A THEORY AND COMPUTATIONAL MODEL

OF

AUDITORY MONAURAL SOUND SEPARATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By

Mitchel Weintraub

August 1985

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

(Principal Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Approved for the University Committee on Graduate Studies:

(Dean of Graduate Studies and Research)

Acknowledgements

I would like to acknowledge the following:

Earl Schubert for supervising this research and providing me with a great deal of encouragement and support;

Marty Tenenbaum and Richard Lyon for many key discussions and helpful suggestions;

Robert White, and Thomas Kailaith for serving as readers;

Larry Leifer, and Thomas Cover for their guidance and advice during my early years as a graduate student;

Doree Weintraub, Zachary Weintraub, Leonard Weintraub, Mom and Dad for providing me with the love and emotional support to complete this thesis.

Contents

Chapter 1

Introduction

Man has wondered how the human mind works for as long as history has been recorded. It is only in the last century that scientists and engineers have seriously begun to focus on the mechanisms that underlie human thinking and perception. The way that people perceive their environment has been a mystery that we are just beginning to unravel.

Scientists have seriously studied our hearing mechanism for approximately 100 years. While we are very knowledgeable in some areas of acoustics and hearing, in other areas we are just beginning to understand the amazing complexity and capabilities of the auditory system. One of these capabilities is the ability of people to listen to one person speaking in the presence of other speakers.

This thesis is concerned with how a person can listen to one person speaking in the presence of an interfering talker using a monaural recording of the conversation. Of course people have two ears, and the directional capabilities that a person gains from using two ears to focus on one talker are very important. However, even using only one ear, a person's capability to focus on a single sound is still far beyond what is achievable with today's technology.

This thesis represents an important step towards an understanding of how the auditory system accomplishes this selective listening task. The theory and models that are discussed in this thesis could not have been developed without relying on the large body of auditory literature. It is my hope that students of audition will use the knowledge in this thesis to continue to build and increase our

understanding of the mechanisms and inner workings of the auditory system.

1.1 Problem Statement

1.1.1 Definition of Terms

The popularity of computers coupled with the possibility of communicating with a machine through speech input and output has made speech recognition a popular and growing field. Speech scientists, engineers and other professionals have been working on speech recognition and synthesis for many years.

The goal of *speech recognition* is to transform the recorded signal of a person speaking (a sequence of numbers which represent the pressure variations in the air) into the text which represents what was spoken. *Speech understanding* is concerned with how a computer can comprehend the meaning of what was said by a speaker and make an appropriate response.

Since the late 1970's, *speech enhancement* has been a rapidly developing field of study. "Thus the main objective of speech enhancement is ultimately to improve one or more perceptual aspects of speech, such as overall quality, intelligibility, or degree of listener fatigue" [Lim 1983]. A related goal is *speech restoration* which is concerned with processing a speech signal in noise to compute an estimate of the speech signal which is as close as possible to the original isolated speech signal.

A new field of *sound separation* is concerned with the processing of an acoustic signal which is a combination of different environmental sounds, and the transformation of this signal into an internal representation that can be used to recognize the different sounds that are present. This writer defines sound separation as 'the processing of an incoming acoustic signal which assists in the recognition of each of the sounds that are present in the listener's environment'.

Sound separation is different from speech restoration. In attempting speech restoration, one may not be able to accurately estimate the spectrum of the speech signal because the background noise is too loud. A very loud background noise which lasts for a short period of time may make it impossible to accurately estimate the speech signal during that time. However, it does not necessarily mean that a

sophisticated recognition system will not be able to recognize what has been said. Contextual information from surrounding words, along with timing information of how long the noise masked the speech signal may allow such a recognition system to continue to work even when speech restoration is impossible.

What is important in the preceding example is that the processing system know that the desired signal has been masked, and to do the best that it can given the circumstances. Sound separation is concerned with the interpretation of the incoming signal, and determining which parts of the sound signal were generated by which sound sources. Sound separation consists of accurate estimation of the speech and interfering signals whenever possible, and using interpolation and other mechanisms when it is not possible to obtain reliable estimates.

Sound separation is concerned with the accurate estimation of each of the sounds present in the environment. Sound separation is also concerned with the accuracy of each of the spectral estimates, since this information can be very useful to a system concerned with recognizing what sounds were said. Although sound separation may try to achieve accurate spectral estimates of each of the sounds present, the more practical current goal of sound separation is to assist a sound recognition system in interpreting incoming sounds.

1.1.2 Need for Sound Separation

In most environmental conditions, the sound that a computer records will contain not only the speaker's voice, but other sounds that are also present. Many of today's speech recognition systems are based on matching spectral templates of the input signal with those stored in the recognition system. If other sounds are present besides the person speaking, then the spectral representation of the incoming sound will be a combination of the spectrum of the person speaking and of the interfering sound. The spectral distortion caused by the interfering sound will cause the speech recognition performance to decrease.

The desire to use speech recognition systems in environments where there are background noises has generated an interest in how computers can recognize speech sounds in the presence of other interfering sounds. Although it is desirable to

eliminate the interfering sounds from the listening environment so that a computer may more easily recognize incoming speech, it is not always practical or possible to pursue this option. In many situations, external noises will be present that cannot be eliminated. Unless one requires that the speech to noise ratio be extremely high, computers will be faced with the problem of interpreting incoming speech in the presence of other interfering sounds. Sound separation is therefore an important issue if speech recognition is to become a viable mode of communication with computers.

Although it is desirable to build a speech recognition system that can function in the presence of interfering sounds, it has been hard enough to build a recognition system that works in a quiet environment. The recognition of speech is a difficult task and has had only limited success over the past decade. This difficult problem is made even harder when other sounds are present. The ability of the human auditory system to recognize sounds in either a quiet or a noisy environment is still far above the performance of any computer.

In an effort to improve recognition performance, recent research has focused on how the human auditory system works. It is hoped that if one models the algorithms used by the human auditory system, the computer's performance can approach the level of a human listener.

To understand how people separate sounds and how the auditory system functions is a challenging and fascinating subject. There is an overwhelming volume of information about the human auditory system which needs to be put together into a conceptual framework. The puzzle of how sounds are separated and interpreted in the auditory system will be solved only by the steady inquiries that researchers will continue to make in years to come. The superb ability of the auditory system to interpret incoming sounds challenges research scientists to understand the mechanisms that allow people to hear so well, and to use this understanding in the construction of machines that recognize sounds.

1.2 Relationship of Speech Enhancement to Sound Separation

Speech enhancement is concerned with making an incoming signal that contains speech plus interfering background noise more intelligible. The goal is to process the speech signal which is not very intelligible, and make it more intelligible for a person listening to the processed output sound.

What algorithms can one use to process speech corrupted by an interfering sound to make the output more intelligible to a listener? The basic approach taken by people working in speech enhancement is to compute an estimate of the speech signal, and to resynthesize this estimate for a person to listen to. If the estimate of the speech signal is very accurate, then when the estimate is resynthesized, it will sound like the original speech before the addition of the interfering signal, and will therefore be more intelligible than the speech plus noise case.

This section will briefly review the different techniques that have been presented in the literature on how to estimate the speech signal's spectrum in the presence of interfering background noise. This section makes the important point that speech enhancement techniques are not capable of handling the masking of one person speaking by a nonstationary background noise (such as another person speaking).

Speech enhancement techniques can enhance speech only in a limited class of background noise signals. A mechanism is needed to separate speech from interfering signals that are more complex than stationary background noises. That mechanism will be provided later in this thesis when the processing of the human auditory system will be discussed.

1.2.1 Speech Enhancement Techniques

Many current speech enhancement systems have tried to increase the intelligibility of a speech signal that has been corrupted by an interfering noise source. The incoming degraded speech is processed and an estimate of the speech signal's spectrum is computed. This estimated speech spectrum is used to resynthesize

a waveform which is then played to a human listener to determine whether the processed speech is more intelligible than the degraded speech.

A good review of the different speech enhancement techniques in the literature can be found in Lim [1983]. This section will briefly review some of the major approaches to speech enhancement. The different techniques used for speech enhancement are listed below:

1. Speech Spectrum Estimation through Noise Subtraction

   The major techniques which use subtraction of the noise signal are 'power spectrum subtraction' and 'correlation subtraction'. [Nawab et. al. 1981, Peterson and Boll 1981, Boll 1979, Berouti et. al. 1979, Lim 1978] In power spectrum subtraction, it is assumed that the noise's power density spectrum is known beforehand. This technique can be applied when the background noise power spectrum is stationary, and can be measured when the person is not speaking. This knowledge allows the noise power spectrum to be subtracted from the total power spectrum to yield an estimate of the speech power spectrum.

   The spectral subtraction technique is applicable only in situations where the noise spectrum is constant. If the spectrum of the noise changes over time, these changes will result in corresponding errors in the estimation of the speech signal. Lim [1979] has shown that spectral subtraction techniques result in a higher signal to noise ratio, improved speech quality, but demonstrated no increase in the intelligibility of the speech signal.

2. Speech Enhancement of Voiced Speech using Periodicity Information

   This technique makes use of the periodicity of voiced speech to separate the speech from the noise. [Hanson et. al. 1983, Parsons 1976, Lim et. al. 1978] In this situation, the noise spectrum need not be stationary, and may be either nonperiodic or periodic with a different period of repetition from the periodic speech signal. Two different techniques used for estimating the speech spectrum are 'adaptive filtering' and 'harmonic selection'. In adaptive comb filtering, the period of repetition of the voiced speech is estimated,

and the speech plus noise is passed through a comb filter that enhances frequencies near multiples of the fundamental frequency and suppresses other frequency regions (which are not multiples of the fundamental frequency of the voiced speech). The output from this adaptive comb filter is the enhanced speech signal. In harmonic selection, first the period of repetition of the voiced speech is estimated, and then the spectral amplitude of harmonics of the fundamental is estimated and used for the resynthesis of the enhanced speech.

Both adaptive comb filtering and harmonic selection rely on the periodicity of the voiced speech for enhancement. This method cannot enhance nonperiodic speech in a background noise, since there is no periodicity information present in nonperiodic speech. These methods typically compute the period of repetition of the speech signal using noise-free speech. Since the pitch has been estimated with reasonable precision and reliability from noise-free speech, these techniques can focus on how intelligibile the enhanced speech is with correct pitch information. By using an 'accurate' estimate of the pitch track from the noise-free speech, an upper limit to the enhanced speech's intelligibility can be measured (since errors in the pitch will result in mistakes in the estimation of the speech signal). The results of the adaptive comb filtering show that the signal to noise ratio increases, but the intelligibility of the processed speech decreases as the filter length increases (from 3 to 13 pitch periods) due to the nonstationarity of the speech signal [Lim 1978a].

3. Speech Estimation using an All Pole Model of Speech

This technique of speech enhancement consists of estimating the parameters to an all pole model of the speech signal, followed by the resynthesis of the speech signal from these parameters. [Lim and Oppenheim 1979, Grenier et. al. 1981, Lim 1978] The estimation techniques used to estimate the parameters of the all pole speech model are maximum likelihood estimation, minimum mean squared error estimation, and maximum a posteriori estimation. One major assumption in the parametric estimation techniques is that

the background must be white gaussian noise. It is claimed that this does not result in any restrictions since the noise can be whitened by passing the speech plus noise through a filter which will whiten the noise [Lim 1978]. However, in order to whiten the noise, the noise spectrum must be known beforehand which results in the same stationary noise condition discussed under the noise spectrum subtraction technique. The results of using an all pole model to enhance speech in a white background noise environment show that the speech quality is improved at various signal to noise ratios, but no claims of improved intelligibility of speech are made [Lim 1983].

1.2.2 Limitations of Speech Enhancement Processing

In some situations, there is so much interfering noise that people have trouble understanding what is being said. The objective of speech enhancement is to process the incoming signal so that people are better able to understand what is said. Many speech enhancement systems have tried to increase the intelligibility of a speech signal that has been corrupted by an interfering noise source. The incoming degraded speech is processed, an estimate of the speech signal is computed, and the speech signal is then resynthesized and played back to a human listener to determine whether the processed speech is more intelligible than the degraded speech.

Current speech enhancement techniques are designed to enhance the speech signal by using some acoustic property which differentiates the speech from the noise. Assumptions typically made are that the noise is stationary (which allows for spectral subtraction of the noise from the total signal) or that the speech is periodic and the noise is not (enabling the amplitude of the speech harmonics to be estimated). These assumptions limit the complexity of the sound separation task, and focus on the acoustic differences between the speech signal and the noise signal for the estimation of the original speech parameters.

The speech enhancement techniques that were discussed earlier are applicable only in certain situations. The sounds to be separated must differ along some dimension so that a technique can be developed to exploit this difference.

Speech enhancement techniques exploit known differences between the speech and interfering noise in order to obtain an estimate of the speech spectrum. The dimensions along which speech and noise sounds differ have been constrained to those dimensions that offer theoretical mathematical techniques for their solution.

Even though these techniques have been demonstrated to enhance the subjective 'quality' of the processed speech, they have not succeeded in improving the intelligibility of degraded speech. Human listeners report that the resynthesized speech sounds less noisy, but their intelligibility scores remain at or below the level of the unprocessed speech.

Why has it been so difficult to improve the intelligibility of speech in the presence of other interfering sounds? This writer's opinion is that it is unlikely that one can improve the intelligibility of speech corrupted by an additive background noise. Below are listed three reasons why it may not be possible to improve the intelligibility of speech in the presence of interfering sounds:

1. The auditory system uses the same information (such as known noise spectral density, or periodicity information) to separate sounds as the speech enhancement systems. In order to improve the intelligibility of the degraded speech, a computer must use the same information 'better' than the auditory system uses it. Since the human auditory system's capabilities are currently far above those of any machine, it seems unlikely that a computer could use a particular piece of information better than the auditory system.

2. Although the auditory system uses many sources of information for separating sounds, only one technique is used by any speech enhancement technique. Each of the speech enhancement systems uses only a single type of information (such as known noise spectral density, *or* periodicity information) to estimate the spectrum of the speech signal. The auditory system is free to use both of these information sources as well as many other knowledge sources (use of pitch dynamics, onsets and offsets, amplitude modulation) for separating speech from the interfering noise.

3. Even if speech enhancement techniques were able to estimate the speech

spectrum with the same accuracy as the auditory system, the auditory system uses information about the interfering noise to aid it in its recognition of the speech signal. If the interfering noise is very loud, it may be difficult or impossible to estimate the parameters of the speech signal. The auditory system can use this knowledge that the speech signal is 'masked' by the noise signal to aid it in its attempt to recognize the speech signal. The resynthesized sound of the speech enhancement techniques contains no information about the interfering noise or the uncertainty of the speech estimate.

These difficulties make it uncertain whether speech enhancement will ever be able to improve the intelligibility of speech degraded by an interfering noise source. Speech enhancement has only been shown to improve the quality of processed speech and decrease listener fatigue in normal hearing subjects [Lim 1983].

A more appropriate goal for the processing of degraded speech is sound separation. Sound separation is the processing of an incoming signal which assists in the recognition of each of the sounds that are present in the listener's environment. Instead of aiding a person in the separation of speech from interfering sounds, a sound separation device could aid a computer in its recognition of speech in a noisy environment.

Speech enhancement systems, originally designed to improve the intelligibility of speech in the presence of noise for human listeners, are now being considered as preprocessors for speech recognition systems. Recognition systems currently work by classifying sequences of incoming spectral slices as one of the possible words in the allowable lexicon. Speech enhancement systems could be added to a speech recognition system to provide estimates of the speech spectrum in the presence of interfering sounds. Even though speech enhancement systems have been unable to improve the intelligibility of degraded speech for a human listener, a computer that is trying to recognize speech may find the enhanced speech to be of great benefit over the unprocessed signal.

An important point that is emphasized in this thesis is that sound separation and sound recognition in the auditory system are not disjoint systems that work in a serial fashion, but work together in order to interpret incoming sounds. The

integration of separation and recognition processing into a joint interpretation model has advantages over the previous approach of a cascade model of separation and recognition.

Sound recognition can provide sound separation mechanisms with feedback to improve the capability and performance of the separation system. Besides providing a recognition system with a spectral estimate of the speech signal, it can also provide the recognition system with information about the accuracy of the spectral estimate and an estimate of the spectrum of the interfering noise. These quantities are not provided as output because current speech recognition systems do not use this information, since they have been designed for sound recognition in a quiet environment where no other interfering sounds are present.

In summary, current speech enhancement techniques are designed to enhance the speech signal by using a single acoustic property which differentiates the speech from the noise. The dimensions along which speech and noise sounds differ have been constrained to those dimensions which offer theoretical mathematical techniques for their solution. Limiting assumptions (such as known noise spectral density, or that the speech is voiced) are made which allow the speech enhancement techniques to exploit known differences between the speech and interfering noise in order to obtain an estimate of the speech spectrum.

Even though these techniques have been demonstrated to enhance the subjective 'quality' of the processed speech, they have not succeeded in improving the intelligibility of degraded speech. It is not clear to the author if improved intelligibility is an achievable goal. Instead of focusing on speech enhancement, emphasis should be directed towards sound separation. Instead of focusing on helping people recognize speech in a noisy environment (which they already do quite well), sound separation focuses on aiding a computer to recognize a person speaking in the presence of other interfering sounds.

1.3 Research Goals

The capabilities and performance of the human auditory system in interpreting incoming sounds are superior to those achievable by a computer. Any

computer system that could claim to separate and recognize sounds as well as the auditory system would be an instant success and in high demand. Since the auditory system is capable of such a high level of performance and since we do not know currently how to achieve this level, this thesis focuses on how the auditory system separates sounds.

This research focuses on developing a conceptual approach concerning what knowledge and information the auditory system uses to separate sounds, and how the auditory system uses this information to separate them. This research is also concerned with the construction of a detailed model of the sound separation processing. The next two sections will discuss in more detail what this thesis has set out to accomplish.

1.3.1 How Does the Auditory System Separate Sounds?

A detailed theory that explains how the auditory system separates sounds does not currently exist. The goal of this research is to understand how the auditory system separates sounds using acoustic information present in the incoming signal. One objective of this research is to understand what information is used by the auditory system to separate sounds. A second objective is to discover what transformations and representations the peripheral auditory system performs on the incoming sound. A third objective is to learn the ways in which this information is used by the auditory system to separate and interpret the incoming sounds that it hears.

The development of a theory of how the auditory system separates sounds encompasses many different areas of auditory research. The relationships of sound separation with these different fields have been carefully reviewed by this writer. The diverse areas of auditory research which provide insight into the separation mechanism are:

- Mechanics of the cochlea and the transduction of sounds
- Representation and encoding of sounds by auditory nerve fibers

- Theories and psychological experiments to determine how the auditory system perceives and uses periodic information

- Psychoacoustic experiments concerned with when the auditory system will perceive two sounds, and when two acoustic stimuli will fuse into a single percept

- Theories and psychological experiments on selective attention, and how the auditory system is able to focus its processing on a single sound source

- Similarities between the interpretation of sounds and the interpretation of visual information

- Gestalt psychology and how the mind organizes, reasons with, and interprets information

This wealth of information about the auditory system has shown that the separation mechanism is an extensive and complicated process. It is hypothesized in this thesis that sound separation operates on several different 'levels' of processing and interacts with sound recognition and sound understanding to jointly interpret incoming information. This research focuses on a single part of the overall separation mechanism - how the peripheral auditory system uses acoustic information to separate sounds. The focus is on how the auditory system separates sounds using 'bottom up' or 'data driven' processing. Each of the above areas of research has contributed to an understanding of how the auditory system uses acoustic information in its interpretation of the sounds that it hears.

The writer has developed a theory of how the auditory system uses acoustic information for the separation of incoming sounds. This theory deals with the goals of auditory sound separation as well as the mechanisms it uses to achieve its goals. The information, representation, and transformations that the auditory system uses to separates sounds are hypothesized. This theory along with other relevant information is reviewed in chapter 2, where the operation of the auditory system is discussed in some detail.

