, and Ambikairajah 2001) is to make the synthesis filters exactly the same as the IIR analysis filters, while time reversing both the input signal of each synthesis filter $y_k(t)$ and the output $\tilde{x}(t)$ to achieve a linear phase response. Due to the causality of Gammatone filter, the

output $\tilde{x}$ at time t depends on the future values of the input signal $y_k(t)$. Thanks to the limited effective duration of impulse response of Gammatone filter, for real-time application, this time reversal process can be implemented in a frame-by-frame fashion as illustrated in Figure 2.5. In the following, we summarize the synthesis implementation algorithm developed for our system:

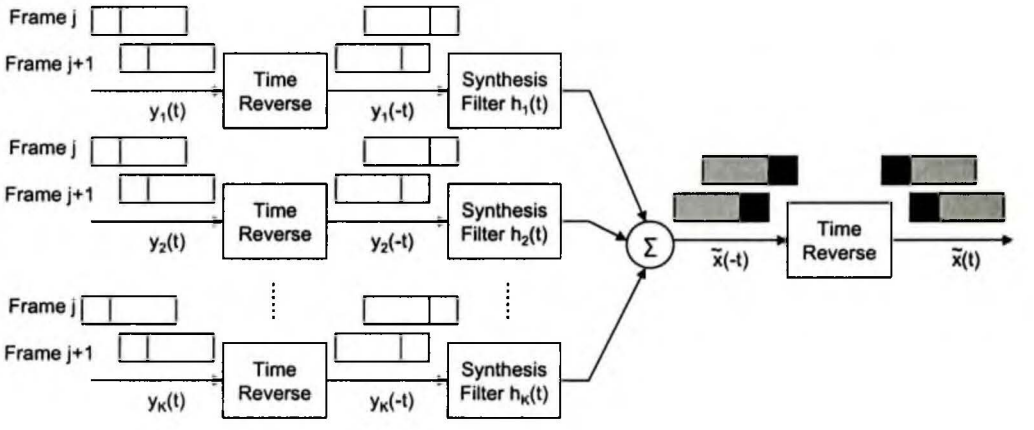


Figure 2.5: Resynthesis.

1. The input signal $y_k(t)$ of each synthesis filter is decomposed into a sequence of frames. The choices of window size and the frame rate will be discussed later.

2. For each channel, the data of j-th frame is time reversed.

3. The reversed signal is filtered through the Gammatone synthesis filter, whose transfer function is exactly the same as in the analysis filter.

4. The output signals are summed up across the channels and then time reversed to produce one frame estimate of $\tilde{x}(t)$.

5. Given a long enough window size, the beginning segment (black colored in Figure 2.5) of the output $\tilde{x}(t)$ is concatenated with the same segment from the

previous frame to form the final reconstructed signal. The second segment of output data represented in grey color is discarded without enough future data available to give an accurate estimation.

6. Repeat step 2 to continue processing of the next frame.

Given a fixed window size, increasing frame rate leads to better reconstruction and a lower average delay. However it also increases the computational load of the system. In this implementation, the frame rate is chosen as 1000 frames/s.

The window length plays a very important role in the performance of resynthesis. The signal-to-noise (SNR) is one of the most simple and common objective measures for evaluating the accuracy of reconstruction. This is given by

$$SNR = 10\log_{10}\left\{\frac{\sum_t x^2(t)}{\sum_t (x(t)-\tilde{x}(t))^2}\right\} \tag{2.6}$$

where $x(t)$ is the original signal and $\tilde{x}(t)$ is the reconstructed signal. The SNR in terms of window size is plotted in Figure 2.6. With a window length longer than 15 ms, the distortion is perceptually inaudible. For a window length less than 15 ms, the reconstruction accuracy is highly sensitive to the window length. This can be explained by examining the impulse response of the Gammatone filter. Within the low frequency range, most of the impulse responses last up to 15 ms. If the window length is reduced to be less than 15 ms, the accuracy of reconstruction is degraded. On the other hand, the maximum delay of this algorithm is directly determined by the window length. Therefore, the choice of window length is a trade off between reconstruction distortion and algorithm delay. In order to preserve the reconstructed speech quality, a window length of 15 ms is adopted in the current implementation.

An example of a resynthesized signal is displayed in Figure 2.7. The distortion is defined as the difference between the original signal $x(t)$ and the reconstructed signal

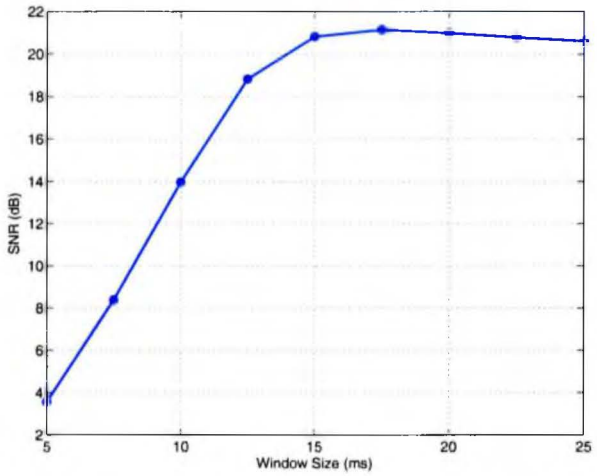


Figure 2.6: Reconstruction accuracy in terms of synthesis window size.

$\tilde{x}(t)$. In this particular case, an SNR of 20.11 dB is achieved. Informal listening test shows that the reconstructed signal is perceptually indistinguishable from the original.

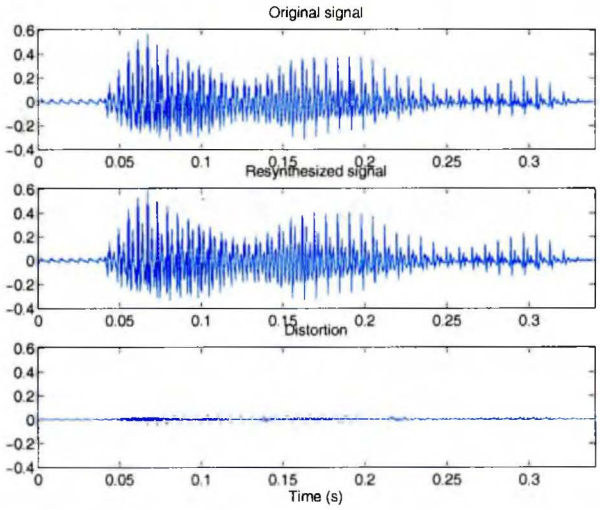


Figure 2.7: Comparison of reconstruction result.

Chapter 3

Acoustic Cues Selection and Extraction

Many different sorts of acoustic cues may be used to derive separate perceptual streams corresponding to the individual sources. This chapter concentrates on those cues used in our model to achieve perceptual segregation and grouping. These are pitch, onset, and binaural spatial cues. For each cue, we will first explain why listeners can use this cue in auditory grouping from the psychoacoustic perspective. Then we will review the existing computational models to extract the cue and describe the specific extraction mechanism employed in our system. After that, we will verify the extraction algorithm both on clean and noisy signals. In the last section, we will further discuss the effects of reverberation on each of these cues.

3.1 Pitch and Harmonic Relationship

3.1.1 Psychoacoustic Evidence

Pitch is the perceptual attribute related to the periodicity of a sound waveform. For a periodic complex sound, pitch is the fundamental frequency (F_0) of a harmonic signal.

The common fundamental period across frequencies provides a basis for associating speech components originating from the same larynx and vocal tract (Bregman 1990; Langer 1992; Meddis and Hewitt 1992). Compatible with this idea, psycholog-

ical experiments have revealed that components that are harmonics of a common F_0 tend to fuse together. Periodicity cues in voiced speech contribute to noise robustness via auditory grouping processes. When pairs of synthesized vowels are presented simultaneously, listeners are able to identify them more accurately if they are synthesized with different F_0s, compared to the cases where

1. both have the same fundamental (Assmann and Summerfield 1994)
2. one is voiced and the other is noise-excited (Scheffers 1983)
3. both are noise-excited (Scheffers 1983)

Similarly, Bird and Darwin (1997), and Assmann (1999) have shown that synthesized target sentences are easier to understand in the presence of a continuous speech masker if targets and maskers are synthesized with different F_0s than with the same F_0.

3.1.2 Existing Models of Pitch Extraction

Robust pitch extraction from noisy speech is a nontrivial process. Numerous computational models have been proposed to account for the periodicity of pitch perception. They can be generally divided into two broad classes: spectral models and temporal models (Moore 2003).

Spectral models assume that the pitch of a complex stimulus is derived from its spectral profile defined along the tonotopic axis of the cochlea. This class of models normally involves two stages. In the first stage, a frequency analysis is performed to determine the frequencies of individual components contained in the input. The second stage is a pattern matching process, where the frequencies of harmonics are compared to internally stored spectral templates. These templates consist of the harmonic series of all possible fundamentals. The model tries to find the template with

an F_0 whose harmonics give a closest match to the spectrum. Therefore, the spectral model is also called the "pattern-match" model in the literature. However, this model is dependent on the spectral resolution of individual components in the stimulus and therefore incapable of explaining the residue pitch associated with the unresolved high-frequency harmonics. To address this difficulty, an alternative mechanism is required.

Temporal models derive a pitch estimate by pooling timing information taken across auditory nerve fibers without regard to the spectral profile. The timing information can be encoded by first or higher order intervals (Cariani and Delgutte 1996a, 1996b; Rhode 1995), or measured by the auto-correlation of the responses (Licklider 1951, Lyon 1984; Slaney and Lyon 1990; Meddis and O'Mard 1997). The auto-correlation analysis is adopted in our model for pitch estimation. Full details of this approach will be given in the next section. Compared with the spectral models, the temporal models provide a unified mechanism to account for a diverse range of pitch phenomena (Meddis and O'Mard 1997), including the residue pitch associated with high order, spectrally unresolved harmonics, as well as the periodicity pitch evoked by low order, resolved harmonics.

3.1.3 Pitch Extraction

The specific type of analysis we use to measure pitch is auto-correlation. It is a process whereby the output at each channel of the auditory-periphery model is correlated with a delayed version of the same signal. At each time instance, the results are visually displayed in a two-dimensional (centre frequency $\times$ lag) representation, named the correlogram. For a periodic signal, similarity is greatest at lags equal to

integer multiples of the period. This results in peaks in the auto-correlation function (ACF) that can be used as a cue to periodicity.

Different definitions of the ACF can be used. For dynamic signals, we are interested in the periodicity of the signal within a short window. This short-time ACF is defined as:

$$\mathrm{ACF}(i,j,\tau) = \frac{\sum_{k=0}^{K-1} x_i(j-k)x_i(j-k-\tau)}{\sum_{k=0}^{K-1} x_i^2(j-k)} \tag{3.1}$$

where $x_i(j)$ is the jth sample of the signal at the ith frequency channel, τ is the lag, K is the integration window length and k is the index inside the window. This function is normalized by the instantaneous channel energy $\sum_{k=0}^{K-1} x_i^2(j-k)$. With this normalization, the dynamic range of results is restricted to [-1,1], which facilities an easier thresholding decision. Normalization can also equalize the peaks in channels whose absolute energy might be quite low compared to other frequency channels. Note that all the minus signs in Equation 3.1 ensure this implementation is causal.

The discrete correlation theorem (Proakis and Manolakis 1995) says that discrete correlation of two real signals g and h is one member of the discrete Fourtier transform pair:

$$Corr(g(n), h(n)) \Leftrightarrow G(k)H(k)^* \tag{3.2}$$

where $G(k)$ and $H(k)$ are the discrete Fourier transforms of $g(n)$ and $h(n)$, and the asterisk denotes complex conjugation. Based on this theorem, we can compute correlation more efficiently using fast Fourier transform (FFT). The particular implementation of the numerator of Equation 3.1 is as follows: FFT a window of signal, $x_i(n)$,$n = j-K+1, j-K+2, \ldots, j$, multiply the resulting transform by the complex conjugate of itself, then inverse transform the product. Normally, the result will be a complex vector. However, it will turn out to have all the imaginary parts zero since

the original signal is real. Meanwhile, the normalization term in the denominator is simply the auto-correlation value at zero lag.

Theoretically, the correlogram should be updated at every sample instance. Considering that the sampling rate of the incoming signal is 16 kHz, updating the value of the correlogram at each sample instance would be computationally inhibitive. Instead, we subsample the correlogram to a more tractable rate of 100 frames per second. Choosing a proper auto-correlation window length is critical. First, the window must span more than one fundamental period to get an accurate estimate. The voiced speech of a typical adult male has a fundamental frequency from 85 to 155 Hz, and the value for a typical adult female ranges from 165 to 255 Hz (Baken 1987). Therefore, the window size must be longer than 11 ms. On the other hand, when we compute the short-time auto-correlation and FFT of a speech signal, a shorter window is required to satisfy the quasi-stationary assumption. Otherwise, the periodicity can be distorted by fast frequency transitions within the frame. In our implementation, we choose a 20 ms rectangular window, which is twice as long as the frame sampling interval. Due to the limited pitch range of speech signals, the interested ACF lag can be narrowed down to 3.1~12.5 ms, which covers a large pitch range from 80 to 322 Hz.

The top panel of Figure 3.1 shows a correlogram for a synthesized vowel /a/ with its F_0 centred at 200 Hz. Because the properties of synthesized sounds are generally well-defined in comparison to natural voices, the properties of the auto-correlation analysis can be clearly demonstrated by applying it to a synthesized sound.

The auto-correlation function reaches the maximum value at zero lag. This value is usually normalized to unity.

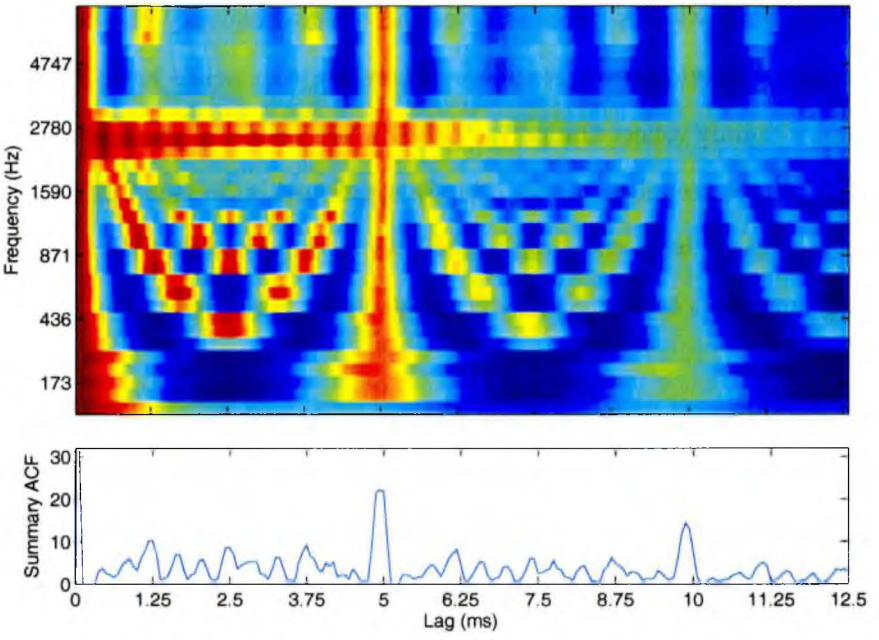


Figure 3.1: Auto-correlation functions (ACFs) (top panel) and summary auto-correlation function (SACF) (bottom panel) for a synthesized vowel /a/ with 200 Hz fundamental frequency.

In the correlogram, the vertical structure at $T_0 = 5$ ms represents the common periodicity across the frequency channels, which is an indication of a 200 Hz fundamental. Since a given fundamental period of T_0 will result in peaks at lags of $2T_0$, $3T_0$, etc., this vertical structure is repeated at lags of multiple periods with comparatively lower intensity.

For resolved harmonics, generally, the nth harmonic displays peaks at a lag of T_0/n and its integer multiplies. This explains the hyperbolic contours shown in the lower half of the correlogram. It also explains the larger number of ACF peaks with increasing centre frequency.

The ACFs for unresolved harmonics are very different in shape from the ACFs for resolved harmonics. Due to the low-pass filtering in the inner hair cell model, the fine structure is removed from high-frequency channel outputs. As a result, the fine

structure information is removed from these high-frequency channel outputs. Only the temporal envelopes are reserved. Therefore, the peaks in ACFs for unresolved harmonics mainly reflect the periodicities in the temporal modulation, not the periodicities of subharmonics. As we already mentioned in Section 2.1.2, this modulation rate is associated with the pitch period, which is represented as a vertical structure at pitch lag across high-frequency channels in the correlogram.

In the correlogram, a common fundamental period across frequencies is represented as common peaks at the same lag. In order to emphasize the vertical structure in the correlogram, a conventional approach is to sum up all the ACFs across the frequency channels. In the resulting summary ACF (SACF), a large peak should occur at the period of the fundamental. Since the ACF peaks have widths proportional to the period, the ACF peaks of very low-frequency channels and unresolved harmonics are generally very broad. As a result, the SACF is inevitably flattened out and unable to provide an accurate estimate of pitch lag. In addition, when multiple competing acoustic sources are present, the SACF may fail to capture the pitch lag of each individual stream. The simplest case is two concurrent vowels: vowel A with fundamental frequency F_{0A} and vowel B with fundamental frequency F_{0B} ($F_{0A} \neq F_{0B}$). In the correlogram, the frequency channels dominated by vowel A display common peaks at the lag of $1/F_{0A}$. However, the other channels dominated by vowel B may display valleys at the same lag. Summarizing the ACFs, these negative valleys introduced by vowel B can cancel out the peaks at lag $1/F_{0A}$, and therefore impair the pitch perception of vowel A. To reduce these undesired effects, we select all the local maxima in each ACF as well as their two immediate neighbors along the lag and sum them up across frequency channels. The SACF obtained in this way displays sharper and more salient peaks indicating the common period of the signal. This implementation also allows for slight peak deviation across frequencies, which often occurs in real signals. Back

to the example of synthesized vowel /a/ in Figure 3.1, the SACF is displayed in the lower panel. In this plot, the common periodicity across frequencies manifests itself as a prominent peak at the lag of 5 ms.

The temporal pitch model was also tested against different noise conditions to verify its robustness. In the three tests, the target periodic signal, a synthesized vowel /a/ with fundamental frequency of 200 Hz, was mixed with three types of masking noises, i.e., white Gaussian noise, concurrent vowel and cocktail-party babble noise. The correlogram of the clean target signal was plotted in Figure 3.1. The left column of Figure 3.2 shows the correlograms of the three maskers:

1. White Gaussian noise: it is a typical aperiodic signal. The correlogram of white Gaussian noise is completely different from the correlogram of a vowel. In each low frequency channel, the ACF response to the narrow-band signal displays peaks characterizing its centre frequency. The peaks vary continuously with increasing centre frequency and do not agree on a common periodicity. As can be seen from Figure 3.2 (a), there is no vertical structure in the correlogram of white Gaussian noise. At high frequency channels, the ACFs reflect repetition in the temporal envelope. Since white Gaussian noise has a flat spectrum, each channel has an equal contribution to the summation of ACFs. Except for the maximum value at zero lag, there are no other prominent peaks shown in the SACF of white Gaussian noise.

2. Concurrent vowels: in this case, the masker is a synthesized vowel /u/ with a fundamental frequency of 150 Hz. Figure 3.2(c) demonstrates the correlogram in response to this masker. It is identical to the correlogram of the vowel /a/ shown in Figure 3.1 except the position of the vertical structure, which is determined by the fundamental period.