Even with the current level of information about how the auditory system works, we are still far away from a precise understanding of the actual operations and transformations that the auditory system uses. Although the theory of sound separation is based on a hypothesis of how the auditory system separates sounds, the actual details of how information is combined and how different quantities are computed and used in the auditory system are unknown. Therefore, the details that are necessary to complete this model of auditory sound separation are not currently known.

Human pitch perception is an example of an auditory process that has been extensively studied for many years now. Experimental data has been unable to distinguish between the different theories of pitch perception. The actual mechanisms that the auditory system uses to compute the pitch of a signal remain unknown. The large effort that has gone into studying pitch perception and the uncertainty that still exists about the pitch processing mechanism has important implications for students of the auditory system. It is this writer's opinion that it will be a very long time before the actual mechanisms of the auditory system are documented and understood. Until these details are uncovered, it is useful to hypothesize and test theories and models of how the auditory system processes sounds.

This writer has developed a computer model which separates sounds based on the theory of human sound separation. The goals and objectives of the computer model will now be discussed in more detail.

1.3.2 The Construction of a Computer Model

The construction of a computational model of auditory processing would be nearly trivial if we knew what operations are performed by the auditory system. The algorithms of the computer model are only estimates of the actual algorithms used, since we do not know the precise details of how the auditory system operates. It is extremely difficult to determine, (out of all the possible mechanisms that could account for the auditory system's behavior) what the auditory system actually uses.

The current research effort has used a large body of experimental literature to develop the theory of auditory sound separation. A computer model which implements this approach to auditory sound separation processing has also been developed. The construction of this detailed model has raised many questions and issues, and has helped to jointly evolve an understanding of what the auditory system is trying to accomplish as well as how it accomplishes this.

When one is trying to model some process, there are several different types of models that one can construct. Since these models can differ in their objectives, listed below are three different types of auditory models that can be constructed:

Literal Model: In a literal model, the model's parameters and output correspond to actual variables and quantities that exist in the original system that is being modeled.

Black Box Model: A black box model computes the same output that the original system computes, but the computational mechanism for arriving at the output may be different from the actual process.

Functional Model: A functional model hypothesizes both the computational mechanism and the output of the system, and tries to functionally simulate what is occurring in the original system.

The computer model presented in this thesis is a 'functional model'. The intent is to compute the same quantities that the auditory separation system computes, and to use them in the same way that the auditory separation system uses the information. The computational model of auditory sound separation is concerned with what is computed by the auditory system and how these computations contribute to the successful separation of sounds. Our current understanding of the detailed computations performed by the auditory system is primarily limited to the peripheral auditory system. Not much is known about the detailed processing of the central auditory nervous system. Both the computer model's output and the mechanisms for achieving this output are hypothesized as the mechanisms and representations that the auditory system uses to separate sounds.

Due to the number of different interacting factors in an auditory model, the model's complexity is too great for it to be understood on paper alone. The use

of computers allows one to simulate how the model will function in different circumstances. In a complex model, not all of a model's behavior can be predicted beforehand. By studying the output of computer simulations of the sound separation process, one may observe different effects not foreseen before the model's construction. A computer model has the advantage of not only specifying precisely what algorithms are used, but of being useful in studying the intricate interaction between the many factors that influence the sound separation process.

The computational model of sound separation presented in this thesis tries to functionally simulate the important steps in the use of acoustic information for the separation of sounds. It is based on the theory that the auditory system computes similar quantities, even if the algorithms and the representations that the auditory system uses differ slightly from those presented in this thesis.

The current implementation is focused on the sound separation process at the lowest levels of auditory processing. It does not make any use of higher level linguistic information used by the auditory system when it separates sounds. A detailed model of the complete separation process is a very large project and is beyond the scope of this thesis. A detailed model of the complete auditory separation process would require the addition of an auditory recognition unit that would interact with the separation mechanism to jointly interpret the incoming information.

How does one evaluate a computer model of the lower levels of auditory separation processing when the upper limit of separation performance is not known and when it is not clear what the optimum solution to the separation problem is? This thesis has developed several techniques to evaluate the accuracy and performance levels of the separation algorithms that have been developed. In addition, the separation output is connected to an existing speech recognition system in a cascade fashion to measure the recognition accuracy of the separated output.

1.3.3 Limitations of the current focus

The auditory system is a complex mechanism that is not fully understood. To limit this thesis to a reasonable size, several aspects of separation processing are not dealt with. Below are listed some of the issues not addressed by this research:

1. How to separate sounds when the noise is stationary and of known spectral density. This thesis focuses on interfering sounds which are complex in their nature and are not known beforehand. It attempts to separate two people who are speaking at the same time.

2. How to separate sounds with binaural information. It is clear that the use of binaural information can improve the performance of a separation system. One could also use a microphone array to focus on a particular direction of incoming sounds. The auditory system performs the separation of sounds sufficiently well with a single ear; it is important first to understand how it accomplishes this without introducing additional input channels.

3. How the auditory system uses high level knowledge to improve the separation processing. Although feedback from a recognition system can help improve separation performance, the separation algorithms employed here use strictly 'bottom up' processing in the separation of sounds.

4. How the auditory system separates sounds that are not independent of each other. How does the auditory system pick out one violin out of the many instruments playing in an orchestra and selectively listen to it? How does the auditory system recognize that there are two voices singing or reciting the same text rather than one voice? This is a difficult issue too complex to be addressed at the current time. It is also not clear that the auditory system can actually accomplish selective separation using acoustic information alone, and it may be that this process relies extremely heavily on the use of predictions of what it expects to hear to achieve this goal.

The goal of this research, then, is to understand how the auditory system separates sounds using acoustic information present in the incoming signal. The

objectives of the separation theory are to understand what information is used by the auditory system, the way that this information is used to separate sounds, and what transformations and representations the peripheral auditory system performs. The computational model of sound separation presented in this thesis is intended to functionally simulate the important steps in the use of acoustic information for the separation of sounds. It is claimed that the auditory system effectively computes similar quantities, even if the algorithms and the representations that the auditory system uses differ slightly from those presented in this thesis.

1.4 Overview

This writer's theory of auditory monaural sound separation is presented in chapter two. An overview of the separation processing will be presented, along with experimental results which will show that: (1) the auditory system uses many different types of information for sound separation, (2) sound separation occurs at different levels in the auditory system, and (3) sound separation and sound recognition work together and can be viewed as a part of the perceptual organization of the incoming data. Even though this thesis focuses only on the use of acoustic information for sound separation, the joint workings of the separation and recognition mechanisms will be stressed to emphasize how they collectively decide what parts of the incoming sound came from which sound sources. After the overview of auditory sound separation, a model of the auditory mechanism used for separation will be discussed in detail. The different representations employed and decisions that the auditory system must face are stressed.

A computational model based on this theory of auditory sound separation is presented in chapter three. It reviews the different representations and transformations used in the separation algorithms. Models of cochlear filtering and the use of periodic information by the auditory system are discussed. The limitations of the first version of the computer model are presented along with a detailed description of the second generation of computer modeling of auditory sound separation. How the system determines how many sounds are present, and how the

spectral estimates of each sound present are computed are also presented.

An evaluation of the current theory and algorithms is presented in chapter four. Experimental results document the accuracy and capability of each component in the computer model. The limitations of the current computational model are reviewed to point out what the problems are and what issues the model leaves unsolved.

Chapter five will discuss the future directions of research on sound separation. Suggestions about how the computational model can be improved will be discussed as well as the addition of other mechanisms such as binaural processing. It will focus on the interface of a separation system with a recognition system, and what requirements and modifications this imposes on a classification system. Psychoacoustic experiments that are needed to better understand the auditory separation system are discussed. A summary and discussion of the potential of this approach to sound separation are also included.

Chapter 2

A Theory of Auditory Monaural Sound Separation

2.1 Need for an Auditory Model of Sound Separation

If one examines any book on speech perception, one will typically find several different theories on how the auditory system perceives speech. Some of the theories usually listed are the *distinctive feature* model, the *motor theory* model, and the *analysis by synthesis* model [Sanders 1972]. These theories of speech perception differ in some ways (e.g., passive versus active models) but are all models of how the auditory system hears a single sound. A typical model of speech perception is shown in figure 2.1.

The model of speech perception that appears in figure 1 shows the processing that occurs during the interpretation of a single sound. Different signal processing operations are performed on the incoming sound, which is followed by a sophisticated hierarchical classification algorithm. What is missing from this model of the auditory system is how extraneous information from another sound present is handled (or recognized as being extraneous and not belonging to this speaker). There are no mechanisms that separate sounds, or which recognize one speech signal in the presence of other interfering sounds. This chapter will present a theory of how the auditory system separates sounds, and how one sound is recognized in the presence of other sounds.

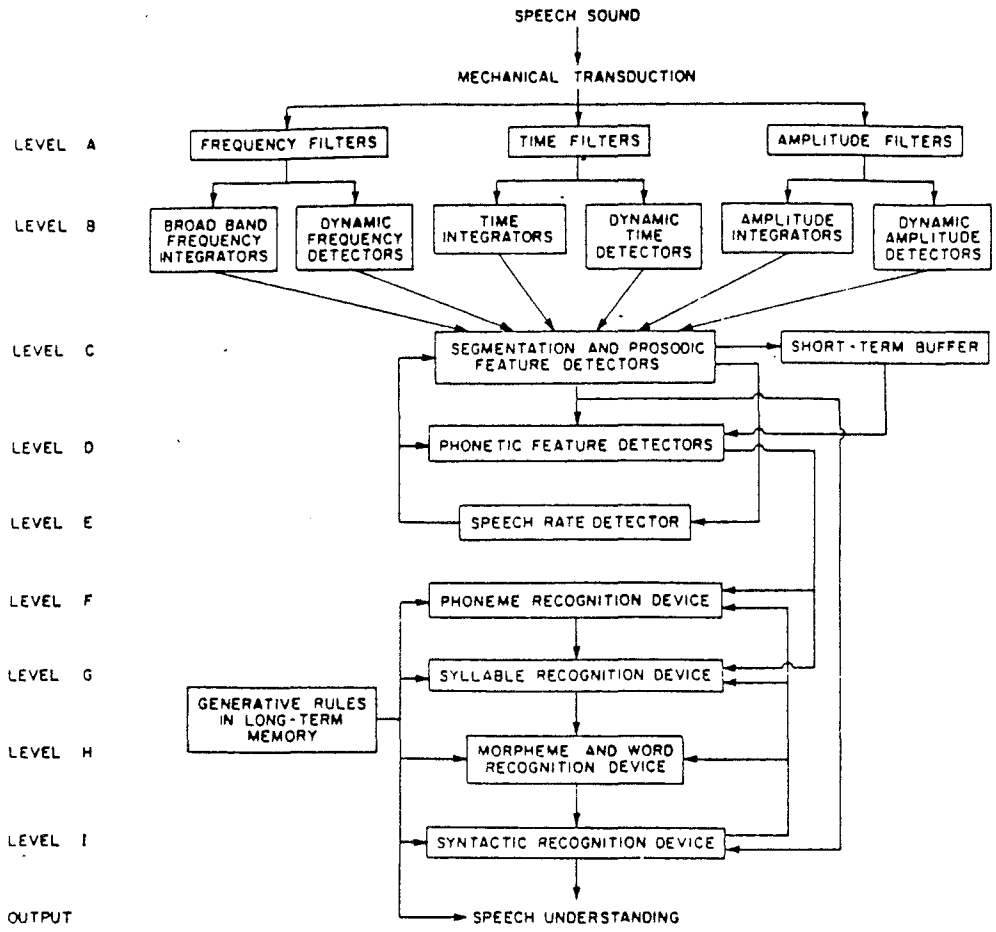

Figure 2.1: A model of the perception of a single sound source [Cooper 1979]

2.2 Goals of Early Auditory Processing

The starting point for auditory processing is the cochlea's transformation of pressure vibrations in the air into a neural representation of the sound that is heard. The cochlea filters the incoming sound into many different frequency regions along the length of the basilar membrane. The organ of corti detects the

vibrations of the membrane, and represents them as neural firings on the fibers in the cochlear nerve. These neural firing events are the only representation of information available to the early auditory system.

Each neural firing can be viewed as having certain properties, such as the time it occurred, and the properties of the neuron from which it came. It can also be viewed as having other properties associated with it, depending on the relationship between this neuron firing and other neuron firings (such as the simultaneous firing of other neurons, or the time between this firing and the previous firing of this neuron). It is the relationship between these neural firings (which compose the representation of the sound) that must be used by the auditory system to separate sounds.

One goal of early audition is defined as **determining what parts of the incoming sound belong together**. The individual neural firings are the fundamental objects manipulated by the auditory system. Based on the properties in a local frequency-time region (computed from the relationship between the different neural events), a determination is made whether the events in this local frequency-time regions are the result of the same incoming sound source, or are from different sound sources.

The auditory system must determine both how many sound sources are present and what each sound source consists of. The separation system can be viewed as computing what sounds must have been present to have caused the auditory representation that is observed.

The process of determining which sound source caused the observed neural firings in a local frequency-time region is a part of the overall interpretation processing that the auditory system performs on an incoming sound. Other, presumably higher-level parts of the interpretation process are the 'classification' of the incoming sound and the 'understanding' (the 'meaning') of what the sound is. The determination of what sound source the incoming neural firing belongs to facilitates the classification of what the sound is. The separation, recognition, and understanding mechanisms work together in the auditory system to interpret the incoming sound.

This view is similar to the role of perceptual organization proposed by Witkin and Tenenbaum [1983] in the context of vision:

> "We propose that perceptual organization is not a description of the image at all, but a primitive, skeletal causal explanation. ... The basis for these primitive inferences is the discovery of similarities – literal spatiotemporal ones – that are extremely unlikely to arise by accident. ... Within the constraints that non-accidental regularities provide, deeper interpretation proceeds by labeling, refining, and elaborating the initial model, discovering new regularities along the way as additional knowledge can be brought to bear." [Witkin and Tenenbaum 1983]

Witkin and Tenenbaum claim that when people view two parallel lines, they can be quite certain that there is some reason for that similarity. It is highly unlikely that two unrelated lines would happen to be parallel. The perceptual system 'knows' that when it sees two parallel lines it is highly likely that there is some underlying causal explanation. It is the similarity between the two parallel lines which implies some relationship between them. This relationship can be attributed to some underlying process responsible for the observed parallelism between the lines.

The same reasoning applies to the perception of sounds. Suppose that at some instant in time, there is a sudden onset in several different frequency regions. It is highly unlikely that two independent sounds started at precisely the same time, and that the simultaneous responses in different frequency regions were caused by independent sound sources. It is more likely that the auditory representation was generated by a single sound source and all the onsets that occurred at that time can be attributed to the onset of a single sound source.

In the processing of incoming sounds, if at some time two different frequency regions have properties in common (similar onsets, similar period of repetition, or other features), then one can conclude that there is probably some relationship between these two regions. The greater the similarity between the properties of the

different frequency regions and the greater the number of properties, the stronger the relationship between the two. Similar features between the two frequency regions implies that there is some relationship between them that is unlikely to arise by accident. Typically, this means they have come from the same sound source, although it is possible they have been generated by different sources (as in an orchestra or a choir when the sounds are not independent). Whether or not they have actually been generated by a single sound source, they are perceived as a single unit, as if they actually did come from the same source.

The physical processes that generate sounds obey the laws of physics and are therefore limited in the different sounds that can be produced. The time and bandwidth limitations of the sounds that we hear depend on the properties and mechanisms of the sound source. The rate at which spectral changes can occur is one constraint on natural sounds. Different frequency regions generated by natural sounds are constrained to have similar properties such as onsets, offsets, and periodicity. The auditory system uses these constraints for separating the different sounds that are present. When two pieces of information are present that are not likely to have been generated by the same sound source, the auditory system will in general hear them as belonging to separate sound sources.

The auditory system also uses the independence between two sounds to help it separate them. Most of the sounds that we hear are independent of each other. There are comparatively few sounds, such as an orchestra or a choir, where the different sound sources are not independent. Each of the sounds we hear has its own dynamics, spectral characteristics, onsets and offsets. Since each of the sounds is typically independent of the others and will have its own properties, it is unlikely that two sounds will have exactly the same information cues [Witkin 1983]. Therefore, the probability that two sounds will have the same pitch contour or onset time is small. When the auditory system sees two different pitch contours that overlap in time, it concludes that each of the contours has come from a different sound source.

Psychologists have used the principles of similarity, proximity, good continuation, and common fate to explain how the perceptual system organizes its input

[D. Weintraub et al 1966]. The separation of sounds is a part of the perceptual organization that the auditory system imposes on the incoming sounds.

The goal of determining what parts of the incoming sound belong together is very different from the goals of speech enhancement techniques, or from the goals of the 'equalization-cancellation' theory of binaural noise suppression. The equalization and cancellation (EC) model of binaural noise suppression consists of the *equalization* of the noise components in the two ears (by using time delays and amplitude scaling), followed by the *cancellation* of the noise components (by subtracting the output of one ear from the equalized version of the other ear). This model has been primarily applied to explain binaural unmasking data (increase in detectability and intelligibility through the use of binaural information).

Both the spectral subtraction techniques of speech enhancement and the equalization and cancellation technique use a subtraction operation which constitutes the enhancement of the desired signal. In neither case is there any need for further analysis of the noise signal by the recognition or understanding mechanisms. Instead of focusing on extracting the signal based on similar features, the focus is on cancelling the noise.

The difference between noise suppression and signal extraction can be illustrated with the following example. The subtraction of the output of one ear from the other ear can only form a null zone for noise coming from a single direction. The EC model will therefore have great difficulty in separating a signal from noise when there are several noise sources coming from different directions. In the limit when the noise signals in each ear are different independent noise sources, the EC mechanism is unable to cancel the noise in one ear from the other ear. However a mechanism that focuses on signal extraction will use the knowledge about which direction the signal is coming from to extract the signal from the noise that comes from a different direction.

To summarize the use of the different information cues by the auditory system: When the auditory system encounters sound patterns that are not likely to have arisen from a single sound source, the auditory system interprets them as coming from different sound sources. It uses the regularity and continuity

that natural sounds contain in order to determine how many sounds are present and what each sound consists of. It focuses on finding common properties in the representation which indicate a causal relationship between the local events in different frequency-time regions.

2.3 Overview of Auditory Separation

2.3.1 Use of Multiple Knowledge Sources in Sound Separation

One objective of this research is to understand what information cues are used by the auditory system to separate sounds. Another objective is to understand both how the auditory system computes these different information cues, and how it uses this information to separate two sounds. This section will present an overview of the auditory separation mechanism and will focus on determining what information is used by the auditory system to separate sounds.

Psychoacoustic experiments will be reviewed that show the auditory system uses many different types of information for the separation of the sounds that it hears. These information cues are: pitch,[1] pitch dynamics, the onset and offset of sounds, spectral continuity, local amplitude modulation fluctuations, visual information (e.g., lip-reading cues), and linguistic information (phonetic transitional probabilities, word transition probabilities, phrasal and message content).

Besides these monaural cues for sound separation, there are also binaural cues that aid the separation of sounds. Binaural information processing is hypothesized as consisting of many levels, just as monaural processing does. The lowest level of binaural processing is the best known and focuses on how the auditory system uses the timing and intensity differences between the cochlear output of the two ears. Binaural information at the higher levels consists of the fusion of monaural processing performed separately in each ear. Although binaural information has been shown to improve auditory separation performance, and the use of binaural information in the computer model would probably result in an increase in separation performance, the goal of this research is to understand how

[1] In much of the literature on speech, the term pitch is used to refer to fundamental frequency.

monaural sound separation is performed.

It is important to understand how the monaural auditory system uses the different pieces of information available to it for sound separation. The results of psychoacoustic experiments (discussed later in this section) can be viewed as follows: when it is highly unlikely that the information pattern (that the auditory system is attempting to interpret) was generated by a single sound source, the auditory system will hear two sound sources. Information cues that are highly unlikely to have come from the same sound source are heard as coming from different sound sources.

2.3.1.1 Use of Pitch in Sound Separation

Research on the use of periodic information for the separation of sounds dates back to experiments by Broadbent and Ladefoged in 1957. The perception of periodic information has been extensively investigated in the literature and constitutes the best known cue for the separation of sounds.

In Broadbent and Ladefoged's experiments, when two formant resonators (locations of peaks in the spectral contour) were excited by pulse trains with different periods of repetition (different fundamental frequencies), they failed to fuse into a single sound image and two sounds were heard. Experiments by Cutting [1976] showed that formant patterns presented in dichotic listening tasks will fail to fuse into a single sound image when the difference in fundamentals between the two ears is as small as two Hz (100 Hz fundamental in one ear and 102 Hz fundamental in the other). Results by Darwin [1981] also confirm that two sounds are heard when formant resonators are excited by different fundamental frequencies.

In a different series of experiments, Darwin [1977] showed that if the fundamental frequency changes abruptly during the synthesis of a continuous formant pattern, two sounds will be heard by the auditory system. While the spectral shape changed continuously over time, the pitch contour changed discontinuously between two different steady state values. At any point in time, there is only one fundamental frequency present. Each of the different frequency regions (at any any moment in time) will have the same periodicity and both are heard as coming

from the same sound source. However regions in time across the pitch discontinuity are not heard as coming from the same sound, but are assigned to different sound sources.

Presumably this happens because the human vocal system is not capable of producing abrupt pitch discontinuities during voiced speech. It is also not capable of producing different frequency regions having different fundamental frequencies. Therefore, when the auditory system encounters a situation where the sound pattern could not have been generated by a single speaker, it believes that two sound sources were responsible for generating the observed periodic information.

These two series of experiments have different implications about the auditory system's use of periodic information for sound separation. In the first example (two simultaneous formants with different fundamental frequencies), different frequency regions are heard as coming from different sound sources. The auditory system has determined that there are two periods of repetition present at the same time, and that these periodic sounds could not have been generated by the same sound source. Each frequency region that has one period of repetition is assumed to come from one sound source, but those frequency regions with a different periodicity are assumed to come from a different sound source. This experiment demonstrates that two simultaneous frequency regions can be interpreted as coming from different sound sources if there is more than one period of repetition.

In the second example (pitch discontinuity of a continuous formant contour), different time segments of the sound are heard as belonging to different sound sources. This experiment demonstrates that different nonoverlapping segments of a sound can be assigned to different sound sources if the pitch changes abruptly and the resulting pitch tracks could not have come from a single sound source. The difficulty here is to determine what are the possible pitch contours that could have been generated by a single sound source. If the pitch change had been very gradual instead of abrupt, the auditory system might have assigned the whole segment to a single person speaking. This would imply that there exists a boundary for the rate of pitch change: if the pitch changes faster than this boundary rate, the

auditory system concludes that two sounds are present; if it changes slower than this rate, the auditory system hears only a single sound present. [Note: this also raises questions about how the auditory system perceives diplophonic speakers.]

Shadowing experiments also have shown that pitch continuity is important for sound separation. In a shadowing experiment, a listener has a different message played to each ear, and is told to repeat what is heard in one ear as quickly as possible while ignoring what is heard in the other ear. If the message that a person is shadowing (the message that the person is trying to isolate) suddenly switches to the other ear, the listener will continue to follow the message that is now in the wrong ear for a short period of time [Treisman 1960]. Experiments by Simmonds and Darwin [Darwin 1978] showed that a listener would follow the wrong message depending on whether the intonation pattern was continuous. If the intonation pattern in the shadowing ear was continuous across the semantic break (when the message switched ears), the listener would hesitate but correctly shadow the incoming message. If the intonation pattern switched ears along with the message, the listener would mistakenly follow the message in the wrong ear. These experiments show that pitch-continuity information is an important cue when a person is listening to a message.

The experiments discussed above have shown that two sounds will be heard when two frequency regions have different periods of repetition, or when the period of repetition of a frequency region changes too abruptly. Periodic sounds produced by the human voice are constrained to have only a single period of repetition at a single time, and are also constrained to continuous changes in the pitch dynamics. Regions in frequency and time that are in conflict with the single sound hypothesis will be assigned to different sound sources.

2.3.1.2 Use of Pitch Dynamics in Sound Separation

Researchers have begun to study the effects of fundamental frequency dynamics on sound separation. When the fundamental frequency of a natural sound changes, the frequencies of the harmonic components of that sound will also change proportionally to the change in the fundamental component. Experiments by

McAdams [1984] indicate that when frequency components do not exhibit coherent frequency movement, two sounds are heard. The components that exhibit different frequency dynamics will stand out and be heard as a separate sound source.

The experiments performed by McAdams used different types of frequency modulation such as vibrato (periodic modulation), jitter (aperiodic modulation) and pitch glides. In one of his experiments, when 15 harmonic components of a 16 component tone are modulated coherently (with a random change in the fundamental frequency) and one of the harmonics is modulated incoherently, that harmonic component is easily heard as being separate.

Rasch [1978] performed a series of experiments on the detection of a softer note in the presence of a louder note. He found that a frequency vibrato on the pitch of the test note (depth= 4 percent, frequency= 5 Hz.) decreased the detection threshold (the amplitude that the weaker note could be detected) relative to the masking note by 17.5 db. These results indicate that the auditory system can use fundamental frequency dynamics to improve the separation of one tone from another.

When two periodic sounds are present, some of the harmonics from each sound will be close in frequency to those of the other sound. The independent motion of the fundamental frequency of the different sound sources can improve separation since the auditory system can use such cues to prevent the assignment of harmonic energy to the wrong sound source.

2.3.1.3 Use of Onsets and Offsets in Sound Separation

Amplitude changes in different frequency regions can be used by the auditory system as an indication of whether the two different frequency regions were created by the same sound source. For many types of sounds, when a sound segment begins, the different frequency regions will all start at roughly the same time. The simultaneous starting and stopping of the cochlear output in different frequency regions can be used by the auditory system to determine when two frequency regions have originated from the same sound source.

Rasch [1978] demonstrated that if two musical notes start at different times, they will not fuse into a single 'sound object' but will be heard as separate notes. If the starting discrepancies are as small as 30 msec, subject will not hear one note as starting before the other but will hear two separate notes. Rasch claims that "the two notes are perceived as two separate but simultaneously occurring sounds."

Bregman and Pinker [1978b] demonstrated that the relative onset between two pure tones is an important factor in how the auditory system perceives them. A pair of roughly synchronous tones (called B and C) were alternated with another tone (called A) which was approximately the same frequency as tone B. If the tones B and C had simultaneous onsets, they were more likely to be perceived as belonging to the same stream. As the difference in time of onset between the two tones increases, the two tones are less likely to belong to the same sound stream and tone B was more likely to stream with tone A. This experiment indicates that the relative onset time between two different frequency regions is an important cue as to whether they have originated from the same sound source.

Experiments by Darwin [1984b] have shown that if a harmonic of a vowel starts or stops at a different time from the rest of the vowel's harmonics, it will be perceptually segregated from the vowel. This effect is present even at an onset disparity of 32 msec, but longer differences between the onset or offset of the harmonic and the vowel allow better separation of the harmonic from the vowel. In other experiments he showed that "a harmonic that starts at the same time as a short vowel but continues after the vowel has ended contributes almost as little to the vowel's phonetic quality as a harmonic that starts before but stops at the same time as the vowel." [Darwin 1984a]

The experiments discussed above demonstrate that the simultaneous onset and offset of different frequency regions are important factors in the perception of a single sound. A difference in onset times can cause the auditory system to perceive that two sounds are present. If different frequency regions have different onset or offset times, the auditory system may interpret this difference as an indication that they came from different sound sources.

2.3.1.4 Use of Common Amplitude Modulation in Sound Separation

Whereas the pitch dynamics information cue dealt with a common motion of the period of repetition in each frequency channel, common amplitude modulation deals with the common fluctuations in amplitude in different frequency regions. The term 'common amplitude modulation' differs from onsets and offsets of sounds since it is defined as the change in amplitude of an already existing sound. Once a sound has started, the amplitude fluctuations present in each frequency channel can be measured.

Hall, Haggard and Fernandes [1984] have performed some important work which demonstrates that common amplitude modulation can be used by the auditory system to improve the detectability of a pure tone in noise.

> "Detectability of a 400 msec 1000 Hz. pure-tone signal was examined in bandlimited noise where different spectral regions were given similar waveform envelope characteristics. As expected, in random noise the threshold increased as the noise bandwidth was increased up to a critical bandwidth, but remained constant for further increases in bandwidth. In the noise with envelope coherence however, threshold *decreased* when the noise bandwidth was made wider than the critical bandwidth. The improvement in detectability was attributed to a process by which energy outside the critical band is used to help differentiate signal from masking noise, provided that the waveform envelope characteristics of the noise inside and outside the critical band are similar. With flanking coherent noise bands either lower or higher in frequency than a noise band centered on the signal, it was next determined that the frequency relation and remoteness of the coherent noise did not particularly influence the magnitude of the unmasking effect." [Hall, Haggard, Fernandes 1984, p.50]

In order for the common waveform envelopes in different frequency regions to improve the detectability of a tone, the auditory system must be capable of comparing local amplitude fluctuations in different frequency channels. These

experiments also demonstrate that the auditory system can determine if the amplitude modulation contours in different frequency regions are the same. If a single sound were present, it would have the same amplitude modulation envelope. The waveform envelope modulation used in these experiments was low-pass noise (0-50 Hz). The noise used implies that modulation envelopes on the order of 20 msec or longer can be used by the auditory system to improve detectability of a tone in noise.

The similarity in local amplitude fluctuations between different frequency channels can also be used to distinguish between different sounds. Experiments [Warren and Verbrugge 1984] have shown that a person can tell the difference between a bottle which is bouncing from one that has broken. Synthetic sounds of a bouncing bottle and one that has broken upon impact were generated with the same average spectrum, but they differ in the simultaneousness of the local amplitude fluctuations in different frequency regions. Listeners were able to differentiate accurately between these two cases.

These experiments should not be interpreted to mean that a lack of amplitude modulation routinely gives rise to the perception of two sound sources. They have demonstrated that the auditory system can compare the local amplitude modulation envelopes across frequency regions and use this information for sound separation.

2.3.1.5 Use of Visual Cues in Sound Separation

While the previous four sections have dealt with four different acoustic cues for sound separation (pitch, pitch dynamics, onsets and offsets, amplitude modulation), the next two sections deal with higher-level information cues for the separation of sounds. The use of visual and linguistic information are discussed in the next two sections even though they are not included in the current computational model (discussed in chapter three). The computer model focuses on the use of acoustic information for sound separation. The use of visual and linguistic information in a computer model of sound separation is beyond the scope of the current thesis but is included here for the sake of perspective.

Lip reading has long been used by deaf people to understand what other people are saying. They are able to look at the facial motions of someone speaking and understand what is being said. Cherry [1953] suggested that reading lips is helpful for separating voices from interfering sounds.

McGurk demonstrated that visual cues play an important part in the recognition of sounds [McGurk and MacDonald 1976]. The 'McGurk effect' occurs when visual and auditory cues conflict, and the result is some intermediate perception. When subjects see a speaker articulating /ga/ and hear the word /ba/, they often report hearing the sequence /da/. Other experiments [Massaro and Cohen 1983] demonstrated that as the acoustic signal changes gradually from a /ba/ to a /da/, the perception of the /ba/-/da/ boundary shifts when visual information conflicts with auditory information. These experimental results indicate that visual information is used by the auditory system for the recognition of sounds.

Although this evidence suggests that visual cues are used for sound recognition, they do not prove that they are used for sound separation. Visual information can be very useful to the auditory system for improving the separation performance in the presence of interfering sounds. The information obtained from looking at the movement of a speaker's lips is useful not only for determining what the speaker is saying, but for determining when he is speaking. The knowledge about when a person moves his lips can be very useful for knowing that the sound that we are currently perceiving is coming from that speaker. The synchronization of acoustic events with the desired speaker's lip movement can be a powerful cue for determining which speaker an acoustic event belongs to.

Although visual information has not been shown to be used by the auditory system for sound separation, it is reasonable to suppose that visual information can be of great benefit in the cocktail party phenomenon. The visual system provides information about when a person is speaking and what he is saying.

2.3.1.6 Use of Linguistic Information in Sound Separation

Linguistic information is commonly used in theories and models of auditory sound processing. Its typical use is in the recognition of sounds at different acoustic

levels (phonemes, words, concepts).

Experimental results of the intelligibility of speech in noise [Rubenstein and Pollack 1963; Miller Heise and Lichten 1951; Howes 1957] demonstrated that when the predictability of a word increases, the intelligibility of that word also increases. Linguistic information is therefore used somewhere in the system to improve the recognition performance. There are two possible mechanisms for this increase in intelligibility. The first mechanism is the use of context information to improve the spectral estimates of the sounds to be separated, and these improved spectral estimates are responsible for the increase in intelligibility. The second mechanism is the use of context information to allow the recognition system to eliminate spurious word sequences and correctly classify what it hears.

Context information can also operate at the phoneme level. When there are two sounds present and one sound is much louder than another sound, it is extremely difficult to hear the weaker sound. When the softer sound is masked by a loud sound, it may be impossible to estimate the weaker sound using acoustic information. The auditory system can use contextual cues of neighboring regions to interpolate what sounds could have been present in the masked interval.

A series of experiments by Warren [1971, 1972, 1974] demonstrated that if predictable phonemes are deleted from a sentence and replaced by a loud noise, listeners perceive both the loud noise and the missing phoneme. The perception of sounds that are not present has been called 'auditory induction.' The synthesis and perception of the missing phonemes has been called 'phonemic restorations'. When it appears that a sound has been masked, the auditory system supplies the sound most likely to have occurred, based on the linguistic constraints that a sentence provides. By contrast, if the phoneme is deleted and replaced by silence, listeners do not fill in the silent interval and perceive that a phoneme is missing.

Expectations about the different sounds which are present allow listeners to improve their separation performance. This improved performance is possible both in repeated listening to a sound segment and when the listener has a priori knowledge of the sound (e.g. listening to familiar music). Expectations about the desired and interfering sounds, as well as knowledge of what each of the different

instruments sounds like, are used to improve separation performance. Since the auditory system knows what it expects to hear and what the timbre of the different sound sources are, our perception that we are able to clearly separate out one sound from the other sounds present suggests that we are *perceiving* the *model* of a sound and not the acoustic information present in the original signal.

Another way in which the auditory system uses linguistic information can be seen in the following example. Suppose that we are listening to a male and a female voice. The male voice says the digit /three/, and the female voice then says the digit /seven/. The digit waveforms can be spliced together digitally so that the digit /seven/ starts as soon as the digit /three/ ends. When this sequence is played to a listener, he hears a male voice saying the digit /three/, followed immediately by a female voice saying the digit /seven/. However, if we delete the /even/ part of the digit /seven/, what is left is the digit /three/ followed by the /s/ of the digit /seven/. When people listen to this waveform, they will hear the word /threes/. Since the voiced part of the female voice is missing, the auditory system interprets the /s/ as belonging to the male voice. It is only after the listener hears the voiced part of the female digit seven that the /s/ is correctly interpreted as belonging to the female voice.

Two information cues aid the auditory system in determining which speaker is responsible for saying the /s/. The presence of smooth spectral transitions between the /s/ and the surrounding voiced regions is one cue that can help the auditory system determine which speaker said the /s/. Another source is linguistic constraints. Knowledge about phonemes and phonetic transitions can be helpful for determining which speaker produced the fricated segment. Linguistically, the interpretation of the /s/ as forming a part of the digit /seven/ is a better explanation of the incoming sound than assigning the /s/ to the digit /three/.

The determination of which speaker the fricated energy belongs to is a different computation from the problems of spectral estimation. In this example, there are no overlapping speech signals that need enhancement. It is trivial to estimate the spectrum of the speech sound present since the segments are nonoverlapping. The issue is how to determine which part of the incoming sound was generated

by which speaker. This example illustrates a conceptual limitation of the speech enhancement approach.

In the experiments on sound separation performed by Cherry [1953], listeners attempted to separate two simultaneously spoken messages. In the listener's transcriptions of what was said by the intended speaker, it was observed that:

> "No transpositions of phrases between the messages occurred in this example; in other examples extremely few transpositions arose, but where they did they could be highly probable from the text."

This observation supports the hypothesis that the assignment of incoming sound segments to the appropriate speaker uses linguistic contextual information.

Sound separation uses linguistic knowledge about allowable phonetic transitions between speech segments, along with the expectations of what we expect to hear each person say, in order to determine which segment was spoken by which speaker. Linguistic information is also used by the separation system in determining what was said in regions where a masking sound obscures the sound that is being focused on.

2.3.2 Integration of Separation and Recognition

If sounds could be separated solely on the basis of their acoustic information cues, it would not be necessary for separation and recognition to work together to interpret the incoming sounds. Recognition processing would occur after the separation mechanism had separated the incoming sounds. However, the masking of one sound by interfering sounds, and the changing of the characteristics of a speaker's voice (e.g., from periodic to nonperiodic) make it difficult to separate sounds using only acoustic information. The recognition mechanism can work with the separation mechanism to jointly separate and recognize the incoming sound.

This section discusses the relationship between sound separation and sound recognition mechanisms. Experimental results will be presented which demonstrate that the 'recognition' mechanism does much more than classify the incoming sound patterns into categories.

When sounds cannot be separated on the basis of their acoustic cues, it is still possible for the auditory system to identify several simultaneous sounds. Experiments by Scheffers [1979, 1982] presented listeners with two synthetic vowels that could not be separated based on their acoustic properties, since both vowels had similar onsets and offsets and used the same excitation function in the synthesis process (either both excitation functions were periodic with the same fundamental frequency or both excitation functions were the same white noise excitation). He demonstrated that listeners were still able to identify both vowels present remarkably well (each vowel was chosen from a set of 8 vowels; both vowels were correctly identified 45% of the time when both vowels were voiced, 26% of the time when both vowels were unvoiced). This demonstrates that even when the separation mechanism is unable to separate the incoming sound using acoustic information, the recognition mechanism is still able to recognize each of the two sounds. The recognition mechanism is capable of recognizing several simultaneous overlapping patterns.

Other experiments demonstrate that even when the separation mechanism does use acoustic information to separate an incoming signal, the recognition mechanism may put the separated output back together for the classification of the sound (as if the sound had not been separated). Experiments by Darwin [1981] and Cutting [1976] synthesized each of a vowel's two formants with different fundamental frequencies. Even though two sounds were heard, the listener was able to correctly identify the vowel that was presented. If the separation mechanism had assigned each formant to a different sound stream, and if the recognition mechanism had access to only one sound stream, then it would not have been possible to recognize the vowel present. The recognition mechanism must have access to both sound streams, so that it can put the information back together and correctly classify the incoming sound.

Even though the recognition mechanism correctly identified the input as a word, it did not reverse the decision of the separation mechanism that there were two sounds present. Since each formant was excited by a different fundamental frequency, the separation mechanism used the periodic information to decide that

there were two sounds present. Neither formant alone could be classified by the recognition mechanism, but the two formants together made up a vowel. The decision of how many sounds are heard is made based on the acoustic cues of the incoming sound, and although the recognition mechanism may disagree with the results of the separation processing for classification purposes, it does not change the perception about the number of sounds present.

Both the experiment described above and other experiments have shown that the recognition system does not function by simply classifying the 'desired' signal alone. Experiments [Bregman 1978d] have shown that the auditory interpretation system must have access both the 'desired' signal and the 'interfering' signal in order to classify an incoming sound correctly. If a section of a continuous pure tone is chopped out of a signal and replaced by a wide band noise burst (whose onset and offset match the section of the tone that was extracted), the tone will be perceived as continuing through the noise. The maximum length of the noise segment for which the tone will be heard as continuing through the noise is roughly 250 to 300 msec [Rasch 1978]. If a segment of the pure tone is chopped out of the signal and no noise is added to fill in the silent interval, the tone will be heard as stopping and then restarting at a later time. Also, if the noise that is added to fill in the gap does not fill up the whole silent interval (the noise starts after the tone has stopped and stops before the tone starts again), the tone will not be heard as continuing through the noise.