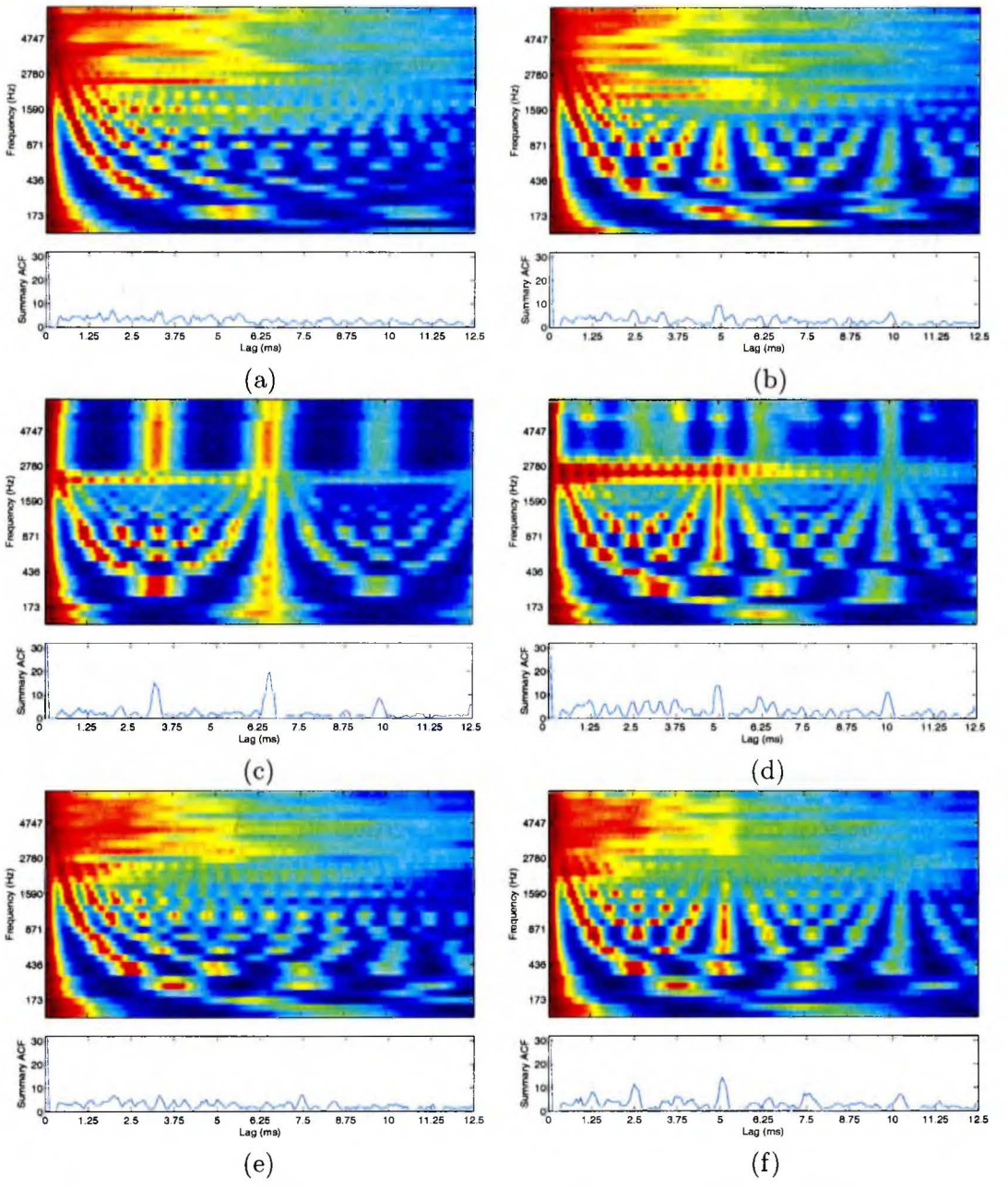


Figure 3.2: ACFs and SACFs computed for (a) white Gaussian Noise; (b) vowel /a/ with a pitch of 200 Hz contaminated by white Gaussian noise (SNR = 5 dB) (c) vowel /u/ with a pitch of 150 Hz; (d) double-vowels: target vowel /a/ with a pitch of 200 Hz mixed with masking vowel /u/ with a pitch of 150 Hz (SNR = 5 dB); (e) cocktail-party babble noise, (f) vowel /a/ with a pitch of 200 Hz contaminated by cocktail-party babble noise (SNR = 5 dB).

3. Cocktail-party babble noise: Figure 3.2(e) demonstrates the correlogram of cocktail-party babble noise. Babble noise is highly nonstationary in spectrum. As a consequence, its correlogram is quite random. There is no common fundamental period that can be detected across frequencies.

The right column of Figure 3.2 illustrates the effect of adding maskers to the target signal. Due to the nonlinear processing in the hair cell model, the correlogram of the mixed signal is not simply a superposition of the correlograms of individual signals, even though the input signals are uncorrelated. Generally, only those frequency components dominated by the target signal show strong peaks at the corresponding pitch lag. This forms the basis of sound separation using the correlogram. Given enough channels dominated by the target signal, the common fundamental period of these channels still manifests itself as a prominent peak in SACF.

3.2 Common Onset

3.2.1 Psychoacoustic Evidence

Onset refers to the beginning of a discrete event in an acoustic signal, which is caused by a sudden increase in energy. The rationale behind onset grouping is that the energy in different frequency components excited by the same source usually starts at the same time. Hence common onsets across frequencies are interpreted as an indication that these frequency components originated from the same sound source. On the other hand, asynchronous onsets enhance the separation of acoustic events. For example, artificially introducing a certain amount of onset asynchrony can even decompose a harmonic stimulus into several partials and dramatically change the perception of the timbre (Mellinger and Mont-Reynaud 1996). This effect is illustrated

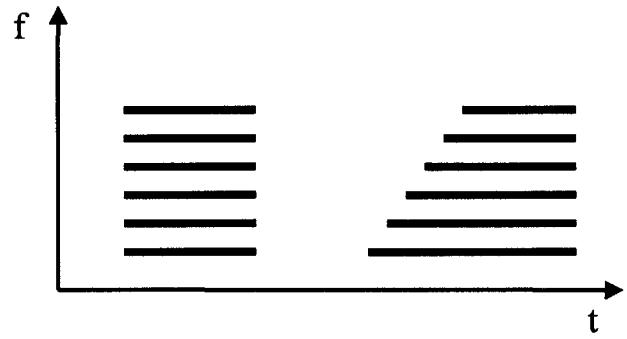


Figure 3.3: The effect of onset asynchronies on fusion. Each horizontal represents a sinusoidal tone. At left, all sinusoids fuse together to a single auditory object. At right, successive tones begin at intervals of 1s and stand out briefly before merging with the rest of the complex (after Mellinger and Mont-Reynaud 1996).

in Figure 3.3. If a short tone is played with all harmonics starting synchronously, a single pitch is perceived. If a delay of one second is introduced between the onset of successive partials, each harmonic stands out briefly as a separate tone before merging with the existing sound.

Since every sound source has attack time, the onset cue does not require any particular kind of structured sound source. It assumes only that the spectral components have reasonably synchronous onsets and tolerably short attack times. In contrast to the periodicity cue, the onset cue will work equally well with periodic and aperiodic sounds. However, when concurrent sounds are present, it is hard to know how to assign an onset to a particular sound source and the system could be prone to switch indiscriminately between emphasizing foreground and background objects. Even for a clean sound stream, it is difficult to distinguish genuine onsets from the gradual changes and amplitude modulations during the sound. Therefore, a reliable detection of sound onsets is a very challenging task.

Sometimes a common offset is also taken into account as a cue. But offset is extremely sensitive to masking noise. As a consequence, offset is not as perceptually important as onset and is not exploited in our current model.

3.2.2 Existing Models of Onset Detection

Most onset detectors are based on the first-order time difference of the amplitude envelops (Bilmes 1993; Goto and Muraoka 1996; Scheirer 1998), whereby the maximum of the rising slope of the amplitude envelops is taken as onset. According to Moore (1995), the smallest detectable change in intensity is approximately proportional to the intensity of the signal. In light of this fact, Klapuri (1999) proposed to detect onset based on the relative difference function instead of the absolute difference function. Recently, a neural model has been proposed by Fishbach and Yeshurun (2001), which can account for numerous physiological and psychoacoustic phenomena. In this model, the first-order time derivative of the amplitude envelope is viewed as an analog to the visual brightness gradient, so that auditory onset can be detected in a way similar to visual edge detection.

3.2.3 Onset Detection

The onset detection model utilized in our system is adapted from the neural model described in Fishbach et al. 2001. First, the auditory periphery response in each frequency band is progressively delayed by an array of neurons with ascending membrane time constants. The kernel of each neuron is characterized by an $\alpha-$function

$$k(n) = \frac{1}{\tau^2} n e^{-n/\tau} \tag{3.3}$$

where τ is the time constant. Then, the first-order time derivative of the amplitude envelope is calculated by differentiating the stimulus along the delay line. This operation is approximated by connecting the outputs of the delay layer to a single onset neuron with excitatory and inhibitory connections. The combination of associated weights forms a first-order derivative of Gaussian function.

In the current model, the delay layer consists of only two neurons. The onset neuron receives the two delayed stimuli, one excitatory and the other inhibitory. The excitatory input has a shorter time constant than the inhibitory input. The overall model can be simply described as

$$h(n) = k_1(n) - k_2(n) \tag{3.4}$$

where $k_1(n)$ and $k_2(n)$ are the kernel functions of the two delay neurons, given by

$$k_1(n) = \frac{1}{\tau_1^2} n e^{-n/\tau_1}, \quad k_2(n) = \frac{1}{\tau_2^2} n e^{-n/\tau_2}, \quad \tau_1 < \tau_2 \tag{3.5}$$

The time constants τ_1 and τ_2 are selected to be 6 ms and 15 ms respectively in order to obtain a bandpass filter $H(z)$. The passband of $H(z)$ covers from 4 to 32 Hz. These frequencies are within the most important range for speech perception for the human auditory system (Drullman et al. 1994a; Drullman et al. 1994b). Figure 3.4 shows the temporal and frequency responses of delay neurons and the whole onset model respectively.

Figure 3.5(a) shows the onset maps for a clean speech sentence. The clean speech says "Don't ask me to carry an oily rag like that". The waveform of the clean speech is plotted in the upper panel. The onset map produced by the onset neurons is plotted in the lower panel. As visible from the result, the onset model can detect the onset of every phoneme, in spite of the intensity and duration of the phoneme. Generally, vowel sounds have much higher intensity and take more time to build

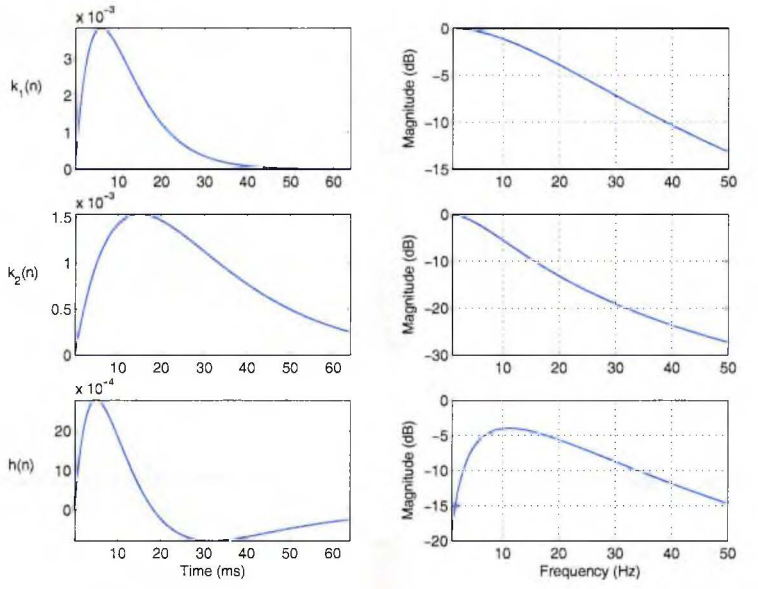


Figure 3.4: Time and frequency responses of two delay neurons (top and centre) and overall onset model (bottom)

up than stop consonants. Compatible with this fact, the detected onset of a vowel sound is comparatively stronger and lasts for longer time. The onset model was also verified under different noise conditions. Figure 3.5(b) shows the onset map for speech corrupted by white noise. The same speech signal was used. The SNR of the mixture is 5 dB. Due to the masking effect of white noise, the onsets of the target speech are largely reduced or completely missing. Only those strong onsets are preserved. Meanwhile, the fluctuations in the amplitude of white noise are sometimes misidentified as false onset detection. Especially when the background noise is turned on at the beginning of the utterance, a strong onset across the frequency bands is produced. Figure 3.5(c) shows the onset map for the same speech mixed with a competing speech signal. The onsets of both speech streams are detected. Therefore, synchronous onsets across frequency bands provide evidence to partition the set of simultaneous spectral components into an active stream, but onset information cannot be used to distinguish the alternative streams. Figure 3.5(d) shows the onset map

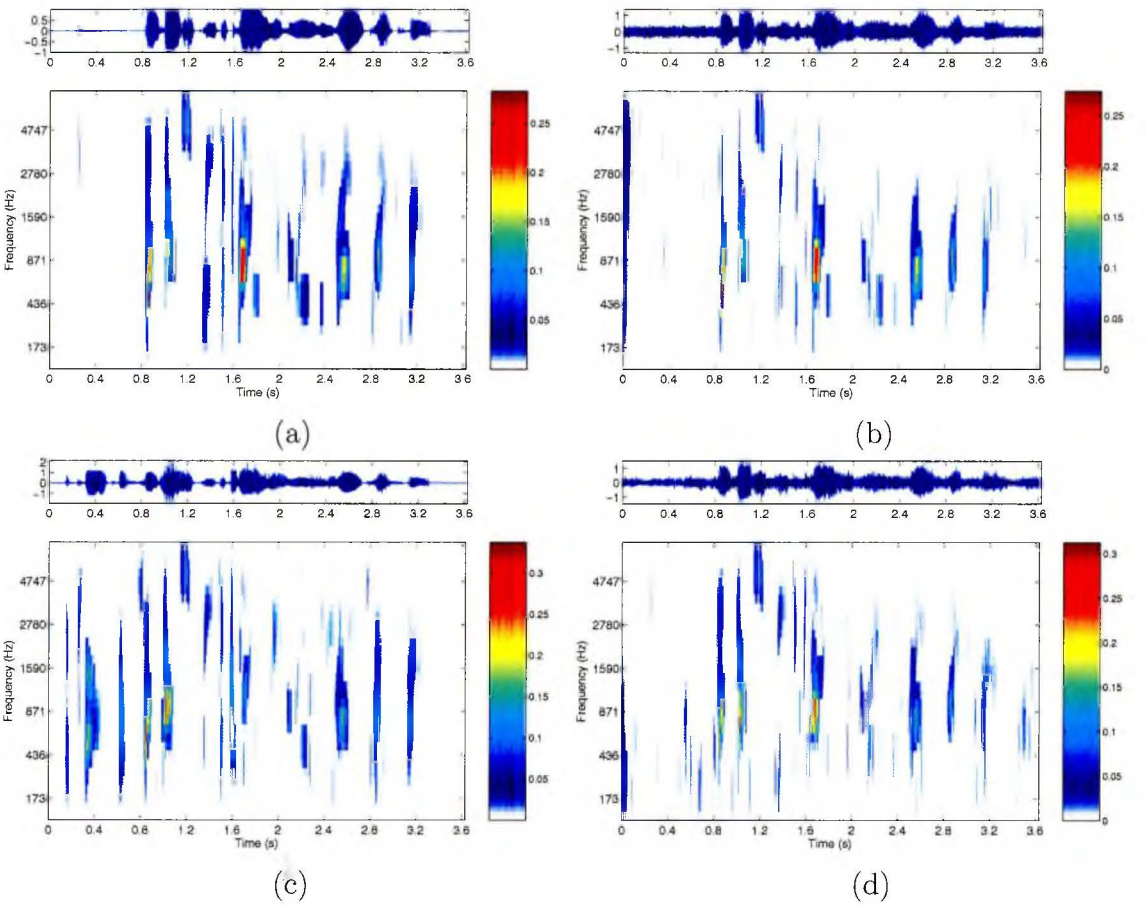


Figure 3.5: Onset maps for a speech "Don't ask me to carry an oily rag like that." under different noise conditions: clean speech (a); corrupted by white noise (b); mixed with competing speech (c); corrupted by babble noise(d). SNR of the noisy speech is 5 dB.

for the same speech mixed with multi-speaker babble noise. Again, the SNR of the mixture is 5 dB. Since the energy of babble noise is mostly distributed over low-frequency bands and the envelope variations of babble noise can be much stronger than white noise, the masking effect reflected from the onset map is pronounced for lower frequencies. But for higher frequencies, the onset map remains undistorted.

3.3 Binaural Spatial Cues

3.3.1 Psychoacoustic Evidence

The cues used in sound spatial localization may also help in the analysis of complex auditory inputs.

Sounds reaching the farther ear are delayed in time and are less intense than at the nearer ear. There are thus two possible spatial cues: interaural time difference (ITD) and interaural intensity difference (IID). Owing to the physical nature of sounds, ITDs and IIDs are not equally effective at all frequencies.

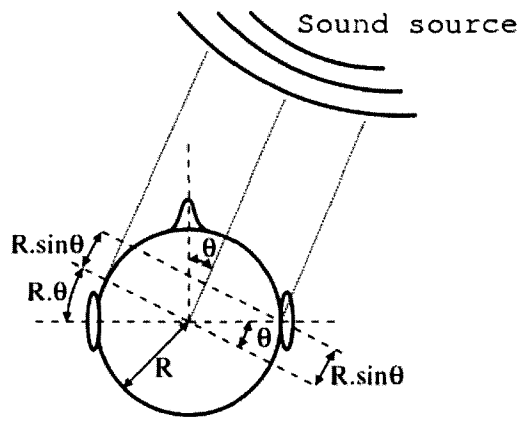


Figure 3.6: Illustration of the method for calculating the difference in arrival time at the two ears.

ITD can be calculated from the path difference between the two ears, as illustrated in Figure 3.6. When the sound source is at an incidence angle θ and the head radius is R, the difference in path length between the two ears is given by $\triangle D$, which follows the simple law

$$\triangle D = R\theta + Rsin(\theta) \tag{3.6}$$

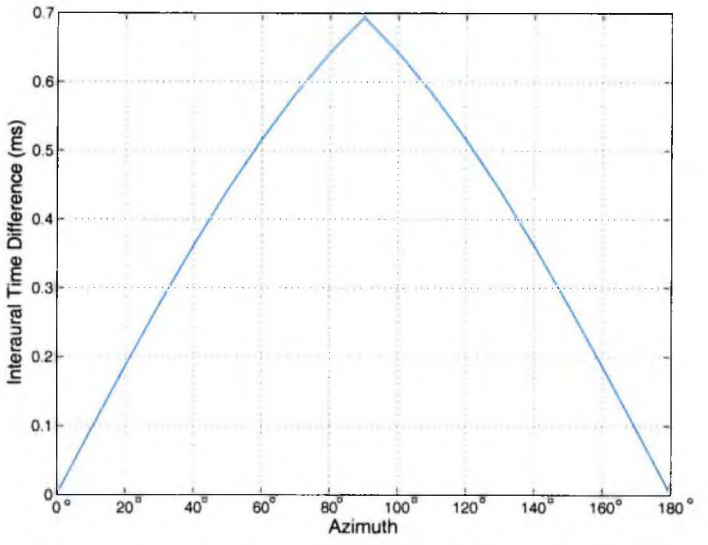


Figure 3.7: Interaural Time Differences (ITDs) as a function of azimuth. Radius of head R is 9 cm.

Knowing that the speed of sound is about 343 meters per second, we can calculate the time difference for the sound reaching opposite ears. Assuming 9 cm head radius, the ITD is plotted as function of azimuth θ in Figure 3.7. ITDs range from 0 (for a sound straight ahead) to about 690μs for a sound at 90° azimuth (directly opposite one ear). For low-frequency sounds with a period larger than twice the maximum possible ITD, the ITD provides effective and unambiguous information about the location of the sound. However, sounds at higher frequencies present ambiguity in determining the correct time delay, which may be greater than the signal period.