These experiments demonstrate that the classification system uses both sounds present to interpret which sounds were there. If the recognition system had access only to the tone, it would not hear the tone continuing through the noise (when the noise completely fills the silent interval). The recognition system must have access to both separated sounds present to conclude that part of one sound is missing because it was masked by another sound.

This point is nicely illustrated in the visual domain in figure 2.2 and figure 2.3 [taken from Bregman 1981]. In figure 2, one can see only one part of the visual representation since the occluding figure is not present. In figure 2.2, it is difficult to determine what is present in the picture. In figure 2.3 where the occluding

figure is present, the picture is much easier to interpret.

These examples demonstrate that the auditory system does not represent the desired signal and ignore the noise as current enhancement techniques do. The auditory system represents and processes both the signal and the interfering sounds to a high level. This point will be also be discussed in section 2.4.5.

Sound recognition consists not only of the classification of what sounds are present, but of determining when some part of the acoustic input is missing, or which acoustic segment is present that does not belong. How a recognition mechanism would determine that a part is missing from the current sound or belongs to the other sound is a difficult question that remains unanswered and is beyond the scope of the current research effort. It appears that the recognition system required by this approach to sound separation is very similar to the recognition system needed for the visual recognition of objects. The recognition of objects with missing or extra line segments in the visual domain is similar to the sound recognition problem of recognizing a sound segment with missing or extra events.

Experimental results have been presented in this section which demonstrate that the 'recognition' mechanism does much more than classify the incoming sound patterns into categories. When sounds cannot be separated on the basis of their acoustic cues, it is still possible for the auditory system to identify the different simultaneous sounds that are present. Other experiments demonstrated that even when the separation mechanism does use acoustic information to divide an incoming signal, the recognition mechanism may put the separated output back together for the classification of the sound (as if the sound had not been separated). The recognition system must have access to both separated sounds present to know that part of one sound is missing because it was masked by another sound. Sound recognition consists not only of the classification of what sounds were present, but of determining when some part of the acoustic input is missing, or which acoustic segment is present that does not belong.

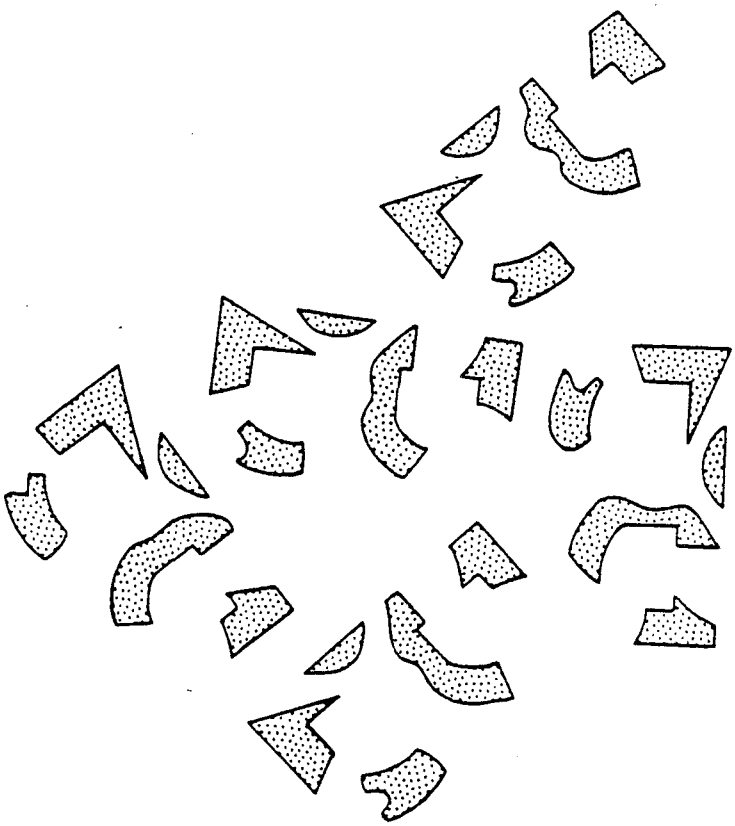

Figure 2.2: Only the picture fragments which were not occluded are shown. [Bregman 1981]

2.3.3 Limits of Auditory Sound Separation

In Cherry's experiments (where people attempt to separate two simultaneous speakers), listeners reported that the task was very difficult. Listeners would need to concentrate very hard on the material, and listen repeatedly to the recording of two people speaking simultaneously. After playing the recording many times, listeners were able to separate the incoming sounds fairly well. If the messages from each of the two speakers were a series of cliches (contained no long contextual strings), "message separation appeared impossible" [Cherry 1953].

The experimental results of Warren [1971, 1972, 1974] (discussed in section

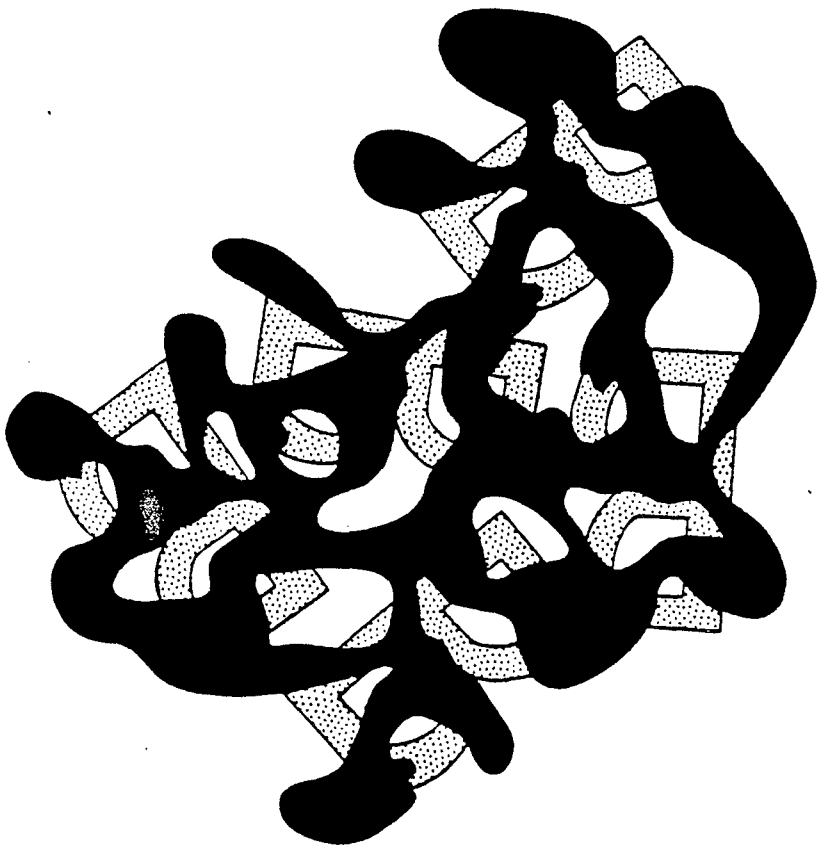

Figure 2.3: Both the picture fragments and the occluding blob are shown. [Bregman 1981]

2.3.1.6) imply that linguistic processing plays an important part in the separation of sounds. If phonemes are deleted and replaced with a loud noise, listeners will hear both the noise and the phoneme that was deleted. When the auditory system has a difficult time in acoustically separating two sound sources, the listener may use linguistic contextual information to perform phonemic restorations of the missing sound. Since listeners perceive the phoneme as if it were actually present, this might lead a listener to perceive that the sounds were easily separated, while in fact the auditory system was unable to estimate the masked sounds using acoustic information. Although people may believe that they can separate sounds from

interfering sounds with apparent ease, these experimental results indicate that sounds which are perceived as separated by the auditory system do not correspond to sounds that can be acoustically separated. It demonstrates that people perceive their models of what sounds are present, and not the acoustic information obtained from the separation processing.

The fact that people perceive their models of what information is present, and not what they actually separate using acoustic information, makes it practically impossible to determine the limitations of auditory separation using acoustic information cues.

Even though it may be difficult to determine the performance level of the auditory system, there are certain sounds that the auditory system cannot separate. If there are two sound sources and each sound source is a steady state sine wave of the same frequency, it is impossible to determine the amplitude of each sine wave. Similarly, two steady state noise-like sounds are impossible to separate. The sounds that are to be separated must differ along some dimension.

This section has shown that the auditory system uses acoustic differences in pitch, pitch dynamics, onsets, offsets, and amplitude modulation to distinguish between the different sounds it hears. It also uses visual information and linguistic information to aid in the separation of the sounds that it hears. Although we may not be able to determine the performance levels of auditory acoustic sound separation, we can explore how the auditory system uses this information to separate the sounds that it hears. The next section will review the mechanisms that the auditory system uses for the separation of sounds.

2.4 Modeling Auditory Sound Separation

The previous section has shown that different information sources are used in the separation of sounds; this section will focus on how that information is used. It will focus on the representations that the auditory system uses, and will discuss different ways that the information cues can be used to separate the incoming sound.

2.4.1 Definition of Terms

This author hypothesizes that there are three different levels of the auditory system's representation of sounds at different points in the separation processing. These levels are called 'Neural Event', 'Group Object', and 'Sound Stream'. These terms are defined below:

Neural Event:	An event occurs whenever a cochlear neuron fires. This neural encoding of an incoming sound is the fundamental quantity manipulated by the auditory separation system. (see section 3.2 where the relation between this event and auditory neural arrays is explained)
Group Object:	A group object is a collection (across both frequency and time) of neural events, having similar properties, that are perceived as a unit. It is an intermediate level in the representation of sounds and corresponds to the natural segmentation of the incoming sound into frequency-time regions that have similar properties.
Sound Stream:	A sound stream is an internal acoustic representation of a particular sound source, which consists of a temporal succession of group objects. Group objects that are assigned to a given sound stream are thought to originate from the same sound source.

Incoming sounds are encoded by the auditory system by using a neural representation. The firing of a neuron is the basic object manipulated by the auditory system in the separation of sounds. Each event has certain information associated with it that depends on the relationship between this event and other neural firings (such as the simultaneous firing of other neurons, or the time between one neural firing and the previous firing of this same neuron). The way that neurons encode an incoming sound is discussed in more detail in the next section.

By describing a sound as consisting of many different local events, a system can analyze the incoming sound by finding relationships between the different parts. It is this similarity in the features of the different events that implies a causal relationship between them. It is this relationship that binds the events

together, which allows them to be interpreted as coming from the same sound source.

Events are not directly assigned to a sound stream. Events with similar properties are joined together into an intermediate representation called a group object. It is these group objects that are assigned to one sound stream or another. The reason for the intermediate representation is that the neural events do not act independently, but act as a cohesive unit.

The presence of an intermediate representation can be seen in several different experiments. In Bregman's streaming experiments (see section 2.3.1.3), pure tones are assigned to either one sound stream or another. A tone is never split where one part of the tone is assigned to one sound stream and another part is assigned to a different sound stream. Each tone behaves as a cohesive 'unit' or 'group' where all its events (which result from the response of the cochlear model to that tone) are assigned to the same sound stream.

The example of a male speaker saying the digit /three/ followed by a female saying the digit /seven/ was discussed in section 2.3.1.6. When the auditory system hears the male vowel /ee/ from the digit /three/, all the events are periodic and form a repeated structure. After the vowel /ee/, the auditory system encounters an onset of fricated energy and a series of events that have no periodic structure to them. These incoming nonperiodic events are grouped together and this group object (whose phonemic representation is an /s/) of nonperiodic events is initially assigned to the same sound stream as the other sounds of the same speaker. When the periodic segment /even/ from the female digit /seven/ is heard, the auditory system changes the assignment of the group object /s/ from the male speaker to the female speaker. The assignment of events (which represent the fricated energy of the /s/) to one sound stream or the other are manipulated as a unit and not individually.

At the highest level of a sound's representation is the sound stream. A sound stream is the internal representation that corresponds to a sound source that humans hear. Group objects are assigned to a sound stream if the auditory system concludes that the group object represents a sound that emanated from

the appropriate source.

2.4.2 The Neural Encoding of Sounds

The transformation of sounds into a neural representation occurs through a series of complex mechanisms. Sound is transformed from pressure variations in the air into mechanical motion at the eardrum (tympanic membrane). The vibration of this membrane moves several small bones in the middle ear cavity. These bones are also attached to the oval window of a fluid-filled chamber (the cochlea or inner ear). The vibrations of this oval window causes the fluid in the cochlea to move, which in turn causes the motion of another membrane (the basilar membrane). Attached to the basilar membrane is an intricate structure of cells (the hair cells of the organ of corti) which are in turn connected to the neurons that encode the incoming sound into neural firings. [Yost & Nielsen 1977].

One end of the basilar membrane responds best (large displacements in the membrane) to high frequency stimuli, while the other end of the basilar membrane responds mostly to low frequency stimuli. The 'place' along the length of the basilar membrane is an important dimension, closely related to the frequency that causes the maximum displacement of the membrane. Attached along the length of the basilar membrane are approximately 30,000 neurons [Chow 1951] which encode the incoming sound.

The length along the basilar membrane is often called the *place* dimension. Neurons along the length of the basilar membrane are organized in a 'tonotopic' manner (i.e. with place mapping to tone frequency). The place dimension along the basilar membrane is preserved through many auditory regions in the central nervous system. "It is unlikely that place along the basilar membrane would be preserved through successive levels of central processing if it were not an important parameter of the internal representation of sound." [Young & Sachs 1979]

At low stimulus intensities, auditory neurons do not increase their neural firing rate above their spontaneous level (the firing rate with no signal present). Rather, they tend to synchronize the spontaneous neural firings with the motion of that place on the basilar membrane [Johnson 1980]. As the sound's intensity

increases, the neurons increase both the synchrony of the firings and the firing rate. Above a certain amplitude level, a neuron will reach a maximum in the 'synchronization index' (degree of phase-locking of neural firing with stimulus) and will approach its maximum rate of neural firing.

The intensity of a tone required to make a neuron change its firing pattern from the spontaneous firing pattern is measured as a function of the tone frequency and is plotted as a 'tuning curve'. Each neuron has a 'characteristic frequency', which is the frequency for which the least amplitude is required to change its firing pattern. The high frequency slope of a tuning curve (the frequency side above the characteristic frequency) is typically 100 to 400 dB/octave; the low frequency side of the tuning curve will typically flatten out at a level approximately 40 dB above the neural threshold [Sachs & Abbas 1974].

The temporal fine structure of a stimulus is maintained in the phase-locking of neurons to the sound [Javel 1980, Rose et al 1971]. Period histograms (the number of neural firings in each time increment of a periodic stimulus) shows a highly significant correlation between the positive amplitude of the stimulus and the number of neural firings recorded at that time.

The observed response of neurons to periodic steady-state vowels [Young & Sachs 1979] can be understood as follows: at low levels, the spectral shape of the vowel can be discerned from either the average firing rate of an array of fibers or from the synchronized rate. At higher levels, the fibers have saturated, and the average rate will no longer yield the spectral information. The synchronization of neural firings with a vowel's harmonics can be used to reveal the spectral shape, even in spite of interfering random noise, whereas average rate will not.

For steady-state vowels whose intensity is less than 60 dB SPL, those neurons whose characteristic frequencies are close to the formants of the vowel will have a firing rate which is greater than other neurons. The spectrum of the vowel can be characterized by a profile of the neural firing rate as a function of place. However, for sounds whose intensity is much above 60 dB SPL, the firing rate profile saturates at the maximum firing rate. Average firing rate no longer reflects the spectrum of the incoming vowel. The phase-locking of the neurons to the vowel

harmonics, however, is maintained at high stimulus levels [Sachs & Young 1980].

These experimental results "indicate that the spectrum of a sound is not conveyed to higher nervous centers of the auditory system by way of the (average) discharge rate in different nerve fibers. It is more likely that such information is carried in the time pattern of the discharges of single auditory nerve fibers" [Moller 1981]. Voigt, Sachs, and Young [1981] also concluded that "The temporal-place representation of vowel spectra is superior to the rate-place representation at moderate to high vowel levels. In addition it retains information about vowel spectra in the presence of background noise. The rate-place representation does not reflect the formant structure of the vowel even at moderate signal-to-noise ratios."

Based on this understanding of the neural encoding of an incoming sound, it is possible to understand why Scheffers [1982] observed that people could identify two simultaneous voiced vowels (even though they were generated with the same fundamental frequency) better than two simultaneous unvoiced (whispered) vowels. For the unvoiced vowels, the auditory system cannot accurately encode the spectrum of the two vowels in terms of neural firing rate. It can encode the simultaneous voiced vowels with more accuracy, since the neurons use timing information to convey spectral information.

Our knowledge about the neural encoding of a sound has been gained through a large number of experiments done over many years. However, our knowledge of how the auditory system uses this information is minimal [Moller 1979]. This lack of information about the central nervous system makes it impossible to accurately model any detailed mechanisms that the auditory system might use in the processing of sounds.

Modern spectral analysis techniques may accurately measure the spectral amplitude of speech sounds, but the auditory system relies heavily on timing information to encode the speech. At moderate levels, it does not rely on rate information, but uses the timing of the neural firings to represent the sounds that it hears. The next section will deal with how the auditory system uses this timing information for the interpretation and separation of sounds.

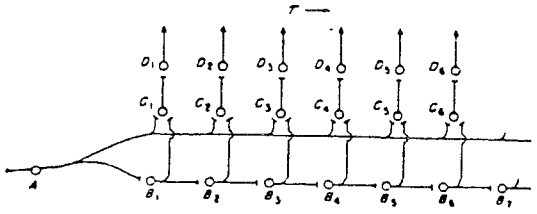

Figure 2.4: Licklider's neural autocorrelator. The original signal is autocorrelated with the delayed signal, and the output is a running integral.

2.4.3 The Processing of Periodicity Information

This section is concerned with how the auditory system computes the period of repetition of a periodic sound. A previous section showed that information about pitch and pitch dynamics is useful for separating sounds. This section will review the mechanisms that the auditory system uses for computing pitch information, and how it uses the information about periodicity for separating sounds.

There are several major theories of how the auditory system perceives pitch. This section will review the theories of Licklider and Goldstein in their models of auditory pitch detection. It will review the experiments which show how Licklider's model is consistent with the use of timing information by the auditory system. It will also show how this author has extended Licklider's model so that this periodic information can be used for the separation of sounds.

In Licklider's theory of pitch perception [Licklider 1951, 1959], the output of each 'place' along the basilar membrane is passed through a neural autocorrelator mechanism. The neural output of a single place is passed through a tapped delay line, and at each tap computes one value of the autocorrelation function of the delay-line output with the current neural output of the same neuron. A diagram of this computation is shown in figures 2.4 and 2.5.

The major difficulty with this theory is that it does not specify the details of how the autocorrelation information in each place location is combined to determine the pitch period. It suggests that a neural net interprets the incoming

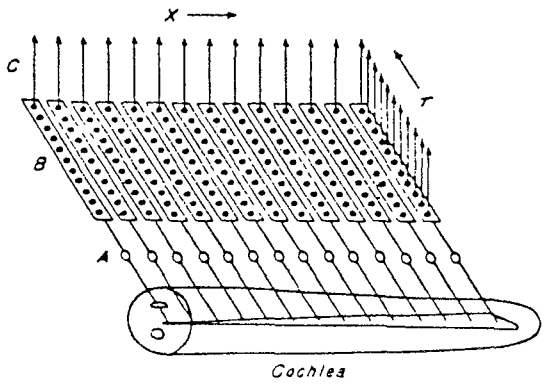

Figure 2.5: Overall autocorrelation mechanism. The autocorrelation of the cochlear output is computed at each place location. The dimensions of this representation are place vs. autocorrelation delay

information to determine the pitch period, but the theory does not specify the algorithms used.

Goldstein's model of pitch perception [Goldstein 1973] estimates the frequencies of the harmonics (obtained from the spectrum analysis) and computes a maximum likelihood estimate of the fundamental frequency from these harmonic values. The constraints that this theory imposes are nicely summarized by De Boer [1975] as follows:

- Only aurally *resolved* components contribute.
- The *phase* relations are irrelevant.
- Only the *presence* of a component is reported - the amplitude is irrelevant (within limits).
- The information about component frequency is basically *inaccurate*; a non-negligible variability is involved.
- It is assumed that the information received corresponds to stimuli in which the components are *successive harmonics*.

There have been several implementations of Goldstein's theory of pitch perception [Willems 1983, Duifhuis et al 1982, Scheffers 1983, Allik et al 1984]. The difficulty with all these implementations of the theory is that they compute the frequencies of the harmonic components from the spectrum using the amplitude of the spectrum along the place dimension. As we have seen from the section on the neural representation of information, the auditory system does not encode the amplitude of a vowel's harmonics in the average firing rate information at moderate to high intensity levels.

Instead of obtaining the harmonic frequencies from the amplitude spectrum, Goldstein [1977] has suggested that the harmonic frequencies are obtained from interspike interval histograms. The timing information at each place location would be used to compute the harmonic frequencies present. The use of timing information in Goldstein's theory of pitch perception makes this theory similar to Licklider's theory of pitch perception. One difference is how the information from different basilar membrane place locations are combined to form an estimate of the pitch period present. Goldstein's theory imposes constraints that the harmonics must be resolvable and must be successive harmonics in order to compute the pitch period.

There is a great deal of experimental evidence which supports the idea that the timing information at each place location is used to compute the period of repetition of the sound. Experiments on the perception of a pitch period for amplitude modulated noise [Houtsma et al 1980, Patterson et al 1977] support the temporal processing of information for the computation of periodic information. Other experiments [Moore 1977] show that the harmonics do not have to be resolvable (as required by Goldstein's theory) for the auditory system to compute the pitch of a harmonic complex. These experiments also show that the relative phase between the unresolved harmonics changes the strength of the pitch percept (as predicted by the temporal theory). The upper frequency limit to the perception of 'residue pitch' of roughly 2500 Hz [Wightman 1973] is explained by the decrease of synchronous firing in the auditory system at these frequencies.

In addition to these psychoacoustic experiments which support the use of

temporal information in the computation of pitch, there is also evidence for Licklider's neural autocorrelation mechanism. Experiments [Langner 1981, Rose & Capranica 1984] show that many neurons are tuned to the perception of amplitude modulated tones.

Licklider's model can be extended to use the periodicity information for the separation of sounds. The computation of an autocorrelation function at each place location has the advantage that the periodic information is a local computation. Therefore, every 'place' can provide information about the periodic information that is present at that location. In Broadbent and Ladefoged's experiments [1957], when two formant resonators are excited by pulse trains with different periods of repetition, the place location that phase locks onto the first formant has a different autocorrelation function from the place location that phase locks onto the second formant. It is hypothesized that this difference in the autocorrelation functions is used by the auditory system to determine that these formants are not generated by the same sound source.

The autocorrelation model also allows for the separation of the higher harmonics *not resolved* by the auditory system. These place locations, (which phase lock onto the AM waveform generated by the higher harmonics), compute an autocorrelation function that shows a peak at the pitch period just like the autocorrelation of the lower harmonics. It is this peak at the pitch period of each place location that is used by the auditory system to separate the incoming sound based on periodicity information.

Research also focuses on the capabilities of the auditory system to separate and recognize two simultaneous vowels with different fundamentals. Experiments by Scheffers [1979, 1982], and Brokx and Nooteboom [1982] chart recognition performance as a function of the difference in pitch between the two vowels. The result of these experiments [Scheffers 1982] indicates that separation and subsequent recognition performance is significantly better for simultaneous vowels with strongly different spectral envelopes than for vowels with relatively similar spectral shapes. However, mo ·ls of sound separation based on pitch mechanisms that assume fine frequency re· ıtion [Goldstein 1973, Parsons 1976] cannot account

for this effect. These methods rely on the resolution of the harmonic components present in the frequency spectrum and the subsequent assignment of these components to the different sound sources. Whenever it is possible to resolve a harmonic component, changing the amplitudes of the harmonics of the other sound will have little effect on the resolution of that harmonic component. Therefore, harmonic selection (resolving the different components by performing a fine frequency analysis, and then computing the amplitude of each of the harmonic components of the periodic sound) cannot account for the observed degredation in performance by the human auditory system when the vowels have a similar spectral shape. This result supports the view that, for vowel recognition, the human auditory system does not perform fine frequency resolution of the harmonic components present.

This section has argued that the auditory system uses the timing information present at each place location to compute an autocorrelation function. In this autocorrelation function is the periodic information present at each place location. Licklider's model does not precisely specify how this information is combined in the auditory system for the determination of pitch.

Each place location that has similar periodicity information can be interpreted by the auditory system as coming from the same sound source. Those place locations that have incompatible autocorrelation functions are interpreted as coming from different sound sources. The next section will discuss how the auditory system uses this periodicity information to combine all the place locations that belong together into the same 'group object'.

2.4.4 The Segmentation of Speech and Group Objects

A group object is a collection of events which, because of similar properties, are perceived as a unit. It is an intermediate level in the representation of sounds and corresponds to the natural segmentation of the incoming sound into frequency-time regions that have similar properties. Group objects assigned to the same sound stream are thought to originate from the same sound source.

Researchers have recently begun to focus on the grouping of the visual perceptual field. Pomerantz [1981] says that "The purpose of grouping is to divide

the perceptual field into units, but what exactly is a unit? ... Any natural unit is defined by its indivisibility. Seldom is this indivisibility absolute, as the unending search for the absolute, fundamental particle in physics well attests. Nonetheless, when a complex structure is broken down into parts, some breakpoints are more likely than others, and these serve to demarcate natural units."

Research is currently exploring the possibilities that there is some innate mechanism for dividing the acoustic flow into discrete segments [Chistovich et al. 1975]. The segmentation of an acoustic input into discrete segments that can be assigned to sound streams is the mechanism used by the auditory system to separate sounds [Broadbent 1977].

At the level of grouping events into group objects, there is no notion of how many sounds are present. Events with similar features are grouped together. After these events have been grouped together into a group object, these objects are then assigned to sound streams based on which sound source is believed to have created these events. Two reasons why events cannot be directly linked to a sound stream will now be presented.

1. The decision concerning which sound stream a segment belongs to can change, and this change reverses the decision on all the events in the group object. In the 'three-seven' example discussed earlier, all the neural events that compose the frication sound /s/ were assigned as a unit to one sound stream, and then later to a different sound stream.

2. Events cannot be directly assigned to a sound stream because a sound stream does not have a single set of properties to which an event can be linked. Speech is composed of short segments that have different properties, such as plosion, frication, periodic regions, and silent intervals between speech segments. Since the characteristics of a speaker's voice will change between the different types of acoustic segments, the properties of the sound stream that represents this speaker will change. It is hypothesized that the auditory system first groups those events in frequency-time regions with similar properties together, and then makes a decision about which sound source this group object belongs to.

Events are grouped into an intermediate representation called group objects, which, in turn, are linked to a sound stream. Acoustic information is used to assign events to group objects. Information about pitch, pitch dynamics, onset, offset, and amplitude modulation are used to assign events to the different group objects. The specific details about how the auditory system creates and uses group objects may not be known for many years. The next three sections will deal with issues and details that are important for the construction of a model of the auditory separation system.

2.4.4.1 Creation of a Group Object

A group object that represents a speech segment extends across both frequency and time. Many details about the creation and use of group objects are not known at the current time. A group object is defined as 'a collection of events with *similar* properties that are perceived as a unit'. The key word in the above definition is 'similar'. What constitutes events that are similar? How does one differentiate between events that are similar and those that are different?

Darwin's experiments [1977] show that if pitch changes discontinuously, each segment with a sufficiently different pitch is assigned to a different sound stream. Therefore, if the events undergo a pitch discontinuity, a new group object is formed. Experiments by Rasch [1978] show that the auditory system will interpret two onsets at different times as two different group objects.

A new group object is created in the following circumstances:

- At the onset of a new segment.
- When the incoming neural events have different properties from any existing group object.

It is difficult to decide group object boundaries. The following difficulties are described along with a proposed solution:

1. A tone is masked during the middle of its duration by a loud masking noise. Does the auditory system represent the parts of the tone before the noise

and after the noise by a single group object or by two group objects? It is hypothesized that the auditory system uses two distinct group objects to represent the segments of the tone. These group objects can then be assigned to the same sound stream or to different sound streams.

2. A periodic segment undergoes a spectral discontinuity (such as a vowel-nasal boundary) but the pitch contour remains continuous. Does the spectral discontinuity cause the input to be parsed into separate group objects? It is hypothesized that the continuity in the pitch dimension is the important factor and will therefore not allow the different regions in time to be assigned to different sound streams even though there is a spectral discontinuity. There may be a phonetic boundary that is perceived at the spectral discontinuity, but for the assignment of segments to sound sources, no boundary exists at the spectral discontinuity.

3. There is a discontinuity in the slope of the pitch conour of a voiced segment. Does a change in pitch dynamics cause the formation of a new group object? It is hypothesized that if the pitch contour is continuous, no segmentation at the sound separation level occurs. The change in the slope of the pitch of a vowel might perceptually segment the two regions but the different periodic regions (with different pitch slopes) will not be assigned to different sound sources.

The higher levels of auditory processing may influence the creation and interpretation of the group objects. If a weak onset occurs during the presence of one sound source, the auditory system may not be sure whether this onset is a random fluctuation from the sound already present or whether it constitutes the beginning of another sound source. The higher levels of processing can influence the decision of when to create a new group object, and can make a difficult situation easier by using more than just the acoustic information present.

2.4.4.2 Simultaneous versus Sequential Grouping

Events across time at the same frequency that have similar properties are assigned to the same group object. Events at different frequencies at the same time which have similar properties are assigned to the same group object. This section is concerned with what happens when the decisions of different information sources which group events at one frequency and time with a group object conflict.

Bregman's experiments [1978b] (described in section 2.3.1.3) show that each of three tones acts as a unit, and each tone is assigned as a unit to one sound stream or another. Experiments by Darwin [1984b] show that if a pure tone (whose frequency is the same as the harmonic of a neighboring steady state vowel) is followed by a steady state vowel, the vowel's harmonic will be perceptually segregated from the vowel. The vowel's harmonic will be a multiple of the same fundamental as all the other harmonics of the vowel. However the difference in onset or offset between this harmonic and the other harmonics indicates to the auditory system that this harmonic does not belong to the same sound stream. Since the auditory system assigns this harmonic to a different sound stream, it must not be assigned to the same group objects as the other harmonics.

These experiments might lead one to believe that the auditory system first assigns events from the same frequency region together, and then assigns different frequency locations that have the same onset, offset, amplitude modulation, and pitch dynamics to the same group object. The difficulties with this approach are:

1. Events at the same frequency location cannot be linked by spectral continuity alone. Experiments by Darwin [1977] show that when the pitch changes discontinuously, the periodic regions on either side of the pitch discontinuity are assigned to different sound streams. Therefore, events at one frequency that are linked through time must use other features, such as pitch, to assign them to the same group object.

2. Experiments mostly deal with the perception of simple sounds. It is very difficult to know how the auditory system links events through time when the spectrum is changing and the pitch is also changing. The auditory system

follows both formant motion (transitions) and pitch change (intonation) at the same time.

This problem of deciding what events at one time belong with to those at the next instant in time is known as the *correspondence problem* [Ullman 1979]. When a harmonic starts at a different time from the rest of the harmonics, if all the events which phase lock onto that harmonic are to be assigned to the same group object, the auditory system must maintain a correspondence through time of the neural events which are responding to this harmonic. It is a very difficult problem to maintain the correspondence from one time to the next of the neural representation of each of two sounds.

In the computational model of auditory sound separation (described in chapter 3), different frequency regions at the same time are assigned to the same group object if their instantaneous properties are all consistent with each other. Different frequency regions that have the same pitch period are assigned to the same group object. A difference in onset or offset of different harmonics will not affect the assignment of the harmonics during the middle of the vowel to the group object. As long as the pitch is continuous through time, the different frequency regions will be assigned to the same group object.

2.4.4.3 Filling in the Gaps

When the auditory system hears a sequence, such as a tone, a noise burst and then the same tone again, it perceives the tone to continue through the noise. When a collection of events (such as the noise burst) is assigned to a different sound stream, it leaves a gap in the other group object present. If the auditory system hears the tone continuing through the noise, it must perceptually synthesize the tone at some level. Does the auditory system perform this synthesis at the acoustic level of sound separation or at the higher levels of separation?

A series of experiments by Warren [1971, 1972, 1974] discussed the synthesis and perception of the missing phonemes which has been called 'phonemic restorations'. The auditory system cannot predict at an acoustic level what the missing phoneme is. The perceptual synthesis must therefore occur at the higher levels of

a sound's interpretation. The higher levels must have access to the representation of both sounds in order to know that the segment is masked rather than missing.

2.4.5 Sound Streams

Work on *selective attention* has dealt primarily with how a person focuses on one sound in the presence of other sounds. Some theories of attention are characterized by the 'early filtering' models (filtering here refers to the separation of one sound from other sounds) of Broadbent [1958] and of Treisman [1960]. The sounds are filtered or separated by focusing on different *functional* channels (e.g., an internal channel which represents the *location* of the desired sound source, *pitch* channel, etc.). In other theories of attention, such as the model of Deutsch and Deutsch [1963], the separation of sounds does not occur until late in the processing (at least the semantic level) of the sounds. All of these models are quite general, and lack specific details on how any of the different operations are performed.

The concept that two speakers could be separated from each other by focusing attention on the output of a functional channel cannot account for how two different speakers can be separated monaurally. Since speech is composed of periodic segments, nonperiodic segments, bursts, and periods of silence, one cannot focus one's attention on a single functional channel, since the sound from a single speaker will change from one channel to another. Phonetic and linguistic knowledge must aid the selection process to determine which segments belong to the same speech stream.

This section will study how the auditory system creates and uses sound streams to represent the different sources that it listens to. It will focus on the number of sounds the auditory system can process at a single time. It will also deal with how group objects are assigned to different sound streams.

2.4.5.1 Is There a Maximum Number of Auditory Sound Streams?

How many sound streams does the auditory system use in sound separation? Does the auditory system have one sound stream for each of the sound sources that are present, or is there one special sound stream which is the *figure*, while

all the other sounds are lumped into the *background* stream? How many sound streams are there when many sounds are present?

If there is one sound present, then only one sound stream is needed to represent this sound. If there are two sounds present, then there will be one sound stream to represent each of the sound sources present. However if there are more than two sounds present, does the number of sound streams increase beyond two? In the figure-ground approach, one sound stream represents the 'desired signal' (which is being focused on) and the other sound stream represents all the other sounds that are present. All acoustic events not assigned to the desired sound stream are put into the 'interfering' sound stream. The two sound stream model is attractive because one sound stream is labeled the desired signal, or figure, and the other sound stream represents the interfering sound, or the background.

Another option is that the auditory system can maintain more than two sound streams at a single time. The maximum number of sound streams would be limited by the processing resources of the auditory system. In this case, if three or more sounds are present and the auditory system has enough processing power (depending on the complexity of the sounds), then the acoustic events can be assigned to the appropriate sound streams that correspond to the sound sources that they have originated from.

At some level in the processing hierarchy, there may only be one sound source that is focused on. This view is held by Moray [1970] in his book on the selective nature of attention in speech and vision:

> "At any moment a listener is sampling only one message. All others are totally rejected." [Moray p. 190]

The fact that one message is being focused on (receiving special processing resources) does not imply that the number of sound streams present is limited to two.

If one is able to determine that the auditory system is capable of modeling many sounds at the same time, then this would imply that there are more than two sound streams present. The fact that many sounds could be modeled would not necessarily contradict the hypothesis that a single sound receives special processing

resources. This is because the special processing that this 'focused' signal receives can occur after the incoming sound is partitioned into the different sound streams.

Although most subjects who participate in shadowing experiments (where subjects are instructed to listen to one message and ignore the other messages present) are typically not able to report much about the unattended message, experimental results indicate that the unattended message is processed at a semantic level even when the subject cannot report the contents of the message. In the experiments by Von Wright et al. [1975], subjects were conditioned by pairing electrical shocks with certain words. The experimenters then recorded the galvanic skin response (GSR) of the subject during a shadowing experiment where no electric shocks were given. They found that the subjects showed a response to the conditioned word, as well as smaller responses to synonyms of that word and acoustically similar words. These results support the idea that even the unattended message is processed at a semantic level, even when a subject cannot report what he has heard. Results from an experienced subject in a shadowing task [Underwood 1974] indicate that a person is able to monitor and respond to two messages at the same time.

One can view Wright's experimental results as showing that each sound stream is processed until at least the semantic level which is similar to the view held by the Deutsch and Deutsch model of attention [1963], the Neisser model [1967], and the Shiffrin and Schneider's model [1977]. Each sound is represented in its own sound stream and is processed to some level by an automated parallel algorithm. This viewpoint does not mean that one of the messages will not receive special processing, but that each message receives separate processing to some level.

Other experimental evidence indicates that separation performance improves with knowledge about the interfering signal. Experiments by Hawkins and Presson [1975] showed that when a masker tone (a strong sine wave which makes it difficult to hear the other sound) was of a known frequency in auditory recognition masking experiments, the performance of a subject improved over experiments where the masker frequency was unknown. They concluded from their experimental results

that:

> "A selectivity process functions in the auditory system to diminish the effects of unwanted input prior to that point in the system at which categorization occurs."

The fact that subjects improved in their performance when certain facts were known about the masking stimulus indicates that subjects used a model of the noise to increase their level of performance in separating the signal from the masking stimulus.

Experiments by Triesman [1964] also indicate that subjects can model more than one sound and can use this knowledge to improve their performance in the separation of the desired message. In one experiment, subjects had to shadow one message in the presence of two interfering messages. The message to be shadowed appeared in one ear, while one interfering message was presented in the other ear, and the second interfering message was presented in both ears. The content of the interfering messages was varied, and the effect on the shadowing performance was measured. Subjects showed slight improvements when both interfering messages were sequences of ascending digits, over the case when both interfering messages were prose. This result supports the view that both interfering sounds were modeled to some depth, and that the information from these models was used to improve the separation performance.

Evidence has been presented that at least two simultaneous sounds can be modeled by the auditory system. Experimental results have also been presented to show that models of the interfering sounds can be used to increase separation performance. At some higher level of a sound's processing, one sound stream might receive special attention at the expense of the other sounds present. At the lowest level of processing, the auditory system can model more than one sound source. It would be difficult for one to claim that one sound stream has a special advantage over the other sound streams. Although the figure-ground paradigm may be appropriate for the higher levels of sound understanding, it is not clear that the figure-ground analogy applies at the lowest levels of acoustic separation.

When a person is allowed to listen repeatedly to the same sound segment over and over again, he is capable of modeling many sound sources at the same time. This improvement in processing ability is facilitated both by additional processing resources (on each pass of the recording) and by having a model of some of the other sounds in the recording when listening to one of the other sounds present. The auditory system is able to hear three sound sources at a time if the sounds are simple enough and do not require much linguistic processing (such as a person speaking, air conditioning noise, and the ringing of the telephone). The ability to hear many sounds (either in repeated listening, or if the sounds are simple and repetitive in nature) led this writer to believe that the auditory system can create more than two sound streams at a time.

2.4.5.2 The Creation of Sound Streams

When does the auditory system decide that there is more than one sound present? There are two situations when the auditory system can decide if two sounds are present. If two *simultaneous* group objects are present (such as two simultaneous periodic sounds), the auditory system will assign each group object to a different sound stream. If two group objects are sequential in time (one follows another), and the auditory system cannot account for both group objects with a single sound model (a single sound source could not have generated these sounds), the auditory system will assign each group object to a different sound stream. An example of this would be a person speaking followed by the sound of a door closing. This determination that the two group objects could not have come from the same sound source uses linguistic or other contextual information about what sounds could be generated from what types of sound sources.