The IID arises from the "shadow" effect cast by the head. Low-frequency sounds have wavelengths which are long compared with the size of the head, and thus, these sounds "bend" very well around the head. Therefore, low-frequency sounds are barely affected by IID. For sound sources that are distant from the listener, IID is negligible below 500 Hz, but may be as large as 20dB at high-frequencies. For sound sources that are very close to the head of the listener, IID can occur even at low frequen-

cies (Brungart and Rabinowitz 1999). Although it is hard to derive a mathematical formula for calculation of IID, we can still empirically measure the relationship between IID and sound incidence azimuth. The method and results will be described in Section 3.3.4.

In summary, the cue of ITD is more reliable at low frequencies, while the cue of IID is more useful at high frequencies. The combination of these two cues has come to be known as the "duplex" theory of localization (Moore 2003).

3.3.2 Existing Models of Binaural Spatial Estimation

A number of binaural models have been developed over the last half century (Jeffress 1948; Lindemann 1986; Gaik 1993). Most of them could be considered a variant of the Jeffress's (1948) neural coincidence mechanism to detect interaural time difference. Essentially, these models have a generic structure including a series of peripheral auditory processing, comparison of interaural timing information using a correlation or coincidence mechanism, computing interaural intensity differences at the outputs of monaural processors, and a subsequent decision-making mechanism. Models of binaural spatial processing are already built into cocktail-party processor and demonstrate particular effectiveness in source separation when the sound sources are spatially separated (Lyon 1983; Bodden 1995; Grabke and Blauert 1998; Roman et al. 2003).

3.3.3 Interaural Time Difference Estimation

In the current model, ITD is determined on the basis of cross-correlation between hair cell channel outputs at opposite ears. This processing is essentially very similar to

the auto-correlation mechanism involved in pitch analysis. Specifically, the interaural cross-correlation function (CCF) is computed as follows

$$\mathrm{CCF}(i,j,\tau) = \frac{\sum_{k=0}^{K-1} l_i(j-k)r_i(j-k-\tau)}{\sqrt{\sum_{k=0}^{K-1} l_i^2(j-k)\sum_{k=0}^{K-1} r_i^2(j-k-\tau)}} \tag{3.7}$$

where $\mathrm{CCF}(i,j,\tau)$ is the cross-correlation at lag τ for the ith frequency channel at jth time instance; l and r are the auditory periphery outputs at the left and right ear; K is the integration window length and k is the index inside the window. As in the definition of ACF, CCF is also normalized by local channel energy estimated over the integration window. This normalization can equalize the contribution from different channels. Again, all the minus signs in Equation 3.7 ensure this implementation to be causal.

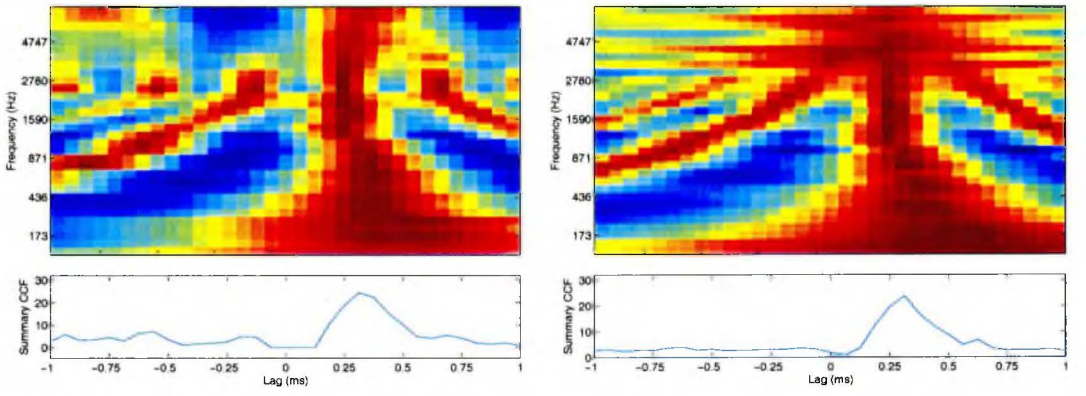
The cross-correlation defined in Equation 3.7 is identical to the auto-correlation in Equation 3.1 used for pitch estimation. The correlation theorem discussed in Equation 3.2 also applies to the computation of cross-correlation. Hence, the normalized CCF in Equation 3.7 can be implemented more efficiently as follows: FFT the two data sets, $l_i(n)$,$n = j-K+1, j-K+2, \ldots, j$ and $r_i(n)$,$n = j-K+1, j-K+2, \ldots, j$, multiply one resulting transform by the complex conjugate of the other, and inverse transform the product. The normalization term is determined by the auto-correlation of l_i at zero lag and the auto-correlation of r_i at zero lag. Note that these two values are already computed in ACF. Repeated computation can be avoided.

As the correlogram in pitch analysis, the CCFs are visually displayed in a two dimensional (centre frequency $\times$ lag) representation, termed the cross-correlogram. The cross-correlogram and correlogram are updated synchronously. For the sake of simplicity, the frame rate and window size are selected exactly the same as the correlogram computation in pitch perception, i.e., 100 frames per second and 20 ms rectangular window. As a result, the FFT values can be reused in both the pitch

model and the binaural model. Lag τ is limited to the range $-1 < \tau < 1$ ms, which includes the range of ITDs encountered in natural listening conditions, as depicted in Figure 3.7.

For a signal without any interaural time disparity, the CCF reaches its maximum at zero lag. In this case, the cross-correlogram is a symmetrical pattern with a vertical stripe in the centre. As the sound moves laterally, the interaural time difference results in a shift of CCF along the lag axis. Hence, for each frequency channel, ITD is computed as the lag corresponding to the position of the maximum in the CCF. Finally, in the same way as the processing in pitch model, all the local maxima as well as their two immediate neighbors in each CCF are picked out and integrated across frequency to produce a summary CCF. The global peak in the final presentation indicates the perceived ITD.

Binaural acoustic signals are required to test the binaural model. The details about how the binaural data are generated will be described in Chapter 5. As an example, the cross-correlograms of a synthesized vowel /a/ and white Gaussian noise signal are plotted in Figure 3.8. Both of the two signals have 30° incidence azimuth.



(a) (b)

Figure 3.8: CCFs and summary CCFs computed for (a) synthesized vowel /a/ with 30° incidence azimuth, (b) white Gaussian noise with 30° incidence azimuth.

At low-frequency narrow-band channels, we see that the CCF is nearly periodic in lag, with a period equal to the reciprocal of the centre frequency. By limiting ITD to the range $-1 < \tau < 1$ ms, the repeated peaks at lags outside this range can be largely eliminated. But for channels with centre frequency within 500~3000 Hz, it is still probable to have multiple peaks fall inside this range. This quasi-periodicity of cross-correlation makes an accurate estimation of ITD a difficult task. At the output of channels with centre frequency higher than 3000 Hz, the fine structure information is removed as a result of low-pass filtering in the inner hair cell transduction model. Only the temporal envelope is reserved. Cross-correlation analysis in these channels gives an estimation of the interaural envelope difference (IED) instead of ITD. In comparison to the CCFs of a vowel, the CCFs of white Gaussian noise contain a lot of disturbance in the high frequency channels due to the aperiodic nature of the input signal.

In both of the two summary cross-correlograms, the maximum is found at a lag of 0.3 ms. According to the ITD-azimuth mapping plotted in Figure 3.7, this ITD gives an accurate prediction of azimuth angle. It is clear that the cross-correlation model works well for either periodic or aperiodic sounds.

To demonstrate the spatial grouping using ITD information, we further tested the cross-correlation model on multiple source scenarios. Figure 3.9 gives the results for two cases. For the left panel, the input signal is a mixture of two concurrent vowels. The target vowel /a/ originating from 0° azimuth mainly dominates in the high frequencies; the interference is a vowel /u/ originating from 30° azimuth, whose energy is mainly distributed in the low frequencies. These two signals are mixed to produce 5 dB SNR. As can be seen from the cross-correlogram in Figure 3.9(a), the high-frequency components show a common peak at zero lag. Meanwhile, some of the low-frequency components show another common peak at a lag of 0.3 ms.

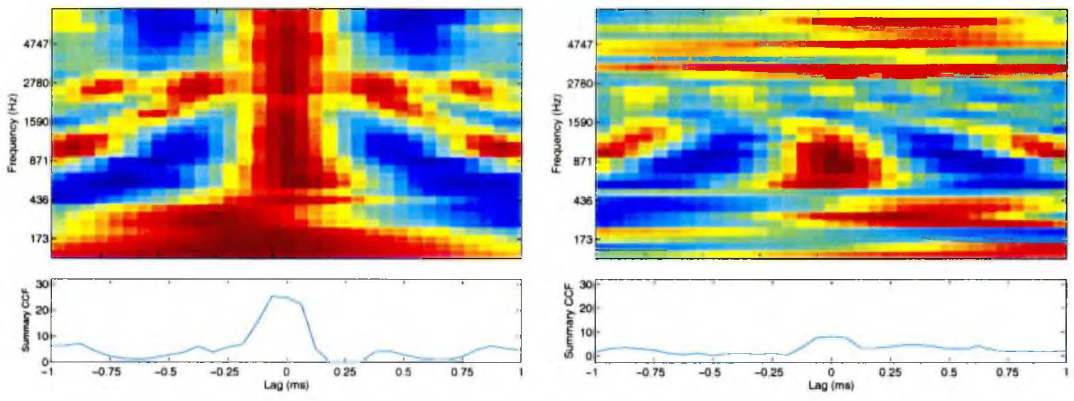


Figure 3.9: CCFs and summary CCFs computed for noisy speech. The target signal is a synthesized vowel /a/ with 0° incidence azimuth. (a): The masker is a synthesized vowel /u/ with −30° incidence azimuth. (b): The masker consists of four streams of babble noises originating from 40°, −30°, 60° and −60° azimuth. The overall SNR is 5 dB in both cases.

Therefore, the two groups of frequency components can be clustered, based on the ITD estimation. In the summary cross-correlogram, the ITD of target signal results in a dominant peak around zero lag. In the second case, the same target is used. The interference is much more complicated. It is a combination of four streams of babble noises originating from 40°, −30°, 60° and −60° azimuth respectively. Again, the overall SNR is 5 dB. In this case, the cross-corrrelogram is greatly disturbed by the spatially distributed interference signal. Only the centre frequency channels (500∼1500 Hz) show common ITD peaks at zero lag. The ITD peaks in the other frequencies are rather random due to the interaction of multiple streams. As a result, the maximum peak in the summary CCF is not as salient as the two-sources case.

3.3.4 Interaural Intensity Difference

IID is defined as the log ratio of the local energy at the opposite ears . For the ith frequency channel and jth time instance, IID can be computed as in Equation 3.8, that is,

$$\mathrm{IID}(i,j) = 10\log_{10}\left(\frac{\sum_{k=0}^{K-1} r_i^2(j-k)}{\sum_{k=0}^{K-1} l_i^2(j-k)}\right) \tag{3.8}$$

where l and r are the auditory periphery outputs at the left and right ear respectively; K is the integration window size, and k is the index inside the window. Again, the frame rate and window size utilized in IID measurement are selected to be exactly the same as in the correlogram computation for pitch perception, i.e., 100 frames per second and 20 ms rectangular window.

As we already mentioned, no simple mathematical formula can describe the relationship between IID and azimuth. However, given a complete binaural sound database, we can empirically evaluate the IID-azimuth mapping. Figure 3.10 is a graphical representation of the IID-azimuth mapping measured from our own data. Note that IID is a frequency dependent value. Therefore, frequency is considered as a variable in the IID-azimuth mapping. From this graphical representation of the mapping function, we can see that sound received at the ear on the farther side is generally less intense.

Now, given an incoming signal, we can calculate the IIDs for each frame of data and then convert them to azimuth based on the IID-Azimuth mapping. Again the performance of IID-based localization model was verified for both single-source and multi-source scenarios. For a single source scenario, two types of signals were tested: a synthesized vowel /a/ with 30° azimuth and a white Gaussian noise presented at $-20°$ azimuth. We find that, for single fixed source, the IID values remain constant along the time axis. The estimation results for one frame are given in the Figure 3.11. The

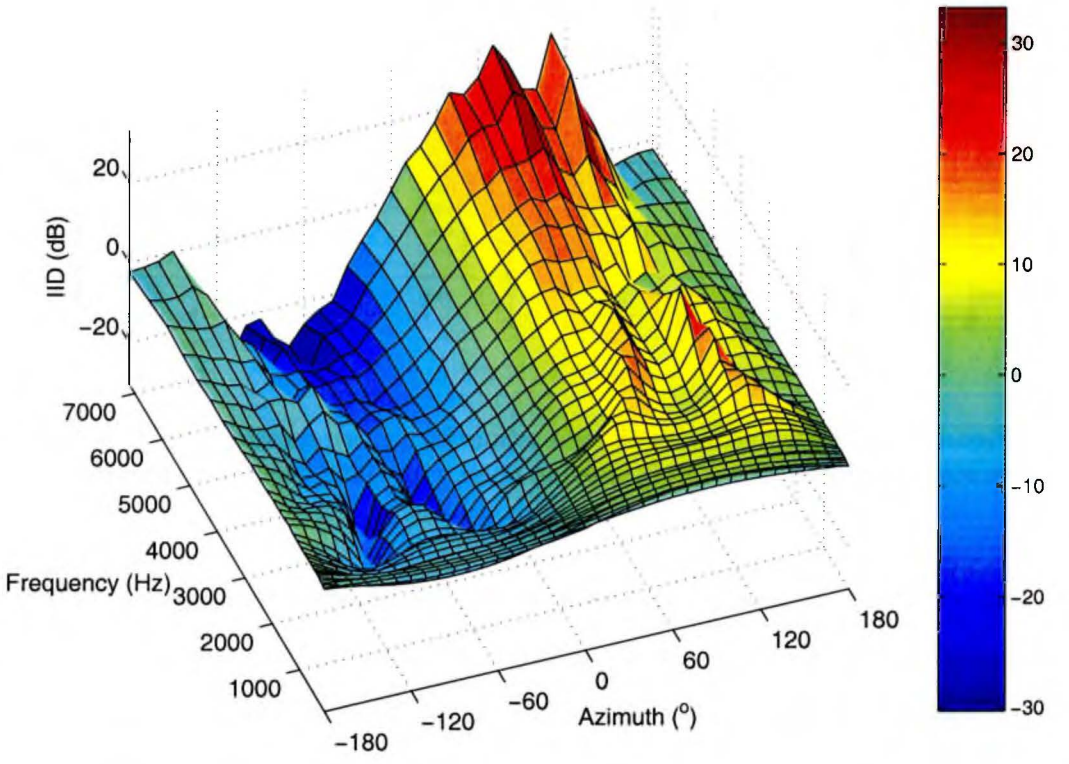
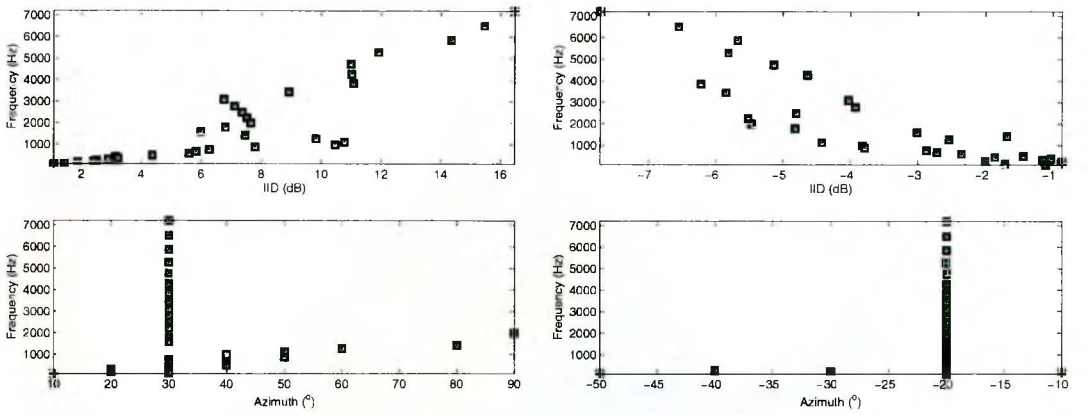


Figure 3.10: Interaural Intensity Differences (IIDs) as a function of azimuth and frequency.

upper panel clearly shows that IID greatly varies with frequency. Once converted to azimuth, a consistent localization estimation across frequencies is obtained, especially in high-frequency channels. In low-frequency channels, as we expect, the IID-based azimuth estimation is not so reliable.

The model was also tested on multi-source scenarios to demonstrate the spatial grouping using IID information. We used exactly the same data as that tested for ITD model. In the first case, the input signal is a mixture of two concurrent vowels. The target vowel /a/ originating from 0° incidence azimuth mainly dominates in the high-frequency range. The interference is a vowel /u/ originating from 30° azimuth,



(a) (b)

Figure 3.11: IIDs and azimuth estimation for (a) synthesized vowel /a/ with 30° incidence azimuth, (b) white Gaussian noise with −20° incidence azimuth.

whose energy is mainly distributed in the low-frequency components. These two signals were mixed to produce 5 dB SNR. In the second case, the same target signal is masked by multi-speaker babble noise. Again, The target signal is originating from 0°. The babble noise is a composition of four streams originating from 40°, −30°, 60°, −60°. The overall SNR remains 5 dB. Figure 3.12 gives the IID and azimuth estimation results for these two cases. In the double-vowel case, the two vowels in the mixture are both synthesized with stationary intensity, common onset and offset. Therefore the IID values do not vary with time. Figure 3.12(a) displays typical result from one frame of data. Overall, despite its unreliable estimate at low-frequency channels, IID gives a very accurate estimate of the target source location, especially at high frequencies. But for the multi-source babble noise, the result is quite random as can be seen from Figure 3.12(b). This result can be explained by noting that IID is obtained by comparison of the overall intensity received at left and right ears. The intensity difference caused by one stream can be cancelled out by the difference caused by another stream originating from the opposite side. Hence, in the presence of multiple spatial-separated streams, IID information becomes unreliable.

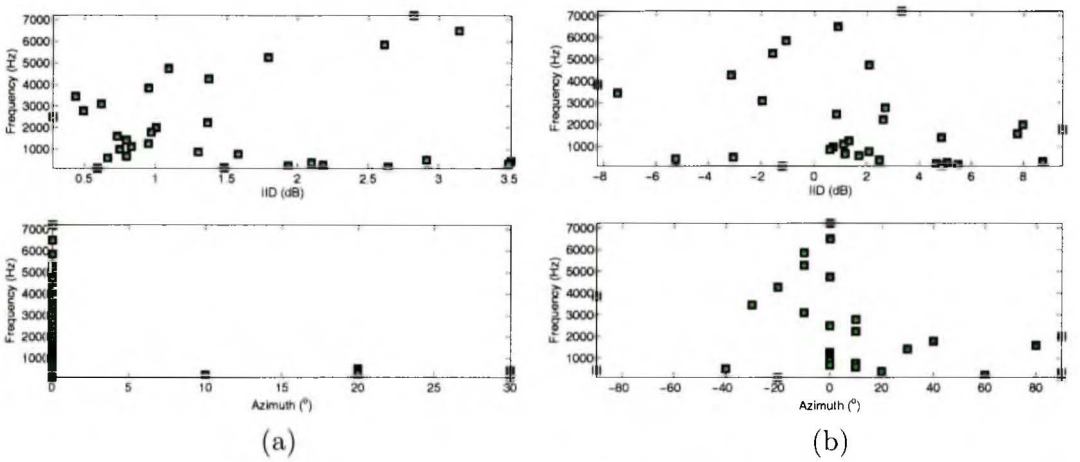


Figure 3.12: IIDs and azimuth estimation for noisy speech. The target signal is a synthesized vowel /a/ with 0° incidence azimuth. (a): The masker is a synthesized vowel /u/ with −30° incidence azimuth. (b): The masker consists of four streams of babble noises originating from 40°, −30°, 60° and −60° azimuth. The overall SNR is 5 dB in both cases.