2.4.5.3 The Assignment of Group Objects to Sound Streams

After the events have been assigned to different group objects, the group objects are assigned to sound streams. The assignment of group objects to sound streams uses the information sources discussed in the beginning of this chapter. The ways that group objects are assigned to sound streams are summarized below:

Chapter 3

A Computational Model of Sound Separation

The construction of a computational model of auditory processing would be fairly trivial if we knew what operations were performed by the auditory system. The computer algorithms described in this chapter are only estimates of the actual algorithms used, since we do not know the precise details of how the auditory system operates. In the absence of exact knowledge, it is extremely difficult to determine from all of the possible mechanisms that could account for the auditory system's behavior, which one the auditory system actually uses.

This chapter will describe the computer model that was developed to separate two simultaneous talkers. The model is based on the theory of auditory separation described in chapter two. The construction of this detailed model has raised many questions and issues, and has helped to evolve both an understanding of what the auditory system is trying to accomplish, and how the system accomplishes its processing.

3.1 Cochlear Filtering, Compression, Detection

The input to the sound separation algorithms is a computer model of cochlear processing developed by Lyon (1982). In the cochlear model, an incoming sound signal (that is sampled at 16 khz) is filtered by an 85 channel filterbank. The filterbank, originally a series of second order canonic sections organized in a

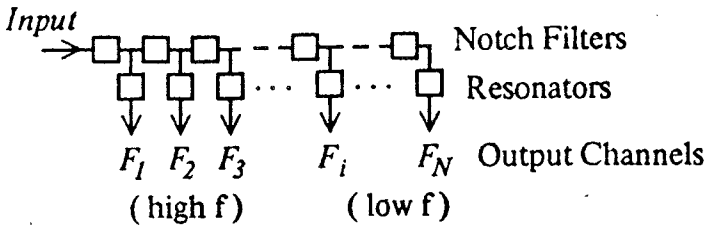

Figure 3.1: Schematic diagram of the cascade-parallel filterbank

cascade-parallel form, is shown in figure 3.1.

By rearranging the poles, the filterbank has recently been reorganized as a cascade only form [Lyon 1984]. The use of a cascade-form filterbank allows for extremely rapid high frequency rolloff (greater than 200 db/octave) with a minimal amount of computation. The transfer function of each filterbank output resembles the shape of an auditory neuron's tuning curves [Sachs & Abbas 1974]. The bandwidth of a frequency channel's output was chosen to match the measurements of critical bands in the auditory system [Zwicker 1962]. The spacing between the center frequencies of each pair of filter sections is a parameter of the model, and can be set depending on how many channels are desired (the current spacing is one quarter of the frequency channel's bandwidth, or approximately one twelfth of an octave at high frequencies). Each filterbank output is maintained at the full sampling rate (16 khz).

The amplitude of an incoming sound signal can vary over many orders of magnitude. To compress this tremendous dynamic range of the input, the output of each filterbank is then processed through a coupled automatic gain control (AGC) mechanism [Lyon 1982, 1984]. The adaptive mechanisms of the peripheral auditory system are functionally modeled by several stages of AGC with different time constants at each stage. The four stages of AGC in the computer model have time constants of 640 msec, 160 msec, 40 msec, and 10 msec. The outermost AGC mechanisms have the longest time constant to adjust the overall sound level, while the innermost AGC loops have the shortest time constants for compressing fluctuations on a smaller time scale.

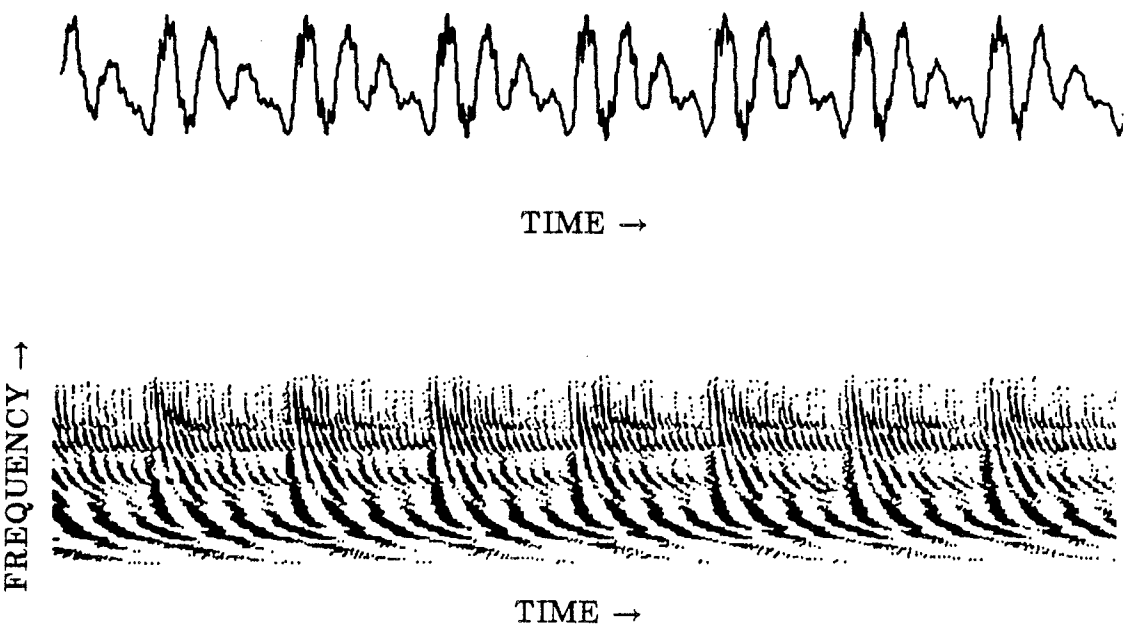

Figure 3.2: Top: A segment of a periodic speech waveform. Bottom: The output of the cochlear model. The waveform is time aligned with the cochlear output.

Since auditory neurons respond only when the basilar membrane moves towards the scala-media, the compressed filterbank output is passed through a half wave rectifier before the neural encoding stage.

An example of the output of the cochlear model are shown in figure 3.2. In this picture, the amplitude of the positive output of each frequency channel is represented by the degree of blackness.

The output of the cochlear model then, is 85 frequency channels, with each filterbank output remaining at the original sample rate of 16 khz. The output is kept at a high sampling rate to preserve the information present in the **fine time structure** of the output. This fine time structure will be used by the separation algorithms. By contrast, most other filterbank designs are concerned only with the envelope, or short-term average level of each frequency channel's output.

3.2 Event Representation

The goal of the neural encoding of a sound is to preserve the timing and intensity information in the output of the cochlear model. Rather than model the neural encoding by a stochastic process that uses many neurons to probabilistically encode the amplitude and timing information of a signal, a deterministic neuron model is used. The output of each frequency channel of the cochlear model (which is the compressed and half wave rectified filterbank output) consists of a series of positive waveform peaks. Each waveform peak looks approximately like the positive half of a sine wave, since each frequency channel's output (before rectification) is a fairly narrow-band signal. The positive waveform in each frequency channel (between zero crossings) is encoded as a single **event**. The location of the event corresponds to the local peak in the channel's output. The amplitude of the waveform at the peak location and the area under the waveform between zero crossings are stored along with the peak time as properties of the neural event.

Since the computer model's event encoding is not a probabilistic model of the neural firings in the auditory system, it does encode the timing and intensity information present in the cochlear model's output into an event representation. There are two important differences between this event encoding and a probabilistic neural model. One difference is that the event encoding output resembles the behavior of an array of neural fibers. A neural model requires many different nerve fibers to adequately encode the intensity of an incoming signal while the event encoding used in this computer model explicitly represents the area and amplitude of the cochlear output. The second difference is that this event encoding does not have a 'refractory period' (minimum time between neural firings) and will encode all the peaks in each frequency channel, while a neural model only fires at a rate below some maximum rate and decreases the synchrony of its firings with frequencies above 1 khz. The event encoding used in this model has the advantage of representing the timing and intensity information present in each frequency channel with little computational effort.

An example of the event encoding of the simulated cochlear output is shown in figure 3.3. The amplitude and area features that are associated with each

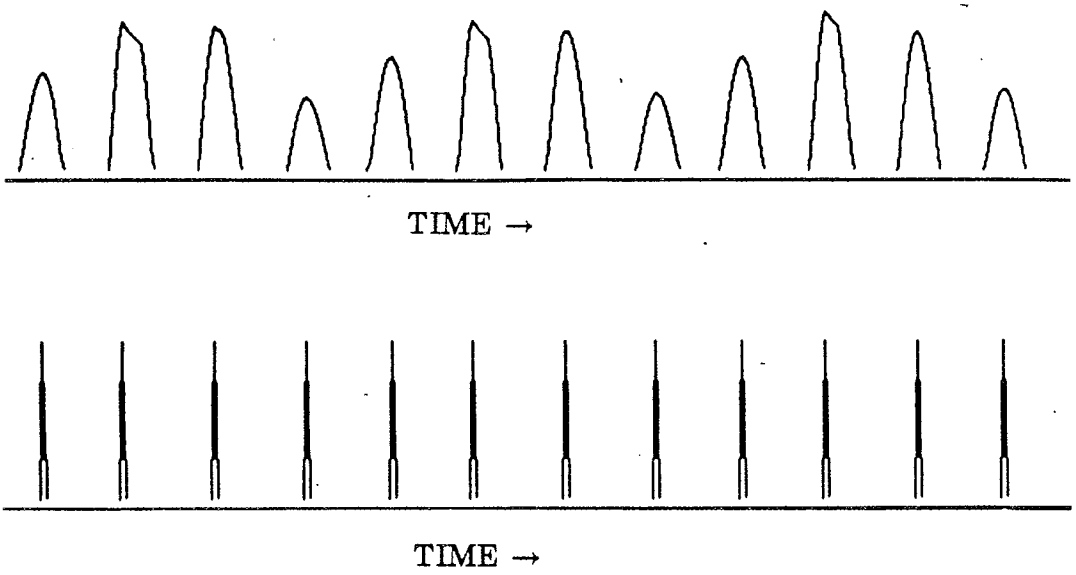

Figure 3.3: Top: The cochlear model's output for a single frequency channel. Bottom: The event representation of that frequency channel. The properties of each event (amplitude and area of the corresponding waveform peak) are not shown.

event are not shown in this figure. Figure 3.4 shows the transformation of the cochleagram into an event cochleagram.

3.3 The Computation of Periodicity and the Coincidence Representation

Periodicity is an important information cue that is used by the auditory system for separating sounds. Those neural events that have similar periodicity features can be viewed as coming from the same sound source. The computation of a local periodicity feature in this model is based on Licklider's theory [1951] of pitch processing in the auditory system. According to his theory, a neural structure computes an ongoing short-time autocorrelation function of each frequency channel's output. The neural structure delays each channel's output through a tapped delay line, and at each tap detects the coincidence of a pulse at the delay-line output with an undelayed pulse. A lowpass filter on each coin-

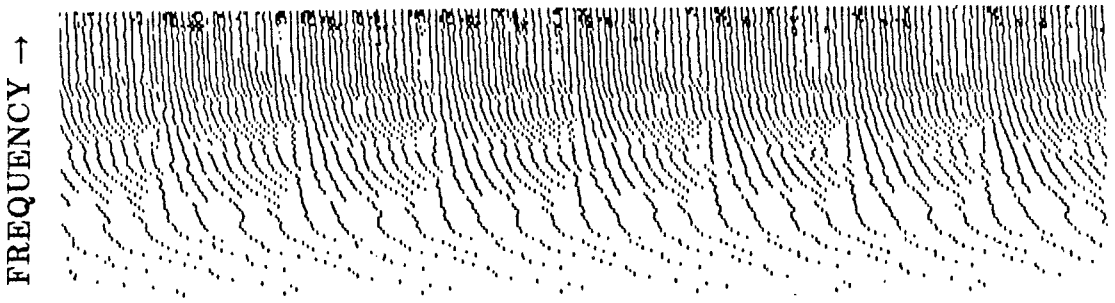

Figure 3.4: An event cochleagram representation

cidence detector output generates an equivalent time window for the short-time autocorrelation function. At a single time instant, the autocorrelation function is a two-dimensional array, parameterized by cochlear place (frequency), and by the delay parameter (repetition period). In Licklider's theory, this entire array representation is then interpreted through an unspecified neural network to determine the pitch period. His theory was developed to explain how the auditory system computes the pitch of an incoming sound, and was not originally intended as a mechanism for sound separation.

This approach is consistent with the evidence on the neural encoding of sounds presented in chapter two. Other theories of auditory pitch perception [Goldstein 1973, Wightman 1973] are mathematical in nature, and are not presented in a form that can easily use the neural encoding of sounds to compute pitch information.

In Licklider's theory [Licklider 1951, 1959], the output of each 'place' along the basilar membrane is passed through a neural autocorrelator mechanism. The details of how this neural autocorrelator works are not given. Another difficulty in implementing the theory is that there are no details for how the periodicity evidence in different place locations is combined to determine the pitch of the

incoming sound. Due to our lack of knowledge about how the auditory system computes the coincidence of neural events (for the purpose of periodicity computation), the way that the events are manipulated remains an open question.

An autocorrelation type of function, called the **coincidence function** (which computes the coinciding of neural events with previous neural events in a delay line) is computed for the event representations of each frequency channel as a measure of the periodicity of the events in that frequency channel. The coincidence of each neural event with previous neural events represents the periodicity information about this event. When an incoming event is detected, the **coincidence** (see Table 3.1 for a definition of coincidence) of this event with all other neural events from this frequency channel within the last 24 msec is computed, and this periodicity information becomes a property of this event. Every 10 msec, a weighted average of the coincidence function of all recent events is computed. Each event's coincidence function is weighted by an exponential factor (time constant = 15 msec) which depends on the time of the event and the time of the current average coincidence function. The channel's coincidence function therefore represents an average of the coincidence information of all recent events in that frequency channel's output.

How are the properties of two events combined to compute the coincidence of two events? During the course of this research, several different formulas for combining the properties of two events were used. Of the four formulas that were used, the last formula in table 3.1 is the version currently used. The reasons for developing each coincidence formula are discussed below.

The value of an autocorrelation function at zero delay is equal to the total energy of the signal being autocorrelated. The first coincidence formula was chosen so that the coincidence function of an event with itself is proportional to the energy of the event (by a factor of $\frac{\pi}{8}$ when the positive peak of the cochlear model's output is approximated by a sinusoid). The value of the average coincidence function for each frequency channel at zero delay could therefore be used as an estimate of the energy in this frequency region.

The frequency spectrum that was provided by this mechanism seemed some-

Version	$C(event_i, event_j)$	$C(event_i, event_i)$	$C(event_i, \frac{1}{2}event_i)$
1	$\sqrt{area_i \times amp_i} \times \sqrt{area_j \times amp_j}$	$area_i \times amp_i$	$\frac{1}{2}area_i \times amp_i$
2	$\sqrt{area_i} \times \sqrt{area_j}$	$area_i$	$\frac{1}{\sqrt{2}}area_i$
3	$avg(area_i, area_j) \times \left(\frac{min(area_i, area_j)}{max(area_i, area_j)}\right)$	$area_i$	$\frac{3}{8}area_i$
4	$avg(area_i, area_j) \times \left(\frac{min(area_i, area_j)}{max(area_i, area_j)}\right)^2$	$area_i$	$\frac{3}{16}area_i$

$C(event_i, event_j)$ is how the properties of $event_i$ and $event_j$ are combined for the computation of the $event_i$'s coincidence function.

Table 3.1: Different ways that the properties of two events can be combined

what flat, and formant peaks were not as sharp as one would like. The reason why this first coincidence formula flatted the formant shape is that the sharp onset at the beginning of a pitch period (which is larger than the cochleagram of the rest of the pitch period), dominated the value of the coincidence function at zero delay. When looking at a picture of a cochleagram, the location of the formant becomes clear during the latter part of the pitch period (after the onset has had a chance to resonate and decay at the formant locations). The second formula for computing the coincidence of events was developed so that the value of the coincidence function at zero delay would be proportional to the average value of the channel's output. The peaks in the frequency representation using this second version were sharper than those obtained using the first version.

In the second version of the coincidence formula the area of an event was used (as opposed to using the amplitude of an event). If the amplitude of an event were used, then two different frequency channels with the same amplitude but different event rates would have different values of the average coincidence function at zero delay. The frequency channel with a higher event rate would have a larger value of the coincidence function at zero delay. By using the area of an event when 'coinciding' two events, frequency channels with different event rates but similar output levels will have similar values in the coincidence function at zero delay.

It was difficult to locate the pitch period in a frequency channel using the first and second formulas. A periodic sound has a peak in the coincidence function at the location of the pitch period. Although this peak was present using versions one and two, many other peaks in the coincidence function were nearly as large. After a careful study of the problem, it was determined that the output of a cochlear filter, which was amplitude modulated, would not form a sharp peak in the coincidence function at the pitch period. Neither of these two versions preserved the modulation depth present in the original cochlear output. Thus, a small amount of amplitude modulation present in the cochlear output would not be preserved in the shape of the coincidence function. A requirement that the modulation present in the cochleagram be preserved in the coincidence function (along with the requirement that the value of the coincidence function at zero be proportional to the amplitude of the cochlear output) led to the development of the third coincidence formula. The first term in this formula lets stronger events influence the coincidence function more than weaker events, and the second term emphasizes the differences in the amplitudes of the two events.

After some use with version three, it seemed there was no reason the amplitude modulation depth needed to be faithfully preserved in the coincidence function. In other words, the modulation depth present at the output of the cochlear filterbank could be increased so that the average coincidence function contained a greater modulation depth than the original cochlear waveform. Version four (the current formula used) enhances the modulation present in the cochlear output to form a sharper peak in the coincidence function.

When deciding how much modulation depth to use in the coincidence function relative to the modulation depth of the original frequency channel output, a tradeoff occurs between (1) emphasizing the amplitude variations in the events, and (2) maintaining the ability to compute the pitch period of a periodic signal. If amplitude changes are emphasized too much, then slight variations in the periodic signal over successive pitch periods will yield very low values for the channel's coincidence function except at zero delay. If amplitude changes are not emphasized enough, the system will not be able to differentiate between the peak in the

coincidence function at the pitch period and the peak at some other location. No quantitative testing was performed to determine how much enhancement of differences in the area of events is optimal to the performance of the pitch computation.

While there is some qualitative evidence on the relationship between a signal's modulation depth and the perceived modulation depth [Mathes and Miller 1947], there is no quantitative evidence on this relationship. More detailed information is needed to determine how the auditory system uses the amplitude modulation output of each frequency channel to compute periodicity information.

There are two advantages for using coincidence formula 4 over using an autocorrelation function to compute periodicity information of a frequency channel's output. These advantages are:

1. $Coincidence(event_a, event_a) \geq Coincidence(event_a, event_b) \quad \forall event_a, event_b$

   The importance of this requirement can be seen in segments where the amplitude of a steady-state periodic segment is decreasing, (i.e., where each repetition of a periodic waveform is successively lower in amplitude). The autocorrelation function of this signal when the overall amplitude is very small will show larger peaks at multiples of the pitch period than at the pitch period, while the value of the average coincidence function at the pitch period will always be larger than the value at multiples of the pitch period. (See appendix one for details.)

2. The coincidence function enhances the modulation depth, the autocorrelation function decreases it. Thus, if an incoming waveform has a certain modulation depth, the autocorrelation function of this signal will have a lower modulation depth to its shape, but the coincidence function will have a greater modulation depth. (See appendix one for details.) By increasing modulation depth in the coincidence function, it is possible to determine the pitch period in each frequency channel from the amplitude modulated cochlear output.

The foregoing discussion makes it apparent that the amplitude modulation

present in the cochlear model's output is very useful for determining the pitch period of the incoming sound signal. This amplitude modulation is a result of the finite bandwidth of each filterbank channel, since the output of each frequency channel is influenced by the adjacent lower-frequency harmonics that are not totally suppressed in the filtering. Thus the sharp high frequency cutoff and the slower low frequency rolloff in the cochlear transfer function result in an amplitude modulated output waveform, and this amplitude modulation in each frequency channel is encoded by the computer model and used for the determination of the pitch of a periodic sound.

Another tradeoff present is the choice of the time constant used in the decay of the recent event's coincidence function (current value is 15 msec). A small time constant allows the average coincidence function of all events to follow rapid changes in a periodic signal, but results in pitch-synchronous activity in the averaged output: sampling the output every 10 msec (100 Hz frame rate) will alias the pitch synchronous information, causing 'beating', if the time constant is too short. A longer time constant will result in smoother transitions between successive coincidence functions (every 10 msec), but will make it more difficult to follow rapid changes in the pitch of a periodic signal.

3.4 Examples of the Coincidence Function

At this point, it should be helpful to present several examples of the coincidence representation of different types of sounds. Figure 3.5 shows the coincidence function when the input is a synthetic periodic sound (all the harmonics of a 100 Hz fundamental). The first vertical stripe (away from the x origin) is the location of the pitch period. Since the output of each frequency region of the cochlear model will have the same pitch period, each frequency region in the coincidence function has a peak at the same location.

Figure 3.6 shows the coincidence function of another periodic sound. This sound is very similar to the sound in figure 3.5, except that the first seven harmonics are missing. The cochlear output of the low frequency channels is very weak

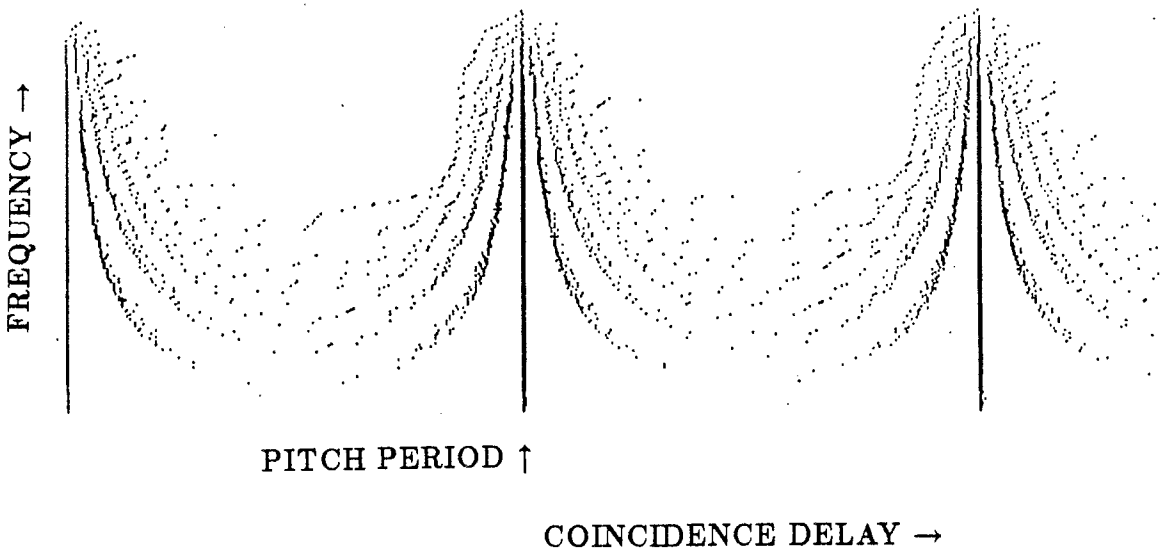

Figure 3.5: The coincidence function of a synthetic periodic sound. All the harmonics of 100 Hz fundamental are present with an amplitude that rolls off as 1/F.

and as a result, the coincidence picture in these frequency regions looks blank. All the frequency channels with center frequencies above the location of the eight harmonic have a peak at the pitch period. The pitch period can be easily determined from this representation as the location of the first vertical stripe (away from the x origin).

The coincidence function of white noise is shown in figure 3.7. Since the output of each frequency channel will be bandlimited noise, the coincidence function dies away as the coincidence delay increases (since the correlation between the channel's output will decrease as the time between the two points increases). Notice the lack of any structure, or vertical stripes (like the periodic sound sources).

The coincidence function of periodic speech is shown in figure 3.8. Each horizontal stripe in this picture represents a concentration of spectral energy. Some harmonics do not show up in the picture (such as the second harmonic) since the amplitude of this harmonic is low. Each frequency region with strong enough energy shows a peak in the coincidence delay at multiples of the pitch period.

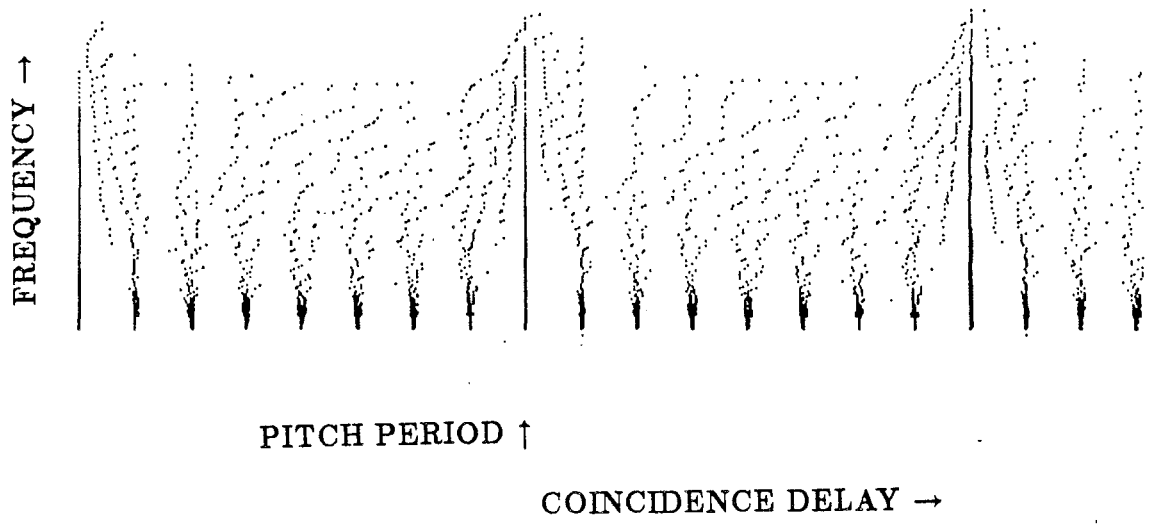

Figure 3.6: The coincidence function of harmonics 8 through 79 of the same periodic sound in figure 3.5

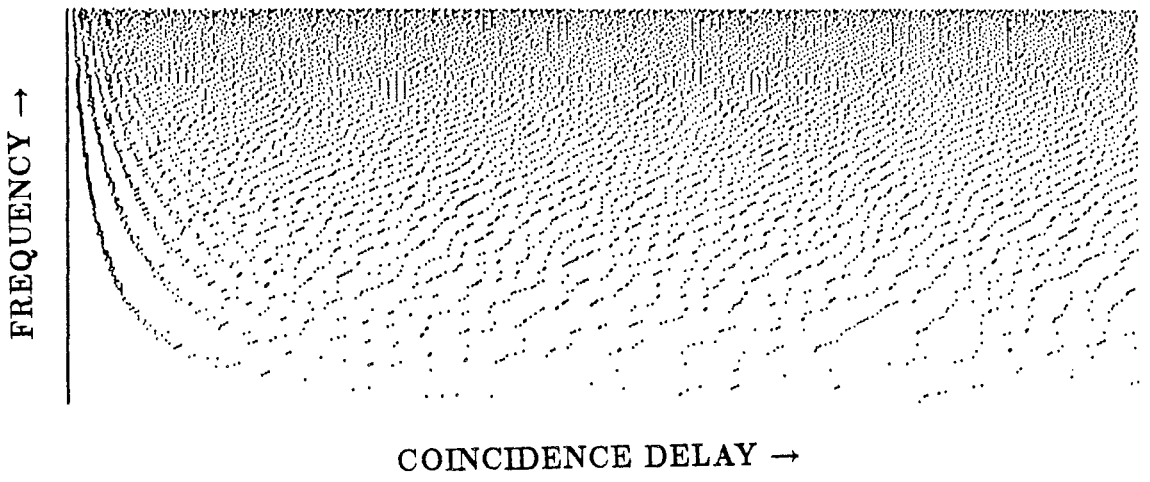

Figure 3.7: The coincidence function of white noise

The last example of the coincidence function is shown in figure 3.9. The input signal is the sum of two sine waves of 100 and 110 Hz. The 110 Hz sine

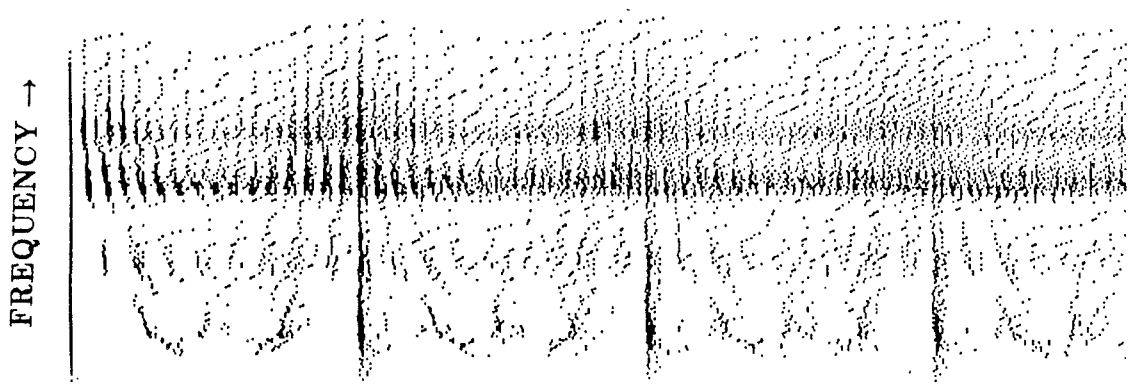

Figure 3.8: The coincidence function of the vowel /I/ in the digit /six/

wave has twice the amplitude of the 100 Hz sine wave. The waveform can be viewed as the sum of a 110 Hz sine wave with an amplitude modulated (10 Hz) sine wave of 105 Hz. When the 105 Hz amplitude modulated sine wave reaches it's amplitude envelope minimum, the 110 Hz sine wave can be clearly seen. The middle and bottom pictures are the coincidence function of a single frequency channel shown over time. As the incoming signal's frequency varies between 110 Hz and some intermediate value (between 105 and 110 Hz), the peak in the this frequency channel's coincidence function will also vary.

This example shows the reason for the development of the smoothing algorithms used (described in section 3.6.1 and 3.6.3). The peak in the coincidence function for two steady state periodic sounds will lie between the pitch periods of the two individual periodic sounds. Each row of the coincidence function is then convolved with a smoothing waveform. If the peak is exactly between the two individual pitch periods, then smoothing the coincidence function will result in an equal amplitude at each of the individual pitch periods. The incoming energy can then be split evenly between the two sound sources. If the coincidence peak is closer to one pitch period than to the other, then the value after the smoothing

operation can be used to assign more of the energy to this sound source than to the other sound source. The smoothing function used in the separation system (described in section 3.6.1, also see figure 3.12) varies with the frequency channel being smoothed; the total width of the smoothing function is rougly equal to the period of a sine wave with a frequency equal to the center frequency of this channel.

3.5 The First Separation System

The algorithms for separating sounds presented in this thesis have been modified many times. The previous section discussed several different versions of the coincidence function which were developed and modified for different reasons. The computer model for separating sounds has also undergone many different changes over time. Two different implementations of the theory of sound separation (presented in chapter two) have been developed. This section will briefly review the first computational model for separating sounds. This model is no longer in use and has been replaced by a newer version of sound separation algorithms. It is described briefly in this section to explain the difficulties that were encountered in implementing the auditory theory. After the limitations and problems with this initial system have been reviewed, the current computational model of auditory sound separation will be presented.

The goal of the first model was to separate two periodic sounds in a way similar to the human auditory system. The acoustic cues used to separate the two periodic sounds are pitch, pitch dynamics, and onset information. These information cues were computed from the average coincidence representation frames (computed every 10 msec).

Since the computation of the average coincidence function for periodic information is local in both time and frequency, two possible outcomes arise: (1) if at some time sound number one was much stronger than sound number two in one frequency region, then the coincidence function would reflect the properties of the stronger periodic sound; (2) if sound two was much stronger than sound one in a different frequency region, then that frequency region would reflect the features

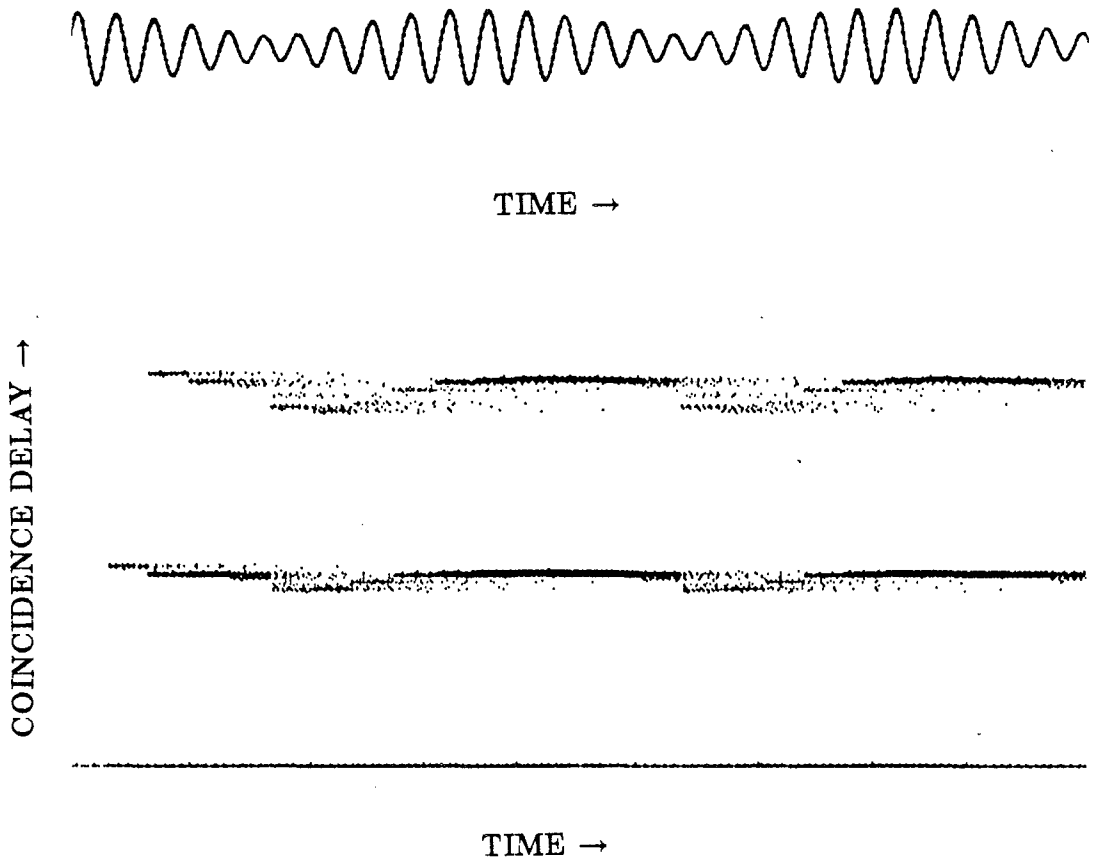

Figure 3.9: Top: Sum of two sine waves (F=100, 110 Hz) of unequal amplitude (the 110 Hz sine wave is twice as large as the 100 Hz sine wave).
Bottom: The coincidence function of a single frequency channel over time. Each column of this picture is the coincidence function of that frequency channel at a different time. This channel's coincidence function is time aligned with the original waveform (above).

of the second sound source. Even though two sounds may have the same overall energy, the two sounds can have the energy concentrated in different frequency locations. The coincidence function in each frequency channel would respond to the sum of the two signals, but if one sound was much stronger than the other sound in that frequency region, the properties of the coincidence representation would reflect only the information from the stronger sound source.

A key approximation in this first system was that 'each channel of the coincidence function will reflect the information of either one sound source or the other'. Since the computations are local in both frequency and time, two sounds that had the same energy could be easily separated if they had different distributions in the frequency-time plane, such that locally one sound was stronger than the other. Using this approximation, the system then focused on determining which sound source the periodic information in a given frequency channel represented. When it determined that there were two periodic sounds present, it decided whether the periodic information present in this frequency channel was consistent with the first pitch period or the second pitch period.

The neural events in each frequency channel at each point in time were assigned to the different group objects present, based on the consistency of the frequency channel's features (pitch, pitch dynamics and onset) and those of the group object. There were two types of group objects in this first system. There were *periodic* group objects and *burst* group objects. The periodic group objects represented steady state periodic sound segments and the burst group objects represented the events at the onset of a periodic segment. The burst groups were necessary because the events at the onset of a periodic sound have nothing in the past to autocorrelate with (and have a very low periodicity value). There were plans to add a *nonperiodic* group object to the system (to represent the nonperiodic speech segments), but this was never implemented.

The system first computed the coincidence representation of the incoming sound. Next, it computed the different information cues (local to each frequency-time region) so that the events could be assigned to the different group objects. The average coincidence function of all recent events in each frequency channel were computed every 10 msec. Since the value of the average coincidence function at the pitch period was sometimes split between two neighboring bins, the amplitude of the coincidence function did not accurately reflect the true degree of periodicity at the pitch location. Therefore, before computing the different local features for sound separation, the coincidence function in each frequency channel was smoothed by convolving it with a gaussian envelope. A gaussian curve for a

smoothing function is just one of the many possible shapes; the standard deviation was equal to 2 samples at a sampling rate of 16 khz. This smoothing operation was added to remove the effect of bin splitting, and to have the amplitude of the coincidence function at the pitch period reflect the periodicity of the signal at that location. The resulting representation was called the *smoothed coincidence function.* The height of the resulting smoothed coincidence function was then used to locate the pitch period.

For each frequency channel, independent features were computed from this smoothed coincidence function (such as the pitch period of the sound in that frequency region). Since it was not possible to determine the pitch period in each frequency channel with 100% accuracy, a list of the possible pitch periods in each frequency channel was computed. A pitch strength was computed for each possible pitch period in each frequency channel. Possible pitch periods were chosen using the following algorithm: all local peaks in the smoothed coincidence function, having an amplitude greater than .7 (an arbitrary number that worked well) multiplied by the maximum value of the coincidence function in the allowable pitch interval, were chosen as possible pitch periods. Using this formula, it was rare for the actual pitch period not to be included among the possible pitch periods, although a channel sometimes contained many false possible pitch periods.

In addition to computing the pitch period of each frequency channel every 10 msec, the pitch dynamics and the percent amplitude change were also computed. The percent amplitude change was computed to determine if an onset of a sound had occurred; the pitch dynamics were computed to aid in the separation of the two periodic sounds.

The next stage in the separation processing was the assignment of events to different group objects. A summary of how the events were assigned to the different group objects is listed below:

1. For each of the group objects that already existed, the system determined how well this group object accounted for the information present in the coincidence representation. When different features of a frequency channel were close to the features of the group object (i.e., when the weighted distance

between the feature vectors was less than a threshold), the group object was said to be compatible with the information present in that frequency channel. A group object could be compatible with some frequency channels by random chance, even though the sound that the group object represented might have already stopped. Therefore, a minimum number of frequency channels had to be explained by any pre-existing group object before any of the channels could be assigned to that group object. The determination whether there were enough frequency channels in the current time frame which were consistent with each group object was made by checking that the number of frequency channels (which were compatible with a group object) was greater than a set threshold. If the system determined that there were enough frequency channels, the individual frequency channel objects (all the events in the past 10 msec window in that frequency channel) were assigned to the group object with a 'link' of a certain strength (based on how close the features of the frequency channel and the group object were).