3.4 Effect of Reverberation

Most everyday listening situations not only consist of multiple sources of sound, but also consist of multiple paths (reflections) that the sounds can take to reach the listener. When the listener wishes to attend to sounds from a particular source and ignore sounds from other sources in a reverberant condition, the acoustic reflections of both signal and masker(s) complicate the listening task even more and adversely affect the signal reception. Recently, some psychoacoustic experiments have been conducted on the effects of reverberation on multi-talker communication (Culling et al. 2002; Culling et al. 2003; Darwin and Hukin 2000). All these works imply reverberation has a variety of destructive influences on listeners's ability to cope with multiple concurrent voices. In this section, we will investigate the effect of reverberation on the acoustic cues.

3.4.1 Pitch

Figures 3.13 (a) and (b) compares the spectrograms of the sentence "The candy shop was empty." spoken in anechoic (upper panel) and reverberant (lower panel) conditions, respectively. Where the F0 contour is relatively flat (e.g., vowel /æ/ in the word "candy", denoted as area "A" in the spectrogram), the harmonic structure is still evident and little affected by reverberation. However, the harmonic structure, where there is fast variation in F0 (e.g., vowel /o/ in the word "shop", denoted as area "B" in the spectrogram), is visibly damaged by reverberation, because each part of the reverberant sound is delayed and superimposed on the following sound.

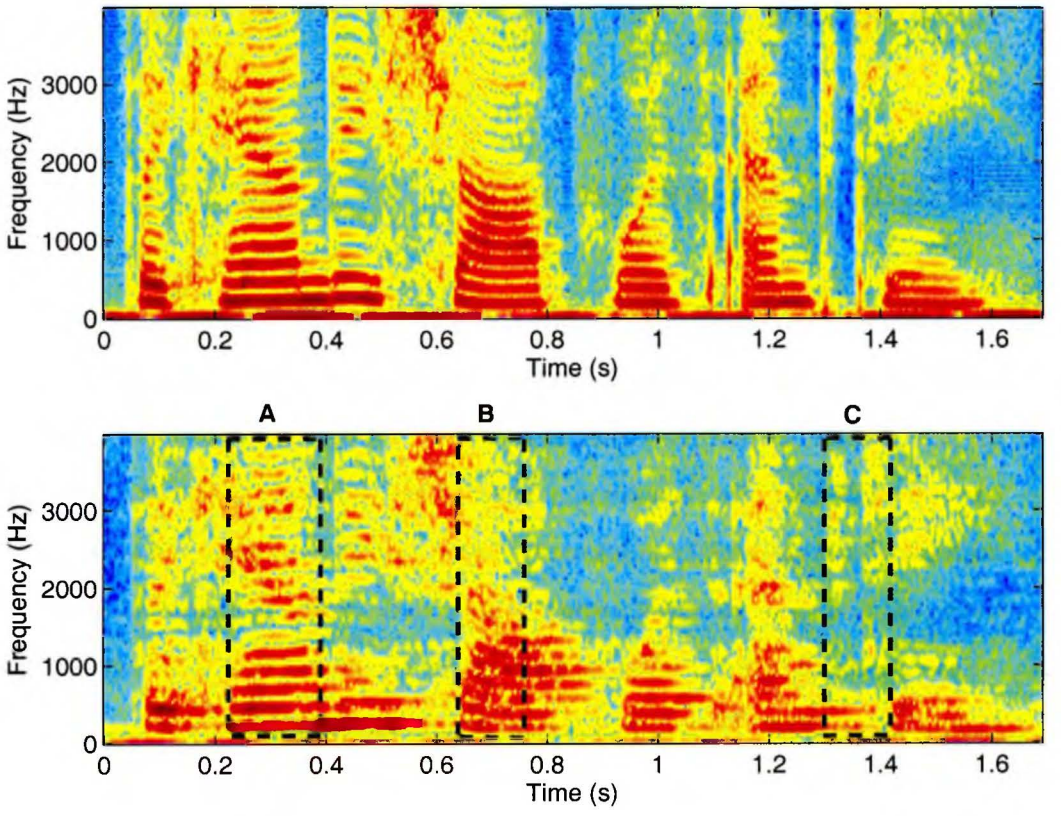


Figure 3.13: Spectrogram of sentence "The candy shop was empty" in anechoic (upper panel) and reverberant condition (lower panel).

This effect is confirmed in the results of autocorrelation pitch analysis. Figure 3.14 shows the correlograms of two reverberant vowel /a/. Both of them have pitch centred at 200 Hz. They are manipulated to have different vibrato, which refers to the pitch modulation. The stimuli used for the right panel has 30 Hz vibrato, which is 3 times faster than the vibrato of the stimuli used for the left panel. For the vowel with relatively stead F0, the correlogram shows that its harmonic structure remains almost intact under reverberation condition. While, for the vowel with a fast varying F0, reverberation makes its harmonic structure completely missing.

Therefore, in the reverberant conditions, the F0-difference between target and masker would no longer benefit the segregation, if the target and masker have fast varying pitch. Similar findings have been reported in Culling et al. 1994.

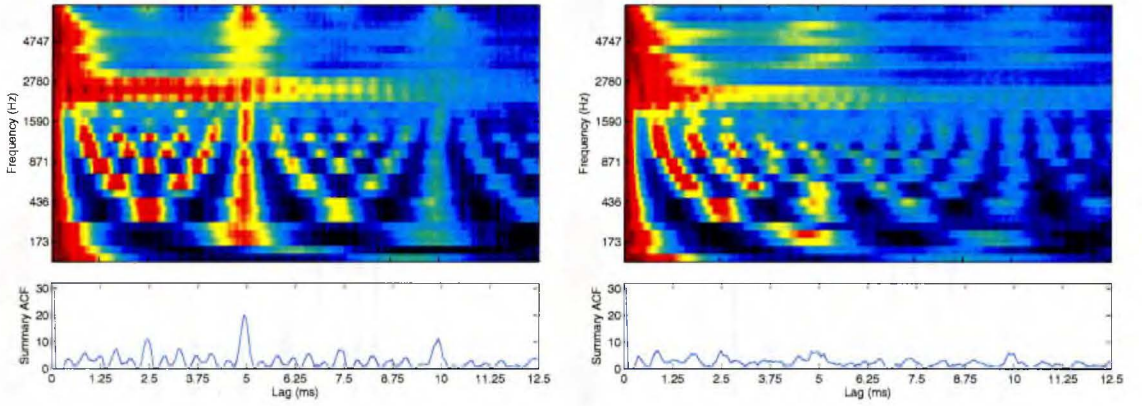


Figure 3.14: ACFs and SACFs computed for reverberant sounds: (a) vowel /a/ with its pitch centred at 200 Hz and a 10 Hz vibrato; (b) vowel /a/ with its pitch centred at 200 Hz and a 30 Hz vibrato.

3.4.2 Onset

Reverberant energy of the proceeding sound can fill the spectral-temporal dips, hence make the onset transients less distinctive. For example, the onsets of two stop

consonants /p/ and /t/ in word "empty" (area C in Figure 3.13) are blurred by the lagging sound of previous phone /m/.

Generally, the signal from the direct path arrives first with relatively strong intensity. It is followed by secondary and multiple subsequent reflections with rapidly decreasing intensity. Overall, onsets are more resistant to reverberation compared with the other cues.

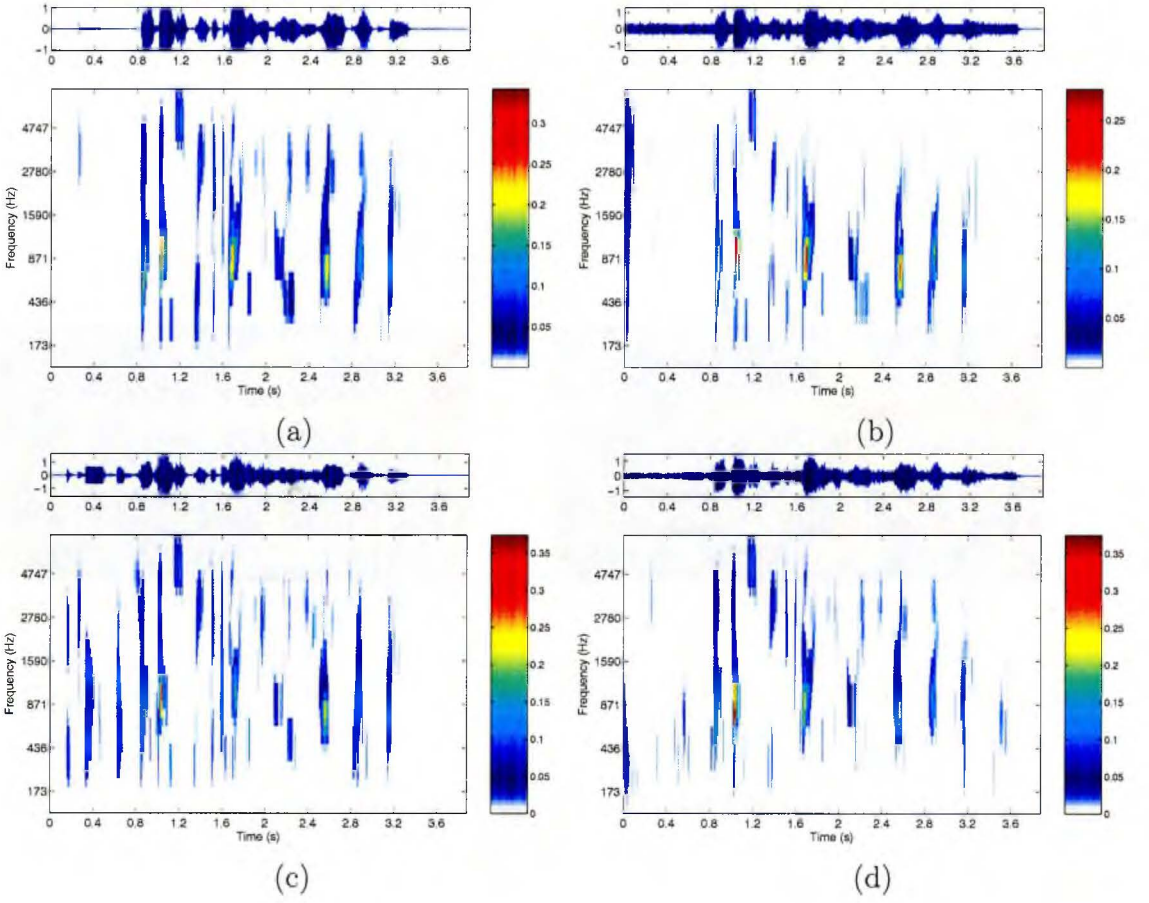


Figure 3.15: Onset maps for a reverberant speech "Don't ask me to carry an oily rag like that." under different reverberant noise conditions: clean speech (a); corrupted by white noise (b); mixed with competing speech (c); corrupted by babble noise(d). SNR of the noisy speech is 5 dB.

Figure 3.15 shows some examples of onset maps of reverberant speech, in which the stimuli are exactly the reverberant counterpart of the stimuli used in Figure 3.5.

Comparing the onset maps of each reverberant-anechoic pair, we come to the same conclusions: although onsets are occasionally masked by the reflection of proceeding sounds, reverberation does not have a substantial impact on onset detection.

3.4.3 Spatial cues

Reverberation introduces potentially an infinite number of sources due to the reflections against the surrounding surfaces. Therefore, the beneficial effect of spatial separation between the target and interfering sources is expected to be largely abolished in the presence of reverberation.

Studies of directional localization in rooms generally show that the localization ability of human auditory system is reduced for a short time following the onset of the leading sound (known as the "precedence effect"). Since the spatial information about the echoes (the lagging sound) is partially suppressed, single sounds with abrupt onsets can be well localized in naturally reverberant environments. Nevertheless, localization of sounds that lack abrupt onsets is seriously impaired by reverberation, because of distortion to both interaural time and intensity differences. As a consequence, reverberation can adversely affect signal separation based purely on directional information. Even modest amounts of reverberation, which do not reduce listener's ability to localize speech presented alone, can reduce listener's ability to exploit localization cues in identifying the target sound presented with spatially separated masking noise.

Compatible to the psychoacoustic results, our simulation results also show that reverberation physically distorts steady-state "directional" cues like ITD and IID. The two localization cues were examined both at the onset and during the steady-state position of the target stimulus. Figure 3.16(a1) shows the cross-correlogram and

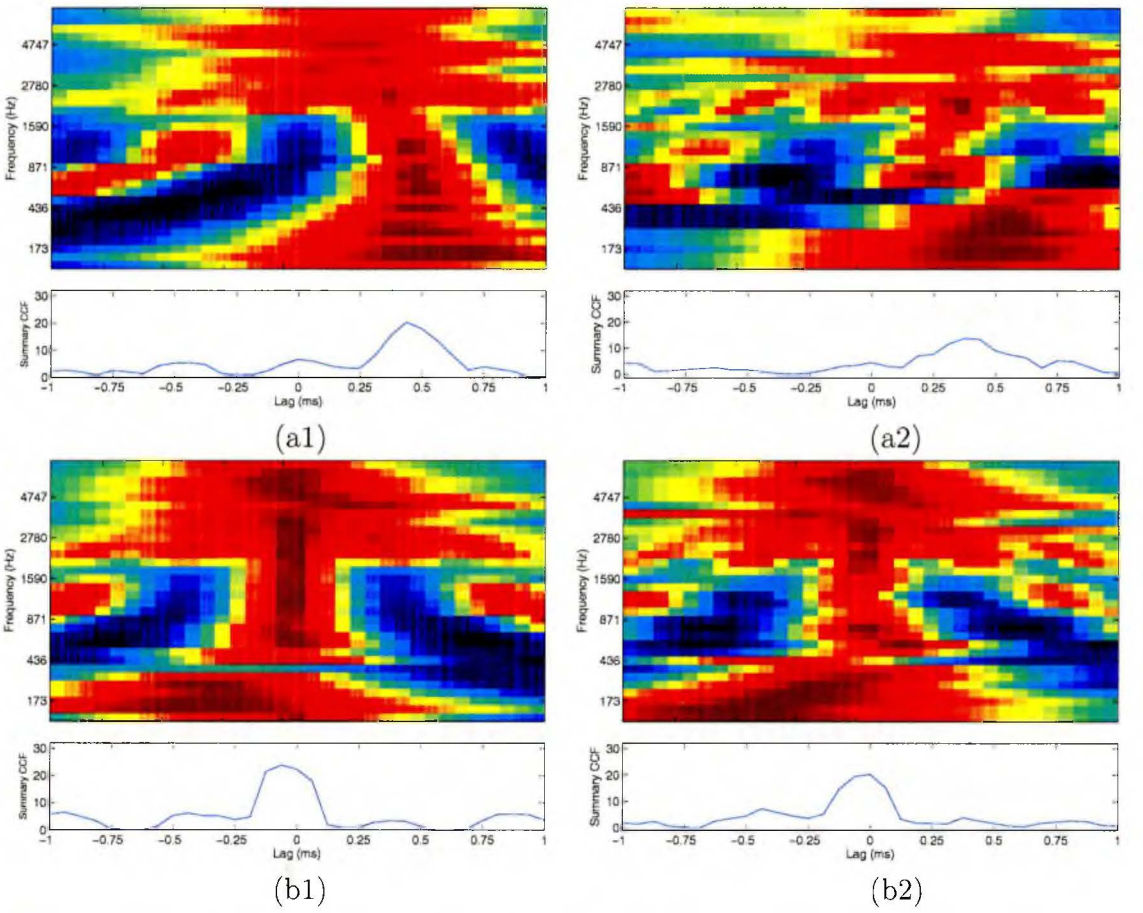


Figure 3.16: CCFs and summary CCFs computed for reverberant sounds: the 1st frame (a1) and the 5th frame (a2) of vowel /a/ with 45° incidence azimuth; the 1st frame (b1) and the 5th frame (b2) of vowel /a/ with 0° incidence azimuth mixed with vowel /u/ with −45° incidence azimuth (SNR of the mixture is 5 dB).

summary CCF at the onset of the reverberant vowel /a/. The stimulus is incident from 45° azimuth. For a reverberant vowel, the common amplitude modulation at unresolved harmonics is inevitably distorted by its own reflection. Therefore, the CCFs of reverberant vowel display much broader and non-periodic peaks in the high-frequency range in comparison to the CCFs of an anechoic vowel plotted in Figure 3.8(a). However, the CCFs of most of the low-frequency channels generally show peaks centring around 0.4 ms. Consequently, a prominent ITD of approximately 0.4 ms is shown in the summary CCF. Figure 3.16(a2) shows the cross-correlogram

and summary CCF at the 5th frame (50 ms after the start of stimulus) of the same stimulus. At this time frame, the stimulus is already in its steady state. The ITD is largely distorted by the echo of proceeding sounds. The random variation in CCF across frequencies makes the peak of summary CCF less prominent and makes the localization estimation less accurate. In Figure 3.16(b1) and (b2), a competing vowel is presented as a masker in reverberant condition. At the onset of stimulus, the correlogram (see panel b1) shows two separated groups of ITD: one group of peaks around 0 ms associated with the target from 0° and the other group of peaks around -0.4 ms associated with the masker from $-45°$. As the sound continues, the ITD grouping becomes less evident (see panel b2).

Similarly, Figure 3.17 examines the IID azimuth estimation for the same set of stimuli. Note that the resolution of IID to azimuth mapping in this implementation is 10°. Due to the quantization error, for a stimulus originating from 45°, azimuth estimation of either 40° or 50° is acceptable. The simulation results show that the azimuth estimation obtained at the onset of stimulus is basically more accurate and more consistent across frequency channels than the estimation obtained at the steady state of sound.

In summary, spatial localization and pitch, the most effective cues to help listener to maintain attention to a particular sound source, are susceptible to the degrading effects of reverberation, especially during steady state portions of stimuli. Meanwhile, the onset cue is more resistant to the effects of reverberation than the other cues. In this sense, the conclusion obtained from our model analysis is compatible with reported psychoacoustic experimental results.

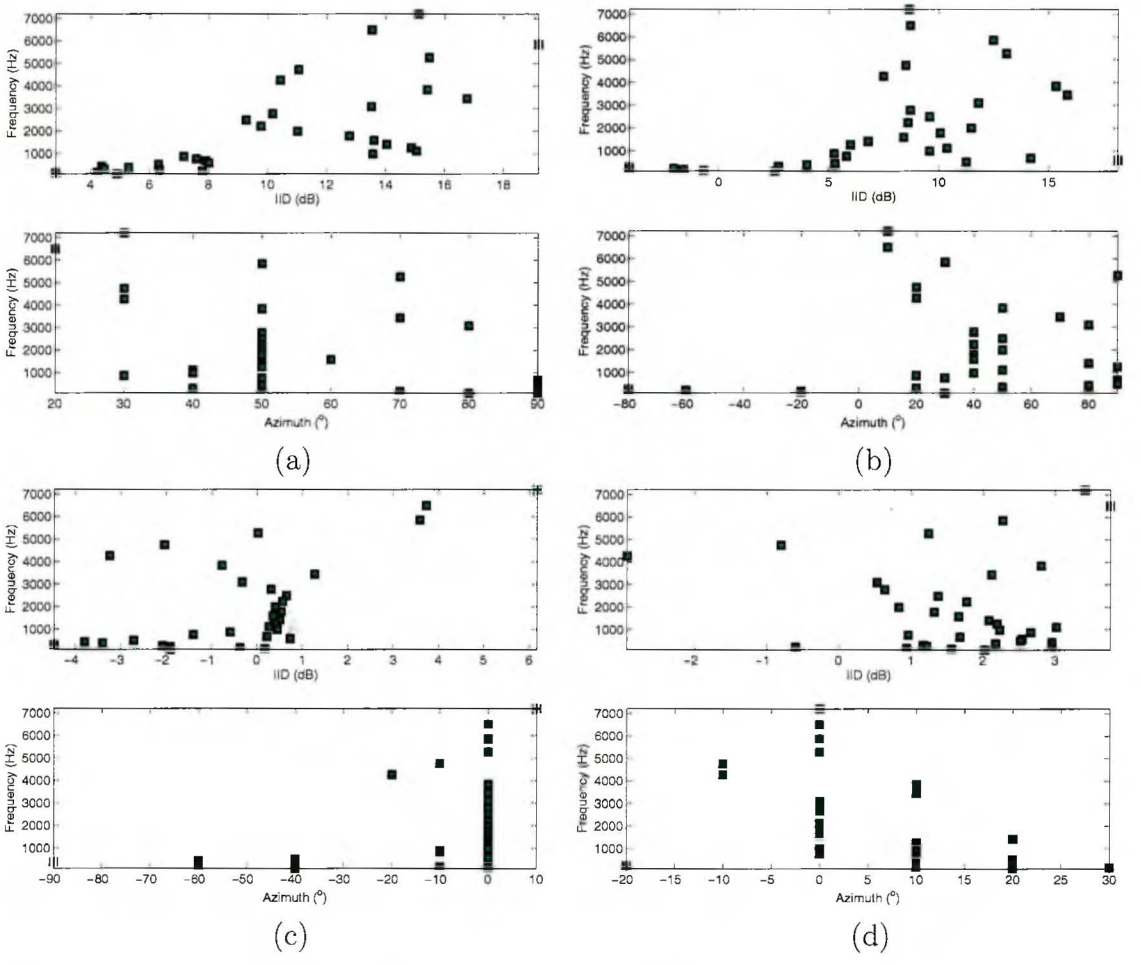


Figure 3.17: IIDs and azimuth estimation for reverberant sounds: the 1st frame (a) and the 5th frame (b) of vowel /a/ with 45° incidence azimuth; the 1st frame (c) and the 5th frame (d) of vowel /a/ with 0° incidence azimuth mixed with vowel /u/ with −45° incidence azimuth (SNR of the mixture is 5 dB).

Chapter 4

Grouping and Segregation

In this chapter, we will further analyze the strength and weakness of each of the cues. Then we will address how to integrate the information conveyed by multiple cues and how to make a grouping decision.

4.1 Motivation for Cue Fusion

Each of the cues we identified for auditory scene analysis has its own limitations, and it becomes unavailable or unreliable in certain scenarios.

Spatial Cues, as long as they can be exploited, have the advantage that they exist all the time, irrespective of whether the speech is voiced or not. There are however some problems in exploiting the spatial cues:

1. The two spatial cues utilized in the current model are unable to separate the sound sources localized in the same vertical (elevation) plane. Even in a horizontal plane, sound localization based on ITD and IID is not strictly accurate, especially for complex sounds. In our implementation, the azimuth resolution is 10° in the vicinity of the horizonal plane.

2. The interaction of multiple concurrent sound streams can produce a false localization estimate. As illustrated in Equation 4.1, l_1 and r_1 are the auditory periphery responses to a sound stream at the two ears. Here the frequency channel index is omitted for the sake of simplicity. Likewise, l_2 and r_2 are the auditory

periphery responses to another concurrent stream. As long as these two streams are uncorrelated with each other, the interaural cross-correlation of the mixed signal can be viewed as a linear combination of the cross-correlation of each individual stream. Therefore the maximum peak in CCF is determined by the ITD of the dominant stream:

$$\begin{aligned} & E\left[\left(l_1(n)+l_2(n)\right)\left(r_1(n-\tau)+r_2(n-\tau)\right)\right] \\ & \approx E\left[l_1(n)r_1(n-\tau)\right]+E\left[l_2(n)r_2(n-\tau)\right] \end{aligned} \tag{4.1}$$

However, the measurement of IID depends solely on the overall intensity of the sounds received at the opposite ears. Suppose, for example, there are the two uncorrelated, competing sound streams. One stream presents at 40° from the right side, whose auditory peripheral responses at opposite ears are denoted as $l_1(n)$ and $r_1(n)$; the other stream originates from $-40°$ azimuth from the left side with its auditory peripheral responses denoted as $l_2(n)$ and $r_2(n)$ respectively. So we have

$$\begin{aligned} & E\left[\left(r_1(n)\right)^2\right] > E\left[\left(l_1(n)\right)^2\right] \\ & E\left[\left(r_2(n)\right)^2\right] < E\left[\left(l_2(n)\right)^2\right] \end{aligned} \tag{4.2}$$

From Equation 4.3, we can see the intensity difference at the two ears is cancelled out and the resulting IID is approximately zero in this case. This could lead to a false perception that there is one stream presented along the centre direction.

$$10\log_{10}\left(\frac{E\left[\left(r_1(n)+r_2(n)\right)^2\right]}{E\left[\left(l_1(n)+l_2(n)\right)^2\right]}\right) \approx 10\log_{10}\left(\frac{E\left[\left(r_1(n)\right)^2\right]+E\left[\left(r_2(n)\right)^2\right]}{E\left[\left(l_1(n)\right)^2\right]+E\left[\left(l_2(n)\right)^2\right]}\right) \approx 0 \tag{4.3}$$

3. Reverberation is another big problem with spatial cues. Multipath echoes from room surfaces inevitably distort the direction information, which was demonstrated in Chapter 3.

Pitch information is particularly effective during the voiced sound segments. Unfortunately pitch detection can be very difficult in the presence of multiple sounds

streams. Similar to the cross-correlation computation, the auto-correlation of a mixed signal is approximately a linear combination of the two auto-correlations of each individual stream, as illustrated in Equation 4.4.

$$\begin{aligned} & E\left[(r_1(n)+r_2(n))\,(r_1(n-\tau)+r_2(n-\tau))\right] \\ & \approx E\left[r_1(n)r_1(n-\tau)\right]+E\left[r_2(n)r_2(n-\tau)\right] \end{aligned} \tag{4.4}$$

Here we give an example when ACF is not reliable as a cue for sound segregation. Suppose stream 1 is a voiced signal, while stream 2 is unvoiced speech or noise at the same time, and the energy of stream 2 is dominant in the local time-frequency (T-F) component. Its uncorrelated nature determines that $E\left[r_2(n)r_2(n-\tau)\right] \approx 0$. Thus the combined auto-correlation still shows a maximum peak at the pitch lag of stream 1, although the height of the peak is much lower than the peak at zero lag. As a result, this portion of T-F components will be falsely allocated to stream 1.

Onset has the advantage that it will work equally well with periodic and aperiodic sounds. However, when concurrent sounds are present, it is hard to know how to assign an onset to a particular sound source and the system could be prone to switch indiscriminately between emphasizing foreground and background objects. Even for a clean sound stream, it is difficult to distinguish genuine onsets from the gradual changes and amplitude modulations during the sound. Therefore, a reliable detection of sound onsets is very challenging.

4.2 Cue Fusion Algorithms

The above analysis shows that there is no single predominant cue, from which the grouping decision can be made. The fusion of information conveyed by multiple cues will certainly produce better performance.

On the other hand, the cues are not independent. For example, psychoacoustical experiments have reported that onset asynchrony can affect the pitch perception (Darwin and Ciocca 1992) and lateralization (Wood and Colburn 1992; Stellmack and Dye 1993). In some circumstances, different cues lead to conflicting decisions. They have to work in a competitive way in order to achieve a correct interpretation of a complex input.

For a computational system aiming to account for various cues as in the human auditory system, a strategy for cue-fusion must be incorporated to dynamically resolve the ambiguities of segregation based on multiple cues. The simplest solution to the fusion problem is called winner-take-all competition. This method requires an analysis to quantitatively identify the confidence on each of the cues. The detailed mathematical definitions of confidence values depend on the specific model used to extract the acoustic cue. When different cues are in conflict, the decision is made exclusively by the dominant cue. Woods (1996) applied weighted-sum mechanism to integrate the estimation arising from pitch and spatial cue. In Kashino et al. (1998), a Bayesian network framework is proposed for the application of music scene analysis, whereby multiple sources of information are integrated in a statistically optimal sense. Neural oscillator is an alternative model to explain the feature binding in auditory organization (Wang 1996; Brown and Cooke 1998). It assumes that the auditory object is represented by synchronously firing neurons in the cortex.

In the current model, we adopt a simple information fusion approach to solve the multi-cue fusion problem. The grouping principle is described as follow:

If all the perceptual cues suggest that a T-F component is dominated by a target signal, we accept this T-F component into the target stream. Otherwise, group it into the interference stream and suppress it.

The fusion rule is performed in a hierarchical process as illustrated in Figure 4.1. In the first stage, given the information from IID, we group the T-F components into two streams (target stream and interference stream). The grouping result is represented by a binary map whose value is one for a T-F unit where the target energy is greater than the interference energy and is zero otherwise. Likewise, ITD segregation can produce another binary map. These two binary maps are combined by the "AND" operation to obtain a spatial segregation map, which is further utilized to estimate the pitch of the target signal or the pitch of the interference. Similarly, a binary map can be produced according to the pitch segregation. If the target is detected as an unvoiced signal, onset cue is also incorporated to group the components into separate streams. At the last stage, all these binary maps are pooled together by the "AND" operation to arrive at the final segregation decision. In the next section, we will describe the detailed implementation of initial segregation by the four acoustic cues individually.

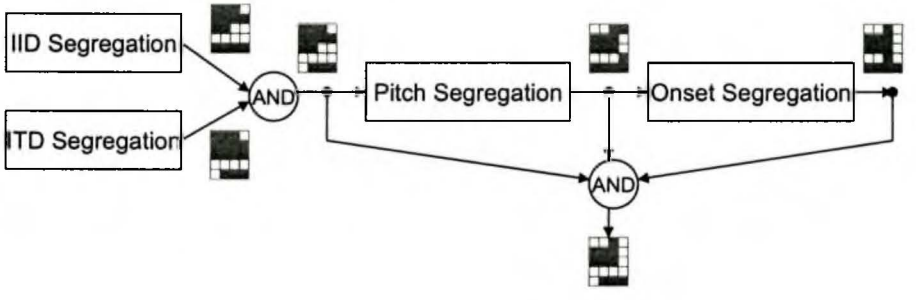


Figure 4.1: Flow chart of cue fusion process.

4.3 Implementation of Segregation Algorithm

1. **IID segregation** is very straightforward. Given the azimuth estimation using IID cue, we simply select those T-F units which have an azimuth within a range of [-10, 10] degrees as the target components and assign one to the corresponding values in the IID binary map.

2. **ITD segregation** consists of three steps.

- Step 1: at ith (i=1, 2, ..., 32) frequency channel, search for a global maximum value $MAXCCF_i$ in CCF.
- Step 2: compare the cross-correlation at zero lag $CCF_i(0)$ with the global maximum value $MAXCCF_i$.
- Step 3: group those T-F units satisfying the inequality $CCF_i(0) > 0.9 * MAXCCF_i$ into the target stream.

3. **Pitch segregation** is much more complicated. The flowchart of the complete pitch segregation process is illustrated in Figure 4.2. The ACFs peaks of very low frequency channels and unresolved harmonics are generally very broad. As a result, the SACF is inevitably flat and unable to provide an accurate estimate of pitch lag. To reduce these undesired effects, in each ACF we only preserve all the values of local maxima as well as their two immediate neighbors along the lag axis and replace the other values by zero (Block 1). ACFs modified in this way have local peaks at most 3 samples wide, which are significantly narrower but still allow for slight peak deviation across channels. Then we detect the common periodicity in the two categories (the interference channels and the target channels) of frequency channels separately. From the correlogram analysis discussed in Chapter 3, we find that the vertical structure as a representation of common periodicity is more evident in the ACFs of resolved

harmonic frequency channels than the unresolved frequencies. Here resolved harmonic frequency channels are defined as those channels with centre frequencies less than 1700 Hz. To obtain a reliable pitch estimate, we must pool the ACFs across a number of frequency channels. Only if there is a significant portion (quantitatively more than 1/3) of resolved harmonic frequency channels dominated by the interference signal, we can further detect the predominant pitch in the interference signal. Otherwise, the detection is viewed to be unreliable, and therefore not to be proceeded with. Summing ACFs across all the interference frequency channels, we get an $SACF_{inf}$ as a function of autocorrelation lag. Searching for the maximum of the $SACF_{inf}$ within a possible pitch lag interval $[MinPL, MaxPL]$, we can get an estimation of common period across interference channels. The search range $[MinPL, MaxPL]$ is determined from the pitch frequency range of human adults, 80~320 Hz. So we have $MinPL = 1/320 \approx 3.1$ ms and $MaxPL = 1/80 \approx 12.5$ ms. The greater the maximum $SACF_{inf}$ is, the stronger the periodicity of the interference signal. ACF is already normalized to [0, 1] and reaches the global maximum value of 1 at the zero time lag. For a quasi-periodic signal, a salient peak of the $SACF_{inf}$ indicates that the signal is more periodic and less noisy. If the global maximum value is above the dynamic threshold $0.25SACF_{inf}(0)$, the estimation of pitch lag appearing in the interference signal is accepted. Then we select those frequency units, showing a local peak around this pitch lag, to be grouped into interference stream (Block 7a). Likewise, if an adequate number of frequency units are initially grouped as target, we repeat this pitch estimation process in the target channels to refine the frequency segregation. The ACFs of target channels are pooled together to get a pitch estimation of the target signal. As we discussed above, an autocorrelation peak around the target pitch lag is not adequate as evidence that this unit is dominated by the target signal. On the contrary, if the target is detected as a voiced signal but the ACF does not

show a peak around this pitch lag, this frequency unit is certainly not dominated by the target signal. Therefore, we do not attempt to select target channels using the periodicity cue. Instead, we use it to exclude those frequency units which are clearly dominated by interference signal.

We define the dominant pitch period in frame j to be the lag corresponding to the maximum of $SACF(j,\tau)$ in the plausible pitch range of target speech [3.1 ms, 12.5 ms], or from 80 Hz to 320 Hz. For those channels where target voiced speech dominates, their ACFs have peaks consistent with the pitch of target speech and the summation of these ACFs generally shows a dominant peak corresponding to the pitch period.

4. **Onset segregation** is only applied when no common periodicity is detected from the target channels. If this is the case, it is probably an unvoiced consonant uttered in the target stream. An unvoiced consonant has the characteristics of a strong onset at the beginning. The consistent onsets across frequency channels are demonstrated as a prominent peak in the summary onset map. By summing up the onset map across the target frequency channels, the common onset can be detected where a peak is find to be higher than 0.5 in the summary onset map. Again, a positive value in onset map is not adequate as evidence that this unit is dominated by the target signal. It could be an onset of the concurrent interference. On the contrary, if an onset time is detected in the target channel but the onset map does not show a positive value at that time, this frequency unit is certainly not dominated by the target signal and should therefore be ignored.

The reader should note that all the threshold values used in the current implementation are determined experimentally.

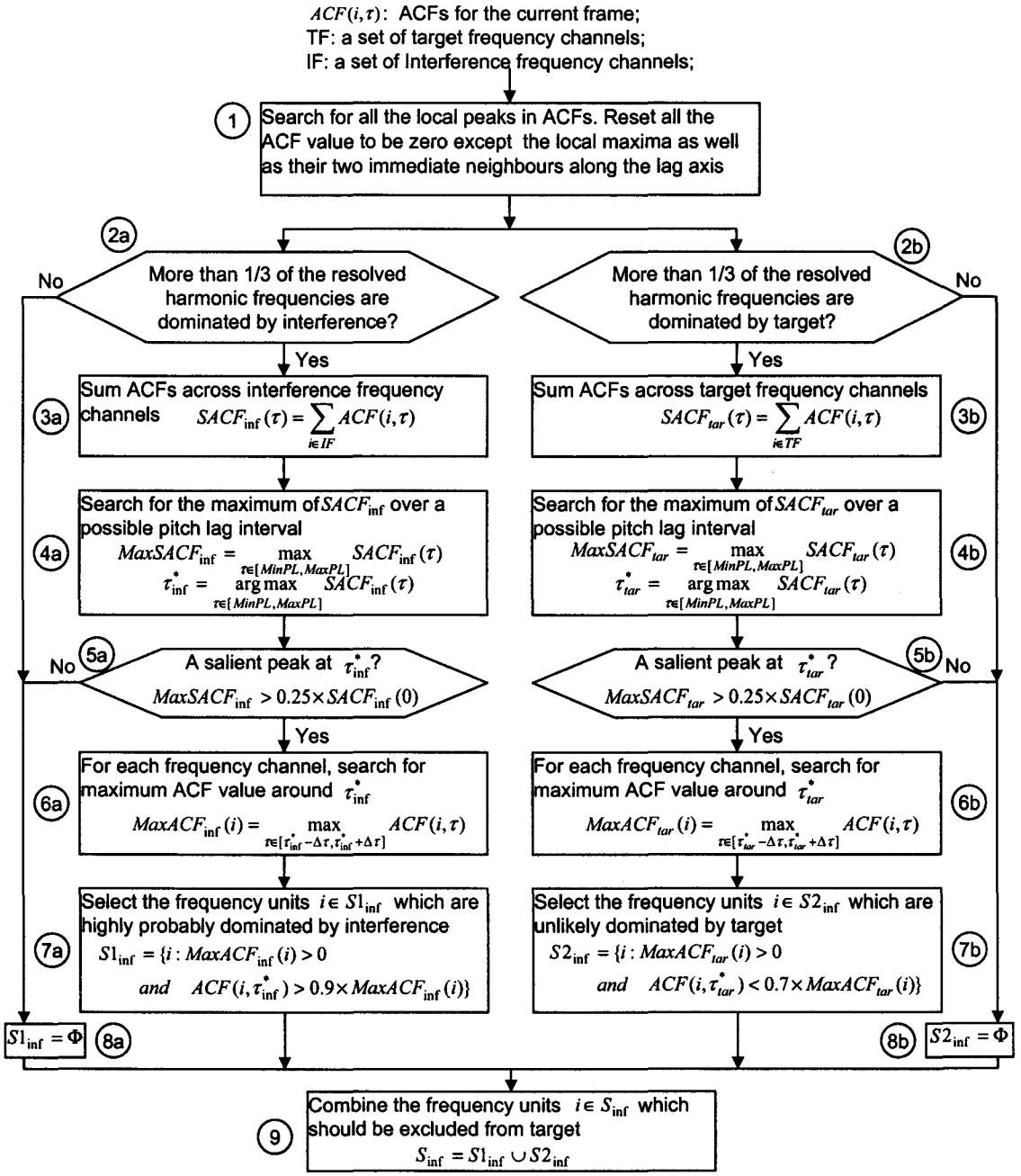


Figure 4.2: Flow chart of pitch segregation.

Chapter 5

Performance Evaluation

In the previous three chapters, we presented a model for speech enhancement using perceptual binaural sound segregation, and described an implementation of a system based on this model. In this chapter, we will examine the performance of this system.

The first part of this chapter is devoted to the background of the evaluation. We will discuss the choice of monoaural sound dataset and describe the binaural spatializing approach to simulate the complex auditory scene. Then we overview the two types of performance measurement. The ideal binary mask is also considered as an upperbound to the capacity of the system.

In the second part of the chapter, a set of systematic testing results will be discussed. First, we will investigate the proper number of frequency bands. Then, we will analyze the robustness of the system against different types of intrusions, different number of intrusions, and reverberation. We will also discuss the contribution of the individual cue in various noise conditions in the last section.

5.1 Testing Corpus

Aiming to compare the performance of our system with other similar studies (Roman et al. 2003; Hu and Wang 2004; Brown and Cooke 1994; Cooke 1993; Drake 2001; Ellis 1996), a corpus collected by Cooke (1993) is utilized to simulate two-competing-

source scenario. The corpus (denoted as Corpus I) consists of a combination of ten speech sentences and ten noise intrusions. In our evaluations, we only selectively tested our model against six noise intrusions including white noise, "cocktail party" babble noise, rock music, telephone ring, female speech and male speech. These intrusions are generally more realistic than the other intrusions, e.g. pure tone, noise burst, siren. Notice that there are two female speech sentences among the ten intrusions. We only tested one of them to avoid repetition of the same type of noise.

Corpus I is useful for comparison with the published results, but it only contains voiced utterances. Here we introduce the second corpus (denoted as Corpus II in the following text), which is composed of natural speech including both voiced and unvoiced speech. Corpus II contains ten natural speech utterances from ten speakers (five females and five males) randomly selected from the TIMIT database. The six intrusions are speech utterances from another six speakers (three females and three males). They are also randomly selected from the TIMIT database. All these target utterances and intrusions have very similar time duration (average duration is 3.172 seconds for a target signal and 3.346 seconds for intrusions) to ensure a significant amount of overlapping in the mixed signal. Corpus II was used to evaluate the model in the multiple-speaker intrusions and reverberant cases.

A full description of the contents of these two datasets can be found in Appendix A.

5.2 Binaural Spatial Synthesis

Binaural acoustic signals are required for testing of the model in this thesis. Recording an extensive test set, which covers a broad range of locations, is unrealistic. An alternative approach is to measure the binaural head-related transfer

function (HRTF) or binaural room impulse response (BRIR) and convolve them with monoaural anechoic signal to simulate binaural recording.

The HRTF characterizes how an impulse arriving at a person's (dummy's) head is smeared out by the diffraction from head and body of the person (dummy). It is used to synthesize anechoic binaural data. HRTF is usually measured by using a dummy head with a microphone mounted at the entrance of each ear. A KEMAR (Knowles Electronics Mannequin for Acoustics Research) HRTF data set is obtained from MIT media lab (Gardner and Martin 1995). In this dataset, the HRTFs were measured every 5° of azimuth in the horizontal plane (0° elevation). The HRTFs were measured at a 44.1 kHz sampling rate.

The BRIR is a set of impulse responses detected at the left and right entrances of the ear channels of a dummy head placed in a room. Unlike the HRTF, which only describes the listener's geometry, the BRIR should include all the information on receiver positions and orientations, source positions and orientations, room geometry, surface materials as well as the listener's geometry (described by the HRTFs). BRIR is used to synthesize binaural reverberant data. A library of KEMAR BRIR data is measured in a mild reverberant room (Infant Auditory Lab with drapes closed) described in Wiklund's thesis, 2003. For each recording session, KEMAR was located in the approximate centre of the room. A single speaker was moved to different locations around the room. There were 12 azimuthes (0°, 22.5°, 45°, 67.5°, 90°, 135°, 180°, −135°, −90°, −67.5°, −45°, −22.5°) measured. The height measured from the floor to the centre of the speakers diaphragm is 5'5", which is exactly the same as the height of microphones in KEMAR mannequin. The distance between the speaker to the microphones is 3'. The waveform of a typical measured room impulse response is plotted in Figure 5.1.

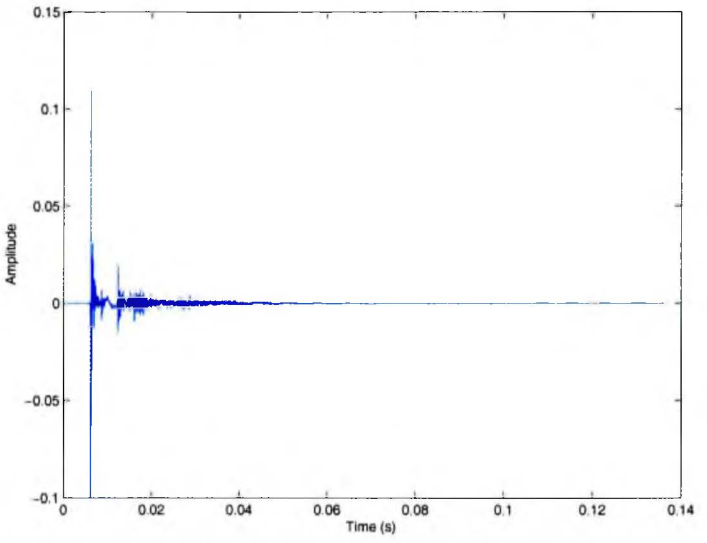


Figure 5.1: A typical room impulse response for Infant Auditory Lab with drapes closed.

In the experiments, each sound stream contained in the input mixture can be binauralized in a certain direction by convolving a monaural signal with an impulse response. For anechoic data the impulse response is obtained from the HRTF data set, while for reverberant data the impulse response is obtained from the BRIR data set. To simulate a complex auditory scene, the input sound is produced by simply adding the waveforms of multiple binaural sound streams.

5.3 Objective Performance Measurement

In this section, the performance measurements considered in this thesis are discussed. The goal for noise reduction in hearing aids is to produce speech that is perceived by the impaired auditory system to be natural and free of degradation. Obviously, a subjective quality measure is the preferable means of quality assessment. However, administrating subjective speech quality tests requires significant time and personal resources. And the results are not exactly reproducible. As alter-

natives, many objective measurements have been developed (reviewed in Deller et al. 2000). Ideally, an objective measure should replicate human performance. In this work, our system is assessed via two objective quality measurements: signal-to-noise ratio, and articulation index.

5.3.1 Signal-to-Noise Ratio

Signal-to-noise ratio (SNR) is probably the simplest and the most widely used measure for assessing noise reduction algorithms. Let $s(n)$ be the clean speech signal and $\hat{s}(n)$ be the enhanced speech extracted from the incoming noisy signal. The error signal can be written as $\varepsilon(n) = s(n) - \hat{s}(n)$. The error energy is then

$$E_\varepsilon = \sum_{n=-\infty}^{\infty} \varepsilon^2(n) = \sum_{n=-\infty}^{\infty} [s(n) - \hat{s}(n)]^2 \tag{5.1}$$

The energy contained in the clean speech signal itself is

$$E_s = \sum_{n=-\infty}^{\infty} s^2(n) \tag{5.2}$$

The resulting SNR (in dB) is obtained as

$$\text{SNR} = 10\log_{10}\left(\frac{E_s}{E_\varepsilon}\right) = 10\log_{10}\left(\frac{\sum_{n=-\infty}^{\infty} s^2(n)}{\sum_{n=-\infty}^{\infty} [s(n) - \hat{s}(n)]^2}\right) \tag{5.3}$$

It is important to note that SNR-based measurements are appropriate for noise reduction systems that seek to reproduce the original input waveform. However, SNR only characterizes the audibility of the speech; it is not adequate to measure the intelligibility of speech.

5.3.2 Articulation Index

The other widely accepted speech quality metric is the articulation index (AI). AI predicts speech intelligibility performance as judged by a human listener. AI was originally proposed by French and Steinberg in 1947 for quality assessment of analog signals. Other researchers (Kryter 1962; Steeneken and Houtgast 1980) subsequently developed the AI measure. AI assumes that the intelligibility of a processed signal is the sum of the component intelligibility losses across a set of frequency bands that span the speech spectrum. The frequency limits for each band are normally associated with the critical bands for the human auditory system. AI assumes that distortion in one band is independent of losses in other bands. Another underlying assumption of AI is that the distortion present in the noisy speech results from either additive noise or signal attenuation. All these independent, linear assumptions are satisfied in our model.

Specifically, the way to measure AI in this thesis is to compute the SNR in five octave bands with centre frequencies of 0.25, 0.5, 1, 2 and 4 kHz, then average the SNRs across these bands to come up with a prediction of intelligibility. Calculation of the AI consists of four basic steps:

1. Decompose the signal into five octave frequency bands.

2. For each octave band, calculate the SNR over the entire waveform.

3. Clip the SNR to ensure a contribution within -12~18 dB:

$$\mathrm{SNR_i} = \begin{cases} +18 & \mathrm{SNR_i} > +18dB \\ \mathrm{SNR_i} & o.w. \\ -12 & \mathrm{SNR_i} < -12dB \end{cases} \tag{5.4}$$

4. Calculate the weighted average of normalized SNR.

$$\mathrm{AI} = \sum_{i=1}^{5} w_i \frac{\mathrm{SNR}_i + 12}{30} \tag{5.5}$$

The weight w_i represents the importance of the i-th octave band for speech intelligibility as listed in the following Table (Marsh 1999).

Centre Frequency (Hz)	Weighting Factor (w_i)
250	0.072
500	0.144
1000	0.222
2000	0.327
4000	0.234

The resulting AI score is a value ranging from 0.0 to 1.0.

5.4 Ideal Binary Mask

The objective of the model is to segregate a target signal from the mixture. From a practical standpoint, what constitutes the target is task-dependent. Ideal binary mask estimation is used in some of the computational auditory scene analysis works (Roman et al. 2003, Hu and Wang 2004), which gives an upperbound on the performance of auditory segregation. In the ideal binary mask, value one is assigned if the target energy is greater than the intrusion energy in the local T-F unit and zero otherwise. The use of ideal masks is supported by the auditory masking phenomenon: within a critical band, a weaker signal is masked by a stronger one (Moore 2003). An ideal binary mask can produce a high-quality reconstructed target for a variety of sounds unless the mixture SNR in the mixture is very low. The other properties of the ideal binary mask are summarized in Wang (2004).

Assuming the original target signals are available, we can construct the ideal mask in the same way as we estimate IID: at the output of each auditory peripheral channel, the energy ratio E is calculated for each 20-ms time frame with 10 ms overlap between adjacent time frames. For the jth frame of the signal at the ith frequency channel, this energy ratio is defined as

$$E(i,j) = \frac{\sum_{k=0}^{K-1} s_i^2(j-k)}{\sum_{k=0}^{K-1} x_i^2(j-k)} \tag{5.6}$$

where s is original target signal and the x is the corresponding mixed noisy signal, K is the integration window size and k is the index inside the window. $E(i,j) >= 0.5$ means that the target is dominant in this local T-F unit; thus the value of this unit in the ideal mask is assigned to be 1. By contrast, $E(i,j) < 0.5$ means that the interference is dominant in this local T-F unit, therefore the value is set to zero. After obtaining this ideal mask, we need to apply the phase delay which is cancelled out at the end of auditory peripheral model and then resynthesize to generate an ideally segregated sound. We calculate the SNR and AI scores after applying this ideal mask and compare them with the performance of our model.

5.5 Simulation Results

5.5.1 Experiment 1: number of frequency bands

An experiment was performed to determine the number of frequency bands utilized in our model. In this experiment, each of the voiced speech (V0 to V9 from Corpus I) is presented at 0° as the target signal. The interference is a female speech (N9 from Cooke's corpus) presented at $-30°$. The amplitude of interference was scaled to give a desired range of SNR.

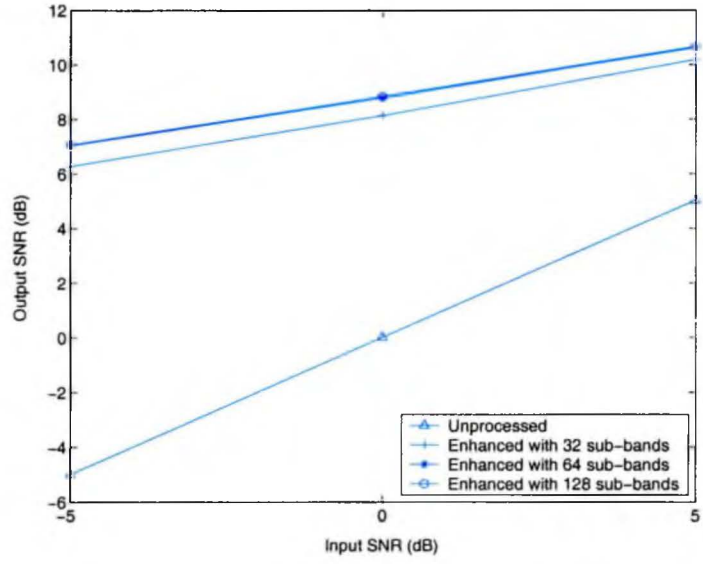


Figure 5.2: Compare the SNR results of enhancement with respect to a varying number of frequency bands (i.e., 32, 64, 128).

Figure 5.2 compares the SNR value of the original mixture and the improved SNR using the proposed model with a varying number of frequency bands. We varied the intensity of noise to obtain three different SNRs (represented as three groups of tests) in the original mixture. Each point in this figure represents the average SNR value calculated across the 10 target signals. The spatial separation between the target and interference makes the SNR values largely different at the opposite ears. For clarity, we only plot the SNR results at the right ear. However, examining the results at the left ear, we arrive at the same conclusion. As can be seen from this figure, processing in 32 frequency sub-bands, our model produces a substantial SNR gain over the original mixture. However, further increasing the number of frequency bands from 32 to 64, the average SNR improvement is only 0.61 dB. From 64 to 128, the performance is indistinguishable. These results imply that the frequency resolution after splitting frequency range into 32 bands is good enough for acoustic cue estimation and auditory segregation purpose. Although a higher

resolution produces slightly better performance, this increment does not justify the much increased computational complexity. Since a fast-operating system is desired for hearing-aid applications, the number of frequency bands is fixed to 32 bands in the following experiments.

5.5.2 Experiment 2: noise type

In the second experiment, the system is assessed against six different types of noise/interference. They are white noise, babble noise, rock music, telephone ring, male speech and female speech from Corpus I. These noises are mixed with each of the ten target utterances from the same corpus. For all the tests, the target speech is fixed at 0° azimuth and the noise is presented at $-30°$ azimuth. We varied the intensity of noise to obtain three different SNRs (represented as three groups of tests). Because the noise intrusion is received from the left-hand side of the target, the SNR at the left ear is always worse than the right ear. In all of the following tests, we evaluate the performance on the two ears separately. The ten target utterances share very similar acoustic characteristics in terms of spatial localization, intensity and duration. Therefore all the SNR and AI performance presented in the following sections, if not specified, is an average value calculated across the ten target signals.

Figure 5.3 shows the performance of our proposed model and compares with the results using the ideal mask. Basically, the AI value of the input signal indicates the difficulty of the segregation task. The variation in performance over the different types of noise is mainly attributed to the energy distribution of the noise in the time-frequency plane and the amount of time-frequency overlapping between the target signal and the noise. For example, the energy of telephone ring is concentrated on a compact area in the time-frequency plane and can be easily distinguished. Therefore,

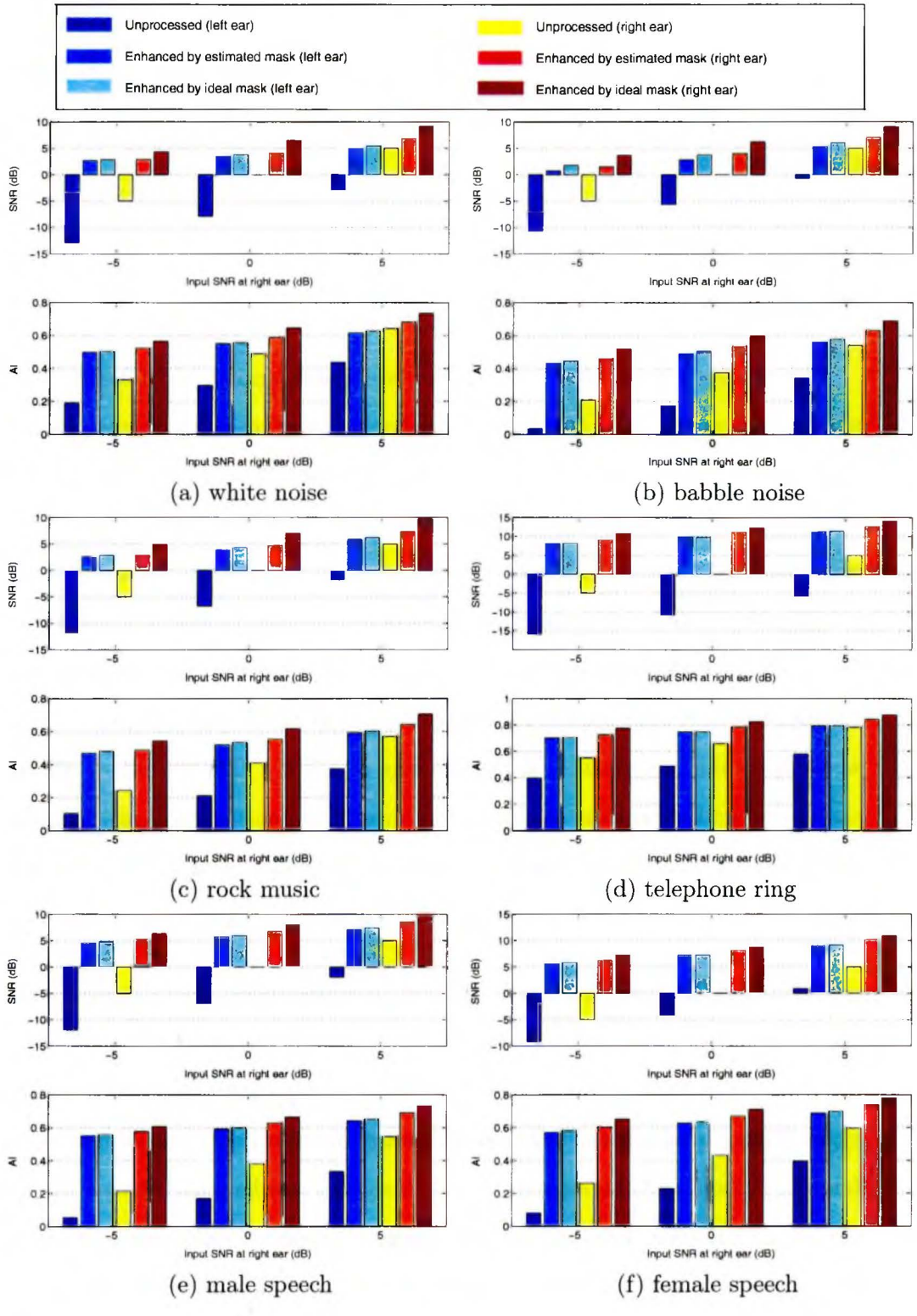


Figure 5.3: Comparison of SNR and AI before and after enhancement in presence of different type of noises.

at the same SNR, speech corrupted by the telephone ring is the most intelligible one. As a result, the enhancement performance is best with the telephone ring. In contrast, acoustic mixtures which are distributed and significantly overlapping in the time-frequency plane present the greatest challenges to the segregation. Consequently, the performance with continuous and broad-band noise intrusions is comparatively poorer. Overall, the proposed model gives a substantial improvement in terms of SNR and AI. Especially at the worse ear (left ear in this experiment), the intelligibility performance after processing is very close to the performance by using ideal binary mask. While at the better ear (right ear in this experiment), our model is inferior to the enhancement by ideal binary mask.

When the SNR of the original mixture is very low, the noise is stronger than the target speech in most of the time-frequency units. Either ideal mask or estimated mask suppresses most of the input signal; hence the output SNR is also low. Therefore, at low SNR conditions, a lower threshold should be utilized to preserve much of the target energy at the expense of increasing the residual noise.

We compare the results with other sound segregation systems using computational auditory scene analysis techniques. All these models are evaluated on Cooke's corpus, which facilitates our comparison.

Table 5.5.2 compares the SNR results of our proposed model with a variety of monoaural segregation models. Each value in the table represents the average SNR gain across all the testing data in a two-competing-source simulation. As can be seen from the table, our system produces a gain of 12.61 dB (left ear, i.e., the worse ear) and 6.64 dB (right ear, i.e., the better ear) over the original mixture, which is much higher than the performance of monoaural system and very close to the performance produced by using binaural ideal mask.

Table 5.1: SNR comparison with monoaural enhancement models

Model	Average SNR gain (dB)
Proposed model	12.61(Left) 6.64(Right)
Binaural ideal mask	12.94(Left) 8.31(Right)
Pitch-and-AM-based model	7.73
Pitch-labeled mask	5.24
A_E-based mask	5.98
True pitch	8.43
Narrow band	5.21
Comb filter	4.65
Wang-Brown system	4.45
Spectral subtraction	3.08
Monoaural ideal mask	10.72

We also compare the evaluation results with another binaural speech segregation model proposed by Roman et al. (2003). Her model is based on sound localization cues (i.e., ITD and IID) without using monoaural cues. In her tests, the reported average SNR gain is 11.31 dB at the better ear, which is nearly equal to the performance produced by the ideal mask. One reason is that she uses sophisticated statistical decision rules to distinguish the target and intrusion components at very high expense of computational cost.

5.5.3 Experiment 3: number of intrusions

In the third experiment, the simulated auditory scene is complicated by gradually increasing the number of concurrent intrusions. Again, each of the 10 target signal from Corpus II is presented at 0° azimuth. There are another six intrusions from Corpus II. All these targets and intrusions are natural human speech signals spoken by different speakers.

Figure 5.4 illustrates the spatial configuration of these sound sources. In total, four scenes were simulated with different combinations of sources:

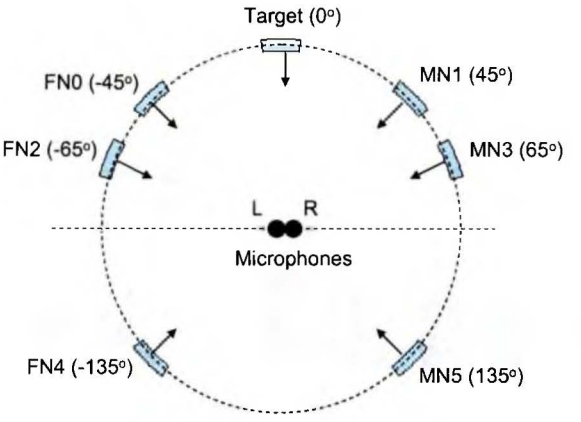


Figure 5.4: Spatial configuration.

- Scene 1: target in presence of FN0;
- Scene 2: target in presence of FN0 and MN1;
- Scene 3: target in presence of FN0, MN1, FN2 and MN3;
- Scene 4: target in presence of FN0, MN1, FN2, MN3, FN4 and MN5.

Comparing the SNR and AI values of the original mixtures, we see that the case of multiple sources scenarios (in Figure 5.5(b, c, d)) is more challenging than the two competing sources tests (in Figure 5.5(a)). According to the analysis in the previous chapter, binaural spatial cue estimation is less reliable when multiple concurrent intrusions are presented from both sides of the target signal. As expected, the estimated mask for multiple intrusions tests is less effective in comparison to the ideal mask. However, after mixing with four or six intrusions, the intelligibility of the input signal is even higher than the two-intrusion case. This can be explained by noting that, in the presence of multiple intrusions, the interference energy is more evenly distributed

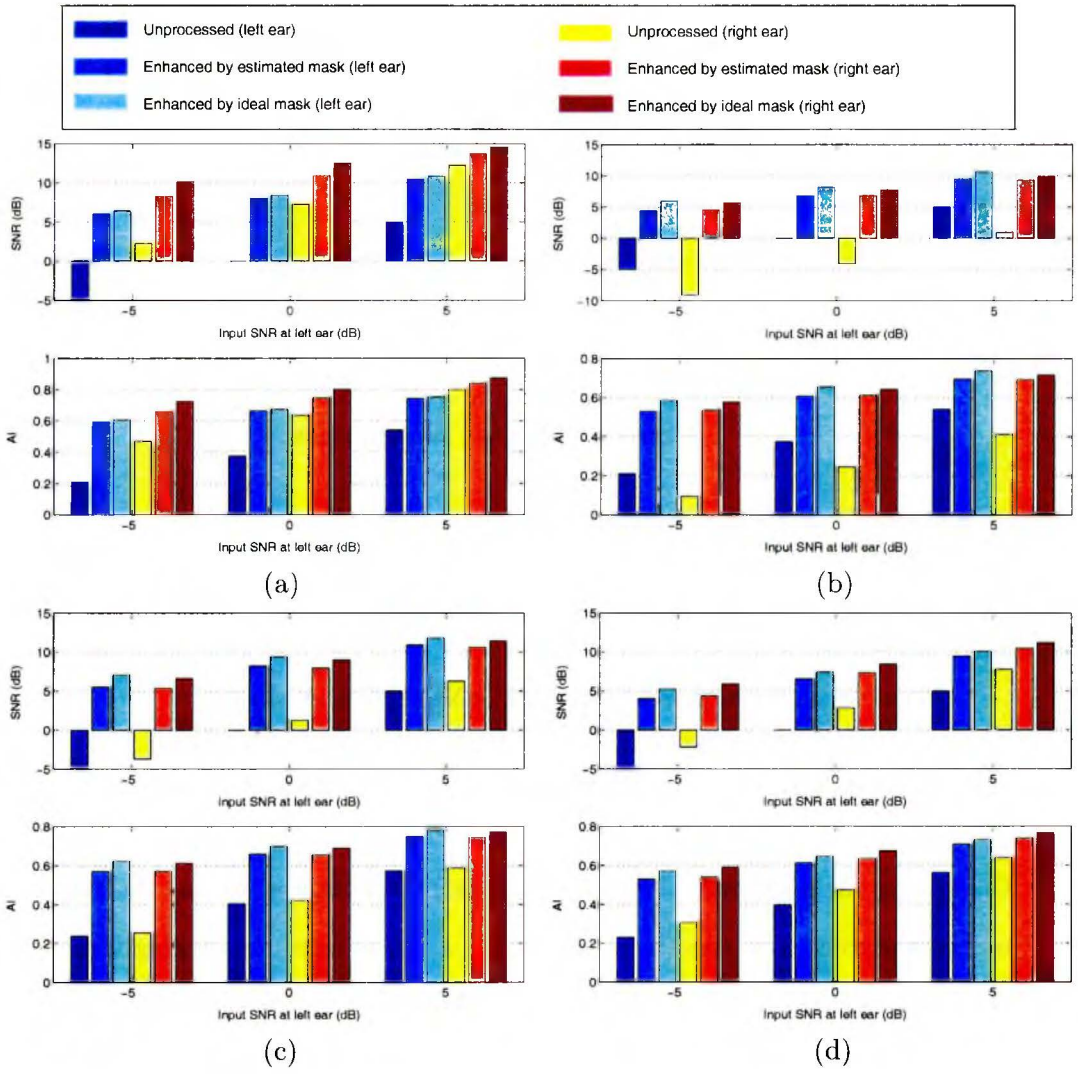


Figure 5.5: Comparison of SNR and AI before and after enhancement with different number of intrusions in *anechoic condition*: (a) with 1 intrusion; (b) with 2 intrusions; (c) with 4 intrusions; (d) with 6 intrusions.

over the time-frequency place; hence more time-frequency regions are dominated by the target signal. Consequently, more target energy can be grouped and extracted. Therefore, the performance is not degraded due to the increasing number of intrusions. On the other hand, the intelligibility gain after enhancement is even higher than the single intrusion case. This result confirms that our proposed model has no constraints on the number of concurrent intrusions.

5.5.4 Experiment 4: effect of reverberation

All the testing data in the first three experiments were synthesized by convolving the monaural signals with HRTFs, which do not take room effects into account. Therefore, the simulated scene is equivalent to that in an anechoic environment. As we know, room reverberation can severely degrade the performance of speech enhancement algorithms, and reported results often neglect this important measure. In this experiment, we repeat the third experiment on reverberant data. The monaural target and noise data remain the same as in the third experiment. The spatial configuration of target and intrusion is the same as illustrated in Figure 5.4. The only difference is that the binaural signals tested in this experiment are synthesized using reverberant room impulse response. The results of our reverberation tests, shown in Figure 5.6, reveal the following:

- The listening task is obviously more challenging when increasing the intrusion number from one to two. From two to six intrusions, the difficulty remains nearly the same.

- The mask estimated by proposed model becomes less accurate when multiple intrusions occur.

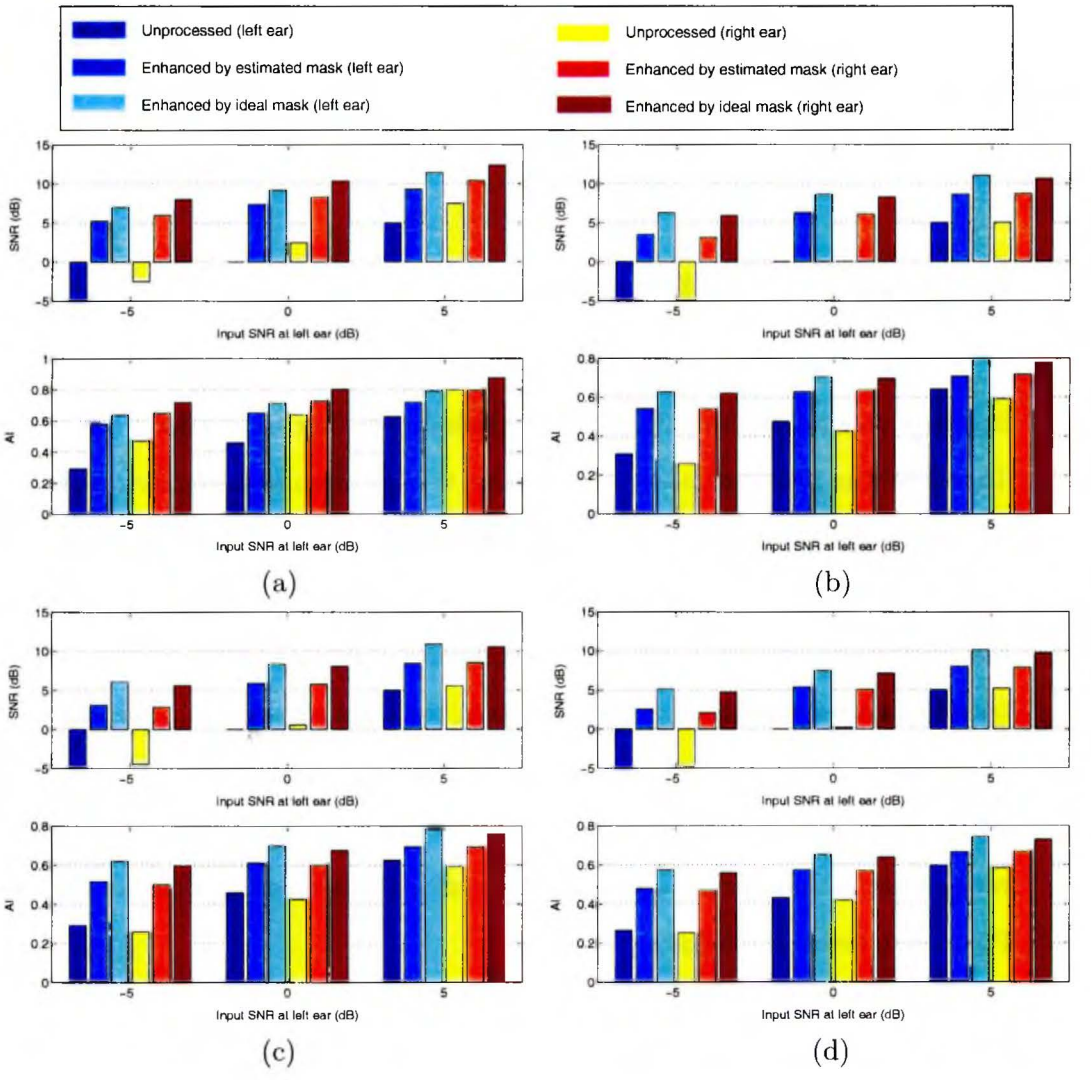


Figure 5.6: Comparison of SNR and AI before and after enhancement with different number of intrusions in *reverberant condition*: (a) with 1 intrusion; (b) with 2 intrusions; (c) with 4 intrusions; (d) with 6 intrusions.

- Reverberation degrades the performance of the proposed model. While the ideal mask is not sensitive to the room effect.

5.6 Discussion

In the previous section, we have only examined the system performance in terms of the two objective measurements. In this section, we will use typical experimental results to explain the contribution of the individual cue in various noise conditions.

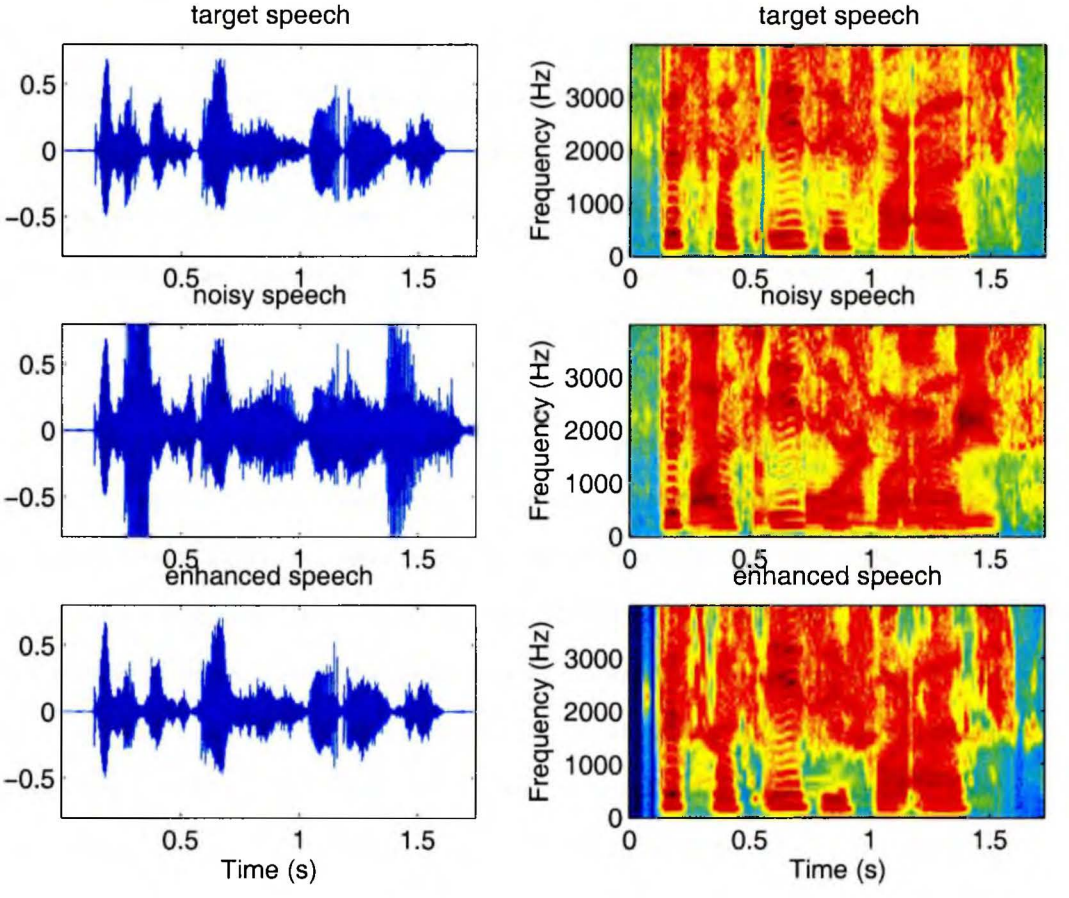


Figure 5.7: Enhancement result for competing speech segregation. The left column shows the waveform of the original target, corrupted and reconstructed speech signals. The right column shows the spectrograms of these signals.

The first experimental example is segregation of two competing speech streams. The results are plotted in Figure 5.7. In this case, the sound input is a mixture of two speech streams coming from 0 and 30 degrees. We suppose the one from the center direction is the target and the other one is interference. The SNR of the mixture signal is 0 dB. After processing, a 10.96 dB SNR gain is achieved. In this competing stream segregation task, the spatial segregation based on the ITD and IID cues is particularly effective in comparison to the pitch and onset cues. As discussed in the pitch extraction analysis, the autocorrelation function defined within the possible pitch lag is highly periodic. The ambiguity inherent in the periodicity of narrow-band auto-correlation functions makes the pitch segregation generally less reliable in comparison to the spatial segregation. On the other hand, since the two streams have similar sound pressure levels, the onset information in both streams can be detected and it is hard to tell which stream the onsets belong to. Therefore, the spatial cues are generally more robust than the monoaural cues when the target and interference sources are sufficiently separated in the space domain.

However, the benefit of spatial cues would disappear if the target source is spatially close to the interference source. Figure 5.8 demonstrates the experimental result of an extreme case. In this experiment, the input signal is target speech corrupted by white noise. The SNR of the mixed signal is 0 dB. Both the target source and noise source are localized at 0 degrees. Since the two streams have the same directional information, they are unable to be segregated by use of ITD and IID cues. Hence, both the target signal and the noise will pass through the spatial segregation. Nevertheless, due to the relatively stationary and statistically uncorrelated property of white noise, both of the two monoaural cues (pitch and onset) are effective in distinguishing the target speech components from the noise. After processing, the overall SNR is increased by 7dB. Still, some speech information is lost in the high frequency

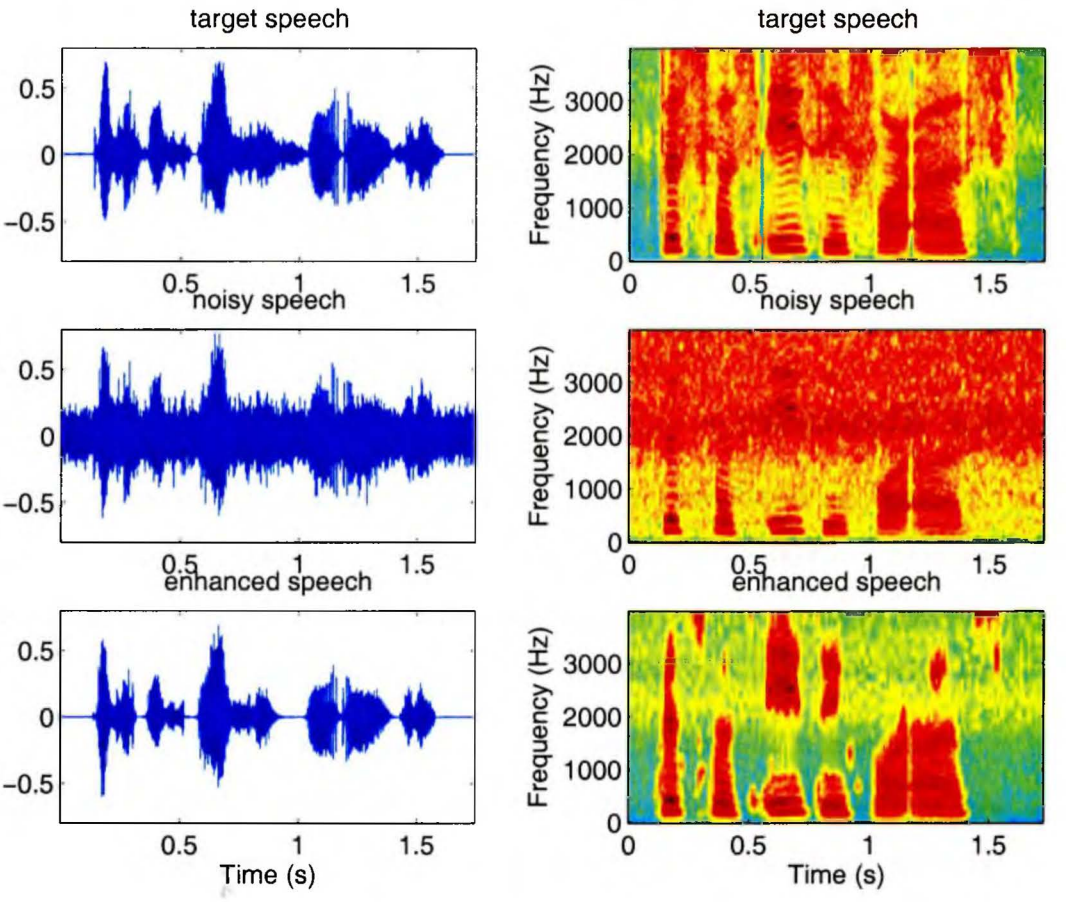


Figure 5.8: Enhancement result for speech corrupted by white noise. The left column shows the waveform of the original target, corrupted and reconstructed speech signals. The right column shows the spectrograms of these signals.

bands because energetic masking occurs in those partials. The reconstructed signal sounds somewhat muffled.

Supposing the auditory scene is further complicated by replacing the white noise with a nonstationary speech interference, it is impossible for the model to distinguish the target speech stream from the interference stream solely based on the monoaural cues. In that situation, the model may fail to produce an acceptable performance.

Chapter 6

Conclusions and Future Work

6.1 Summary

Sensorineural-impaired persons experience a hard time listening to a speech signal in a noisy "cocktail-party" environment. In the current study, we have developed an adaptive hearing system to extract information pertaining to a target speech signal in a noisy background and solve the cocktail party problem.

The model is built on our deep understanding of auditory scene analysis, which is a general process carried out by humans to extract information pertaining to a target speech signal in a noisy background. In a generic sense, the auditory environment may be viewed as a complex scene containing multiple objects. The normal-functioning human auditory system groups the sounds received from these objects into separate perceptual streams based on distinctive acoustic cues. Using advanced signal processing techniques to simulate this grouping process carried out in normal human auditory system, our proposed model appears to be able to make up for the perceptual grouping process missing from the auditory system of a hearing-impaired person.

Our model performs bottom-up segregation of an incoming signal, which is closely related to Bregman's conceptual model of the auditory scene analysis mechanism. In the first stage, the input mixture is decomposed into time-frequency elements through an auditory peripheral model. For each elementary unit of the representation, the

perceptual acoustic cues are estimated. A subsequent grouping process is operated on these acoustic cues in order to identify time-frequency components that are likely to have originated from the same sound source. The components dominated by interference are suppressed. The auditory representation of the target source is preserved and then inverted to resynthesize a time-domain waveform of the target signal.

In Chapter 2, we described the implementation of the auditory peripheral model to obtain a time-frequency representation of the incoming signal. We also proposed a low-delay filterbank inversion method to make a real-time resynthesis feasible at the end of processing. In Chapter 3, we discussed the property and estimation algorithm for each of the important auditory perceptual cue. In light of the experimental results, we analyzed the robustness of each individual cue against various noise conditions and room reverberation. In Chapter 4, we provided theoretical analysis to point out the strength and weakness of each individual cue. To dynamically resolve the ambiguities of segregation, we described a strategy to combine the evidence from multiple cues and thereby make a better grouping decision. The model was systematically evaluated, and the experimental results are presented and discussed in Chapter 5.

6.2 Conclusions

The model exhibits a number of interesting features:

- As we know, the human auditory system is the best-performing sound separation system in existence in terms of both performance and efficiency. Our model is motivated by the mechanisms involved in human perceptual processing.

- Both objective speech quality measurements (i.e., SNR and AI) and informal listening tests show that the proposed model achieves substantial improvement

on the intelligibility of a target signal, while it largely suppresses the unwanted background noise/interference.

- Superior to the monoaural noise reduction algorithms used in the current hearing instruments, our binaural model can deal successfully with a wide variety of noise intrusions, including competing speech signal and multi-speaker babbles, by exploiting the localization information.

- In comparison to the multi-microphone noise reduction techniques (e.g., beamforming, blind source separation), the simulation results confirm that our model has no constraints on the number of intrusions. Furthermore, the binaural system is good for a compact design of hearing instruments.

- Simulation results show that our model has the capacity of working in a mild reverberant environment, though the performance is slightly degraded by the reverberation distortion, which is to be expected. No special effort was made to deal with the reverberation phenomenon.

- The implementation is optimized to make it computationally efficient and low-processing delay, so that the model is suitable for implementation on a DSP chip for real-time processing.

6.3 Future Work

Our model does have some deficiencies, which need to be improved in its future development. Suggestion for future work can be summarized as fowllows:

- We have identified some important cues and shown the effectiveness of the cues. But we still need to improve the methods for the cue-extraction that will be more robust.

- The importance of different cues is dependent on the environment. A uniform cue-integration strategy is inadequate. We may need a scene analysis mechanism as the front-end processing of the model. From such a mechanism, we may then determine the confidence on each of the cues and thereby choose the right strategy to combine the cues.

- The benefit of new acoustic cues, namely those based on modulation effects in speech signals (particularly frequency modulations) deserve serious attention. Here, we may look to current work being done by Kan (2005).

- With regards to degradation in model performance due to reverberation, we need a scheme to overcome the effect of reverberation and make the system work equally well in a reverberant environment.

- In the current implementation, because the target signal is assumed to be arriving straight ahead, we apply exactly the same segregation mask to the signals received at the two ears. But in a real scenario, when the target signal is off-centre, there will be a time delay between the target signals received at the two ears. If this is the case, separate segregation masks should be applied to the two ears.

- In the current model, we aim to completely suppress the interference and produce a maximal SNR improvement by using the binary mask. However, undesirable distortions can occur as a result of the quick on-off switching controlled by the binary mask. The most notable distortion is "musical noise" in which statistical fluctuations in the frequency components of noise lead to random tonal artifacts in the processed signal. To reduce the musical noise, we need to smooth the change in the gain coefficients of the segregation mask. Consequently, the distortion to the target signal will be reduced at the expense of

more residual interference. As long as the model is still able to compensate the SNR loss of the impaired auditory system, the smoothed mask can produce a more pleasant sounding output.

Appendix A

Testing Corpus

Corpus I is collected by Cooke (1993). This corpus consists of a combination of ten speech sentences and ten noise intrusions. These speech sentences only contain voiced sound. In our evaluation, we only selectively tested against six noise intrusions: white noise, "cocktail party" babble noise, rock music, telephone ring, female speech, and male speech. These intrusions are generally more realistic than the other intrusions, e.g. pure tone, noise burst, siren. Note that there are two female speech sentences among the ten intrusions. We only tested one of them to avoid repetition of the same type of noise. This corpus is mainly utilized to simulate the case of a two-competing-source scenario.

Table 6.1: Target Signals of Corpus I

ID	Speaker	Utterance
V0	1	I'll willing marry Marilyn.
V1	1	Why were you away a year, Roy?
V2	1	Why were you weary?
V3	1	Why were you all weary?
V4	1	Our lawyer will allow your rule.
V5	2	I'll willing marry Marilyn.
V6	2	Why were you away a year, Roy?
V7	2	Why were you weary?
V8	2	Why were you all weary?
V9	2	Our lawyer will allow your rule.

Corpus II consists of ten target signals and six intrusions. The ten targets are nature speech utterances from ten speakers (five females and five males) randomly

Table 6.2: Noise Intrusions of Corpus I

ID	Description	Characteristics
N1	white noise	wideband, continuous, unstructured
N3	babble (teaching laboratory noise)	wideband, continuous, partly structured
N4	rock music	wideband, continuous, structured
N6	telephone	wideband, interrupted, structured
N8	male TIMIT utterance	wideband, continuous, structured
N9	female TIMIT utterance	wideband, continuous, structured

selected from the TIMIT database. The six intrusions are speech utterances from other six speakers (three females and three males). They are also randomly selected from the TIMIT database. All these target utterances and intrusions have very similar time durations (average duration is 3.172 seconds for the target signal and 3.346 seconds for the intrusions) to ensure a significant amount of overlap in the mixed signal. This corpus is mainly used in multiple-speaker intrusions and reverberant testing.

Table 6.3: Target Signals of Corpus II

ID	Speaker ID	Utterance
S0	MRJS0	She had your dark suit in greasy wash water all year.
S1	MRJR0	That diagram makes sense only after much study.
S2	MPAM1	Only the most accomplished artists obtain popularity.
S3	MJDH0	Cliff was soothed by the luxurious massage.
S4	MESD0	The tooth fairy forgot to come when Roger's tooth fell out.
S5	FMGD0	The morning dew on the spider web glistened in the sun.
S6	FCMH0	Before Thursday's exam, review every formula.
S7	FDRW0	The causeway ended abruptly at the shore.
S8	FPKT0	To further his prestige, he occasionally reads the Wall Street Journal.
S9	FKMS0	The nearest synagogue may not be within walking distance.

Table 6.4: Noise Intrusions of Corpus II

ID	Speaker ID	Utterance
FN0	FLNH0	If people were more generous, there would be no need for welfare.
MN1	MJFC0	Her wardrobe consists of only skirts and blouses.
FN2	FADG0	She slipped and sprained her ankle on the steep slope.
MN3	MDRM0	Please take this dirty table cloth to the cleaners for me.
FN4	FDAC1	Husky young man, he said with mock distaste.
MN5	MSJS1	The idea of a central tank with lines to each house is not in itself a novelty.

Bibliography

Assmann, P. 1999. "Fundamental frequency and the intelligibility of competing voices", In *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, pp. 179–182.

Assmann, P. F. and Summerfield, Q. 1994. "The contribution of waveform interactions to the perception of concurrent vowels", *Journal of the Acoustical Society of America*, 95, 471–484.

Baken, R. J. 1987. *Clinical Measurement of Speech and Voice.* London: Taylor and Francis Ltd.

Bell, A. and Sejnowski, T. 1995. "An information-maximisation approach to blind separation and blind deconvolution", *Neural Comput.*, 7(6), 1004–1034.

Bilmes, J. 1993. *Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm.* Master's thesis, Massachusetts Institute of Technology.

Bird, J. and Darwin, C. J. 1997. *"Effects of a difference in fundamental frequency in separating two sentences"*, Chapter Psychophysical and physiological advances in hearing, pp. 263–269. Whurr, London.

Bodden, M. 1995. "Binaural modeling and auditory scene analysis", In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, NY, pp. 31–34.

Bregman, A. 1990. *Auditory Scene Analysis.* Cambridge, MA: MIT Press.

Brown, G. J. and Cooke, M. 1998. "Temporal synchronization in a neural oscillator model of primitive auditory stream segregation", In D. F. Rosenthal and H. Okuno (Eds.), *Computational auditory scene analysis*, Mahwah, NJ, pp. 87–103. Lawrence Erlbaum.

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

Brungart, D. S. and Rabinowitz, W. M. 1999. "Auditory localization of nearby sources I: Head-related transfer functions", *Journal of the Acoustical Society of America*, 106, 14651479.

Cariani, P. and Delgutte, B. 1996a. "Neural correlates of the pitch of complex tones. I: Pitch and pitch salience", *Journal of Neurophysiology*, 76, 1698–1716.

Cariani, P. and Delgutte, B. 1996b. "Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch", *Journal of Neurophysiology*, 76, 1717–1734.

Cooke, M. P. 1993. *Modelling auditory processing and organisation.* Ph. D. thesis, Cambridge University.

Culling, J. F., Hodder, K. I., and Toh, C. Y. 2003. "Effects of reverberation on perceptual segregation of competing voices", *Journal of the Acoustical Society of America*, 114, 2871–2876.

Culling, J. F., Summerfield, Q., and Marshall, D. H. 1994. "Effects of simulated reverberation on the use of binaural cues and fundamentalfrequency differences for separating concurrent vowels", *Speech Communication*, 14, 71–96.

Culling, J. F., Toh, C. Y., and Hodder, K. I. 2002. "Effects of reverberation on speech segregation", *Journal of the Acoustical Society of America*, 111, 2421–2422.

Darwin, C. J. and Ciocca, V. 1992. "Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component", *Journal of the Acoustical Society of America*, 91, 3381–3390.

Darwin, C. J. and Hukin, R. W. 2000. "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention", *Journal of the Acoustical Society of America*, 108, 335–342.

Deller, J. R., Hansen, J. H., and Proakis, J. G. 2000. *Discrete-Time Processing of Speech Signals.* Wiley-IEEE Press.

Drake, L. A. 2001. *Sound source separation via computational auditory scene analysis (CASA)-enhanced beamforming.* Ph. D. thesis, Dept. Elect. Comput. Eng., Northwestern Univ., Evanston, IL.

Drullman, R., Festen, J. M., and Plomp, R. 1994a. "Effect of reducing slow temporal modulations on speech reception", *Journal of the Acoustical Society of America*, 95, 2670–2680.

Drullman, R., Festen, J. M., and Plomp, R. 1994b. "Effect of temporal envelope smearing on speech reception", *Journal of the Acoustical Society of America*, 95, 1053–1064.

Ellis, D. P. 1996. *Prediction-Driven Computational Auditory Scene Analysis.* Ph. D. thesis, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA.

Fishbach, A., Nelken, I., and Yeshurun, Y. 2001. "Auditory Edge Detection: A Neural Model for Physiological and Psychoacoustical Responses to Amplitude Transients", *Journal of Neurophysiology*, 85, 2303–2323.

French, N. and Steinberg, J. 1947. "Factors governing the intelligibility of speech sounds", *Journal of the Acoustical Society of America*, 19, 90–119.

Gaik, W. 1993. "Combined evaluation of interaural time and intensity differences: psychoacoustic results and computer modelng", *Journal of the Acoustical Society of America*, 94, 98–110.

Gardner, W. G. and Martin, K. D. 1995. "HRTF measurements of a KEMAR", *Journal of the Acoustical Society of America*, 97, 3907–3908.

Ghoreishi, M. and Sheikhzadeh, H. 2000. "A hybrid speech enhancement system based on hmm and spectral subtraction", *IEEE Trans. Speech and Audio Process.*, 3, 1855–1858.

Goto, M. and Muraoka, Y. 1996. "Beat Tracking based on Multiple-agent Architecture - A Real-time Beat Tracking System for Audio Signals", In *Proceedings of The Second International Conference on Multiagent Systems*, pp. 103–110.

Grabke, J. W. and Blauert, J. 1998. "Cocktail party processors based on binaural models", In D. F. Rosenthal and H. Okuno (Eds.), *Computational auditory scene analysis*, Mahwah, NJ, pp. 243–255. Lawrence Erlbaum.

Hear-it 2004. "500 million hearing impaired people", *http://www.hear-it.org/.*

Hu, G. and Wang, D. 2004. "Monaural speech segregation based on pitch tracking and amplitude modulation", *IEEE Transactions on Neural Networks*, 15, 1135–1150.

Jeffress, L. A. 1948. "A place theory of sound localization", *Journal of Comparative and Physiological Psychology*, 41, 35–39.

Kan, K. 2005. *Hiearchical Sparse Coding of Natural Sounds Learns AM and FM Features.* Master's thesis, McMaster University.

Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H. 1998. "Application of the Bayesian probability network to music scene analysis", In D. F. Rosenthal and H. Okuno (Eds.), *Computational auditory scene analysis*, Mahwah, NJ, pp. 115–137. Lawrence Erlbaum.

Klapuri, A. 1999. "Sound onset detection by applying psychoacoustic knowledge", In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)-1999.*, Volume 6, pp. 3089–3092.

Kryter, K. 1962. "Methods for the calculation and use of the articulation index", *Journal of the Acoustical Society of America*, 34, 1689–1697.

Langer, G. 1992. "Periodicity coding in the auditory system", *Hear Res*, 60, 115–142.

Licklider, J. 1951. "A duplex theory of pitch perception", *Experientia*, 7, 128–134.

Lin, L., Holmes, W., and Ambikairajah, E. 2001. "Auditory filter bank inversion", In *Proc. IEEE International Symposium on Circuits and Systems (ISCAS01)*, Sydney, pp. 537 – 540.

Lindemann, W. 1986. "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals", *Journal of the Acoustical Society of America*, 80, 1608–1622.

Lyon, R. 1983. "A computational model of binaural localization and separation", In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)-1983*, Volume 8, pp. 1148–1151.

Lyon, R. 1984. "Computational models of neural auditory processing", In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)-1984*, Volume 9, pp. 41–44.

Marsh, A. 1999. "Fundamentals of Sound", *http://www.kemt.fei.tuke.sk/Predmety/KEMT320_EA/_web/Online_Course_on_Acoustics/.*

Marzinzik, M. and Kollmeier, B. 1999. "Development and evaluation of single-microphone noise reduction algorithms for digital hearing aids", In V. H. T. Dau and B. Kollmeier (Eds.), *Psychophysics, Physiology, and Models of Hearing*, Singapore, pp. 279–282. World Scientific.

Meddis, R. and Hewitt, M. 1992. "Modeling the identification of concurrent vowels with differnt fundamental frequencies", *Journal of the Acoustical Society of America*, 91, 233–245.

Meddis, R. and O'Mard, L. 1997. "A unitary model of pitch perception", *Journal of the Acoustical Society of America*, 102, 1811–1820.

Mellinger, D. K. and Mont-Reynaud, B. M. 1996. "Scene Analysis", In A. N. P. H. L. Hawkins, T. A. McMullen and R. R. Fay (Eds.), *Auditory Computation*, New York. Springer.

Moore, B. C. J. 1995. *Hearing-Handbook of Perception and Cognition* (2th ed.). Academic Press.

Moore, B. C. J. 2003. "Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms", *Speech Communication*, 41(1), 81–91.

Parra, L. and Spence, C. 2000. "Convolutive blind separation of non-stationary sources", *IEEE Trans. Speech Audio Process.*, 8, 320–327.

Proakis, J. G. and Manolakis, D. 1995. *Digital Signal Processing: Principles, Algorithms and Applications* (3rd ed.). Prentice Hall.

Rhode, W. 1995. "Interspike intervals as a correlate of periodicity pitch in cat cochlear nucleus", *Journal of the Acoustical Society of America*, 97, 2414–2429.

Roman, N., Wang, D., and Brown, G. J. 2003. "Speech segregation based on sound localization", *Journal of the Acoustical Society of America*, 114, 2236–2252.

Rosenthal, D. F. and Okuno, H. G. (Eds.) 1998. *Computational Auditory Scene Analysis*, Mahwah, NJ. Lawrence Erlbaum.

Scheffers, M. 1983. *Sifting Vowels: Auditory Pitch Analysis and Sound Segregation.* Ph.d. thesis, Rijksuniversiteit te Groningen.

Scheirer, E. 1998. "Tempo and Beat Analysis of Acoustic Musical Signals", *Journal of the Acoustical Society of America*, 103, 588–601.

Slaney, M. 1993. An efficient implementation of the Patterson-Holdsworth auditory filterbank. Technical Report 35, Apple Computer.

Slaney, M. and Lyon, R. F. 1990. "A perceptual pitch detector", In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)-1990*, Volume 1, pp. 357–360.

Steeneken, H. and Houtgast, T. 1980. "A physical method of measuring speech-transmission quality", *Journal of the Acoustical Society of America*, 67, 318–326.

Stellmack, M. A. and Dye, R. H. 1993. "The combination of interaural information across frequencies: The effects of number and spacing of components, onset asynchrony, and harmonicity", *Journal of the Acoustical Society of America*, 93, 2933–2947.

Wang, D. 1996. "Primitive Auditory Segregation Based on Oscillatory Correlation", *Cognitive Science 20*(3), 409–456.

Wang, D. 2004. "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis". In P. Divenyi (Ed.), *Speech Separation by Humans and Machines*, pp. 181–197. Kluwer Academic,.

Wiklund, K. 2003. *R-HINT-E: A Realistic Hearing in Noise Test Environment.* Master's thesis, McMaster University.

Wolfe, P. J. and Godsill, S. J. 2000. "Towards a perceptuelly optimal spectral amlipitude estimator for audio signal enhancement", *Proc. IEEE International Conference Acoustic Speech Signal Process.*, 2, 701–704.

Wood, W. A. and Colburn, S. 1992. "Test of a model of auditory object formation using intensity and interaural time difference discrimination", *Journal of the Acoustical Society of America*, 91, 2891–2902.

Woods, W., Hansen, M., Wittkop, T., and Kollmeier, B. 1996. "A simple architecture for using multiple cues in sound separation", In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)-1996*, Volume 2, Philadelphia, PA, pp. 909–912.

Zurek, P., Greenberg, J., and Rabinowitz, W. 1999. "Prospects and limitations of microphone-array hearing aids", In V. H. T. Dau and B. Kollmeier (Eds.), *Psychophysics, Physiology, and Models of Hearing*, Singapore, pp. 233–244. World Scientific.