2. After the events in some frequency channels were linked to the existing group objects, the system determined whether there were any frequency channels that were 'unexplained' by any of the current group objects (frequency channels with no links to any group object or very weak links). If there were a large number of frequency channels that remained unexplained, the system tried to create a new group object that explained the remaining frequency channels. All frequency channels not well explained by any existing group object (links to current group objects less than some threshold) were collected. If there were enough frequency channels that were unexplained by any existing group object, and if there is a new group object which can explain this data to a certain level (explain more than 70% of the remaining frequency channels), then a new group object was created and the appropriate frequency channels were linked to this new group object.

There were several problems with this first separation system. The main difficulties that were encountered with this model are summarized below:

1. The system required that the information features of the group object at the previous time frame (pitch period, pitch dynamics, increase in amplitude) be close to the features of the different frequency channels in the current time frame (10 msec later) for the frequency channel to be assigned to the group object. Sometimes, the pitch of a talker would change from one time frame to the next by an amount greater than the pitch continuity threshold used. This would result in none of the frequency channels being assigned to the appropriate group object, and a new group object would then be created. These two group objects would have similar pitch period features and would then compete with each other to explain the different frequency channels even though both group objects belonged to the same sound source.

2. Sometimes there were channels not explained by any group object, but the channels were not numerous enough to justify creating a new group object. Therefore weak sounds sometimes went undetected because of the lack of sufficient information. By lowering the threshold used to create a new group object, one could decrease the number of times that frequency channels went unexplained. This lowered threshold would result in the creation of false group objects, which did not correspond to any sound source present. The threshold in this system was set at a high level so that false group objects would never be created. This disadvantage was somewhat offset by the backtracking algorithm used. That is, if a new group object was created at some time, a search was undertaken for frequency channels in previous time frames that could be assigned to this new group object.

An example of the separated output from this system is shown in figure 3.10 and 3.11. There are two synthetic vowels present with different pitch periods, different onset and offset times, and formants in different frequency regions. Shortly after one vowel has begun, a second synthetic vowel is started. The system determines both how many sounds are present and which frequency channels were created by which sound sources.

By looking at the cochlear model's output of the higher frequency channels,

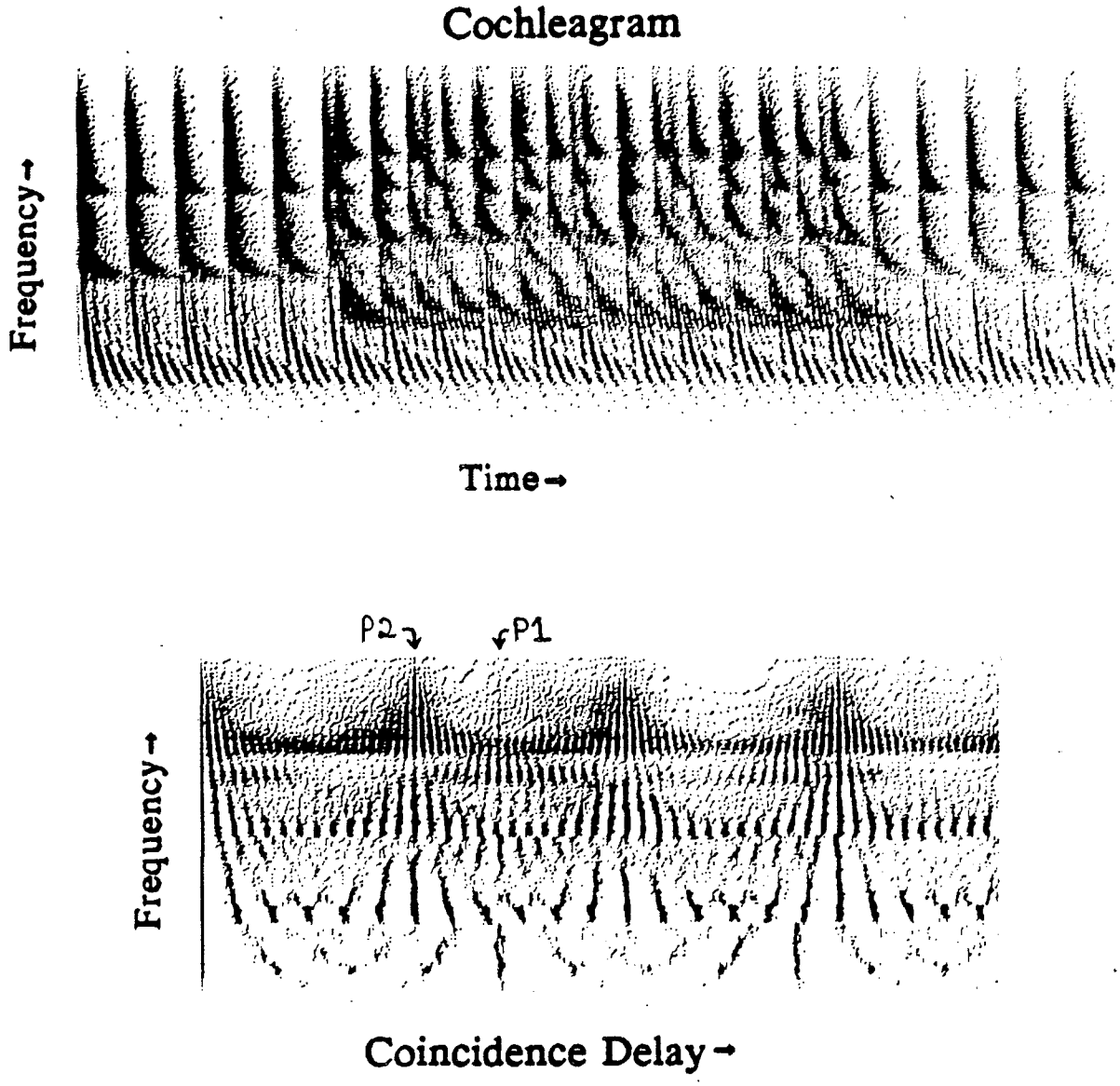

Figure 3.10: Top: The original cochlear representation of the two synthetic vowels. Each vowel has its three formants in different frequency regions. Bottom: The coincidence representation when both synthetic vowels are simultaneously present.

one can see that each pitch pulse from the different sound sources was separately resolved in time. The output of the separation program assigned all the events in each frequency channel in a 10 msec interval either to one group object or to

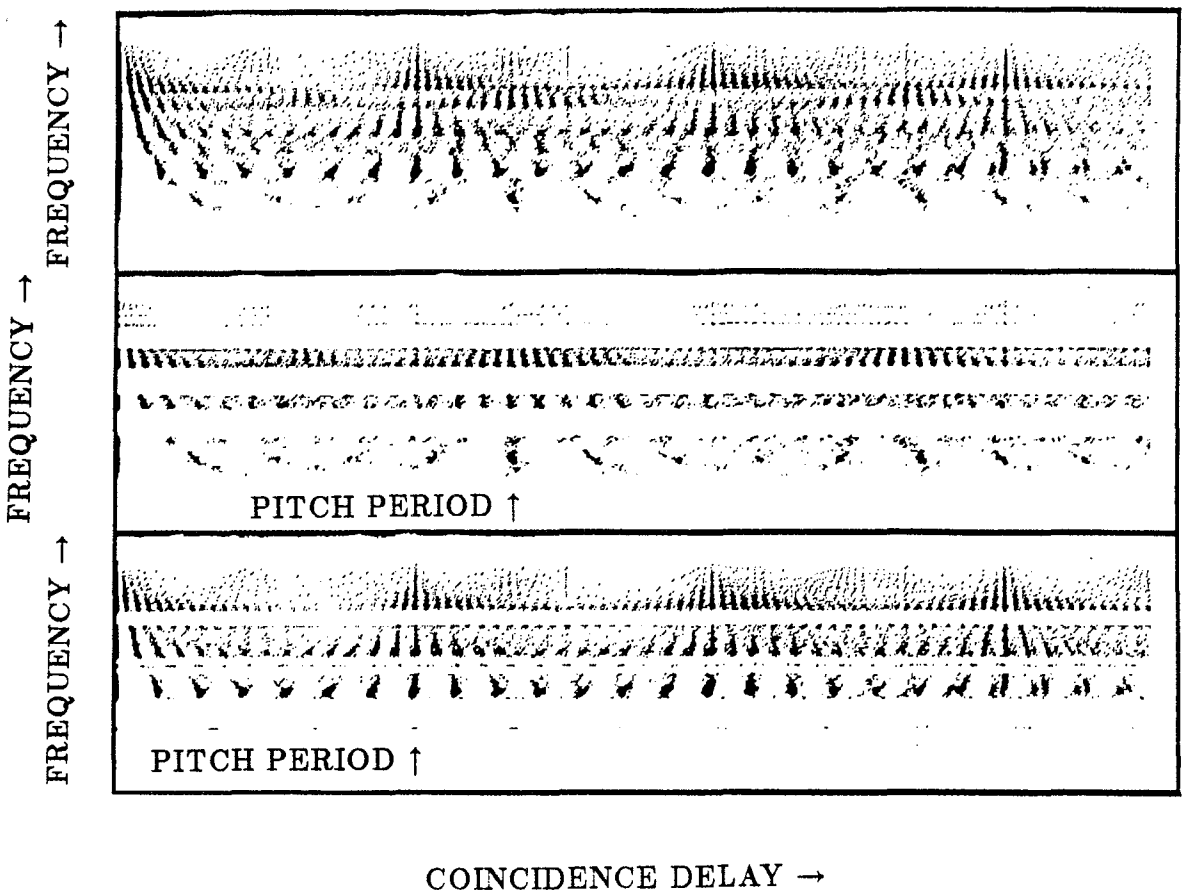

Figure 3.11: Top: The same coincidence function shown in figure 3.10. Middle: Those frequency channels with one pitch period were assigned to the first group object. Bottom: Those frequency channels with the second pitch period were assigned to the second group object.

the other group object. Based on these results, the method used to assign events to group objects in the first system was then modified. Instead of assigning all the events in a frequency channel in a 10 msec time interval to one group object or another, independent link strengths between each event (in each frequency channel) and the different group objects were computed.

There are several difficulties with the algorithms used by this first computational model of sound separation. The separation program contained many different thresholds for making the decisions that were necessary to separate sounds.

All decisions made by the system used local information, and independent decisions were made every 10 msec. The system assumed that the features computed in each frequency channel (pitch, pitch dynamics, increase in amplitude) reflected the features of either one sound source or the other sound source (they were not some intermediate feature vector). The decision when to start and stop group objects was based on satisfying arbitrary criteria, as was necessarily the case with many of the decisions made by this first computational model.

In an attempt to solve these problems, a second computer model of the auditory sound separation process was developed. The next section describes this model.

3.6 Current System Overview

Since many decisions must be made by a separation system, a decision framework for separating sounds was needed. The second, and current computer model attempts to solve the difficulties encountered with the first separation system. The difficulties with this second computational model of sound separation are described in chapter four where the performance of this system in separating two simultaneous talkers is reviewed.

The goal of the current computer model is to separate the simultaneous speech of a male and female talker. The author selected two male and two female talkers from a database of speakers used for a speaker-independent continuous-digit recognition system being developed in the same laboratory [Kopec and Bush 1985]. For these four talkers, the author constructed a database of handmarked speech digit strings. This database was used to train and test the separation performance of the computer model. The output of this separation system was then used as input to the Kopec-Bush speech recognition system.

The database consists of 39 single-speaker digit strings (of seven continuous digits) spoken by two males and two females. It also consists of 38 examples of dual-speaker digit strings (obtained by adding the single-speaker waveforms of a male and a female speaking different digit strings). The database was limited to this size for computational reasons (eg, limits on disk storage, computer time

necessary to process the data).

The current separation system, then, consists of the following four stages:

1. An iterative dynamic programming pitch tracking algorithm determines the pitch period for each of two sound sources. If one or both sounds are not periodic, the corresponding pitch period has no meaning.

2. A Markov model represents the number of sounds present (i.e., one or two) and the type of each sound source (i.e., periodic or nonperiodic). By finding the probabilities of the states of the Markov model given the input, the Viterbi algorithm is used to determine when group objects start and stop, how many group objects are present, and which group objects belong to which sound streams.

3. An algorithm estimates the amplitude (in each frequency channel) of each sound source present, given information on the number of sounds and the type of each sound source. Both periodic information and spectral continuity constraints are used in an iterative algorithm to compute an estimate of each sound source. One constraint, however has been dropped: the approximation that the features in a frequency channel reflect one sound source or the other.

4. An algorithm resynthesizes a waveform from the separated output. In order to interface with the recognition system (at the current time), a waveform of the separated output is computed. This waveform also allows people to listen to the separated results.

The next sections will describe in detail how each of the different algorithms works.

3.6.1 Fundamental Frequency Computation for Two Speakers

When a person is speaking, some of the speech segments can be classified as being 'periodic'; others segments can be classified as 'nonperiodic' (although

from both sounds would be multiples of the resulting fundamental frequency.

To compensate for these false peaks in the average coincidence function (at a multiple of both pitch periods), a scaling factor is applied to emphasize peaks in this representation with a smaller value of the pitch period. This scaling factor is a linear weighting which varies from a value of 1.0 at zero delay to a value of .1 at the end of the coincidence delay line. The peaks at the actual pitch periods are emphasized over peaks at multiples of the pitch periods.

A dynamic programming algorithm computes a pitch track instead of making independent decisions for the pitch values every 10 msec. Two options exist for computing a dynamic programming pitch track for two periodic sounds. These two options are compared below.

- A dynamic programming algorithm traces the pair of pitch periods (P1,P2) through time. Since each pitch period could vary over a range of 384 samples (the length of the delay line), there are 147456 different values for the two pitch periods (P1,P2). A dynamic programming score for each of the 147456 pitch period pairs would have to be computed every 10 msec. Since such extensive computation is not feasible, a beam search would be necessary to reduce the computation load, but such a search might result in the deletion of the actual pitch period pair.

- An iterative dynamic programming algorithm first traces the **dominant** pitch period (the pitch period with a stronger peak in the average coincidence function) through time, and then traces the **weaker** pitch period through time. This algorithm is computationally more feasible, since at each of the two iterations, only 384 possible pitch values need to be considered. The disadvantage of this approach is that if it makes an error while computing the first pitch period, this error might also cause an error in the location of the second pitch period.

Due to the computational considerations mentioned, the iterative dynamic programming algorithm was chosen as the algorithm for computing the pitch pe-

riod. Different dynamic programming algorithms are used to compute the dominant pitch track and the weaker pitch track.

A dynamic programming algorithm computes a pitch score for every possible dominant pitch period (p = 0 to 383 samples) every 10 msec. It is desired that the location of the maximum in the dominant pitch score is equal to the pitch period of one of the two sound sources. The dominant pitch score

$$DynProgScore1(P,T) \stackrel{\text{def}}{=} \begin{array}{l} PitchScore1(P,T)+ \\ \max_{p=0}^{383} \left\{ \begin{array}{l} DynProgScore1(p,T-1)+ \\ TransScore1(P-p,T) \end{array} \right\} \end{array} \quad (3.1)$$

is equal to the sum of the score from the periodicity information at the current time frame, and the score from transitional information (which is based on the dominant pitch score at the previous time). The pitch score is computed as follows:

$$PitchScore1(P,T) \stackrel{\text{def}}{=} ModAvgCoin(P,T) \quad (3.2)$$

$$ModAvgCoin(P,T) \stackrel{\text{def}}{=} \begin{cases} AC(P,T)* LinWt(P) & \text{if } AC(P,T) > AC(P-1,T) \text{ and } AC(P,T) > AC(P+1,T) \\ 0 & \text{otherwise} \end{cases} \quad (3.3)$$

The average coincidence function [AC(P,T) is shown in the bottom picture of figure 3.14] is equal to the average of all the rows of the smoothed, normalized, minus-random coincidence representation in figure 3.13 (bottom right picture). The modified average coincidence function [ModAvgCoin(P,T)] is zero everywhere except at the local maximums of the average coincidence function. This restriction forces the dominant pitch track to pass through a local maximum in the average coincidence representation. The average coincidence function is scaled by the linear weighting function previously described (varies from 1.0 at P=0 to .1 at P=383), to prevent multiples of the pitch period from being chosen as the actual pitch period.

The transition score is a function of the pitch change (from the previous time frame to the current time frame), and is largest when the pitch transition is smallest.

$$TransScore1(deltaP, T) \stackrel{\text{def}}{=} \exp^{(-.1*abs(deltaP))} * \max_{p=0}^{383} ModAvgCoin(p, T) \quad (3.4)$$

The first term in equation 3.4 varies from 1.0 (when the change in the pitch period is equal to 0) to 0.0 (when the pitch period change is very large). The second term scales the transition score by the maximum of the modified average coincidence function (equation 3.3). Since the pitch score varies depending on the height of the average coincidence function, the transition score is scaled to balance the relative importance of the pitch score and the transition score. This transition score favors the pitch period over multiples of the pitch period. When the pitch of a periodic sound changes by n samples, twice the pitch period will change by 2n samples, and will therefore have a lower transition score than the actual pitch period.

The value of the pitch period P which maximizes the dynamic programming score at a time T is computed. The algorithm then backtracks 25 frames (250 msec), tracing the previous pitch period p which gave rise to this maximum in the dynamic programming back in time. The resulting pitch period (250 msec before the current time) is the dominant pitch period at that previous time frame.

Once a decision has been made about the value of the dominant pitch period at every point in time, these pitch periods are then assigned to the sound stream which is believed to have generated them. Each of two sound streams contain the average pitch value of the speakers that are being separated. The dominant pitch value is assigned to the sound stream with the closer average pitch period (provided that the dominant pitch value is within 60% of the average pitch value of that sound stream). The dominant pitch period will oscillate between the two sound streams, depending on which periodic segment is louder at the time.

Since the system attempts to separate a male and a female speaker (whose average pitch values are different), the system can assign the pitch to the sound stream based upon which speaker has the closest average pitch period. This algo-

rithm would not be feasible when attempting to separate two speakers with the same average pitch.

A second dynamic programming algorithm (equation 3.5) is then used to fill in the missing pitch periods in each sound stream. The best pitch track between the known pitch endpoints (when the dominant pitch period switches from this sound stream to the other sound stream and then back again) is then computed. The dynamic programming score for the weaker pitch period

$$DynProgScore2(P,T) \stackrel{\text{def}}{=} \begin{array}{l} PitchScore2(P,T)+ \\ \max_{p=0}^{383} \left\{ \begin{array}{l} DynProgScore2(p,T-1)+ \\ TransScore2(P-p,T) \end{array} \right\} \end{array} \quad (3.5)$$

is also equal to the sum of a pitch score and a transition score. The pitch score for the weaker pitch period is computed as follows:

$$PitchScore2(P,T) \stackrel{\text{def}}{=} \frac{\sum_{frequency} \max\{0, MCoin(f,P,T) - MCoin(f,DomPit1(T),T)\}}{NumFreqChannels} \quad (3.6)$$

MCoin is the coincidence representation after it has been smoothed, normalized, and had the coincidence function of white noise subtracted from it (bottom right picture in figure 3.13). If the average spectrum of the weaker periodic sound is larger than the average spectrum of the dominant periodic sound in some frequency regions, one would expect the value of the modified coincidence representation (in these frequency regions) at the location of the weaker pitch period to be larger than the value at the dominant pitch period. In each frequency channel, the difference in the value of the modified coincidence function at possible pitch period P with the value of the modified coincidence function at the dominant pitch period is used to indicate that P is a possible location of the weaker pitch period.

The transition score for the weaker pitch period

$$TransScore2(deltaP,T) \stackrel{\text{def}}{=} -.1 * (\frac{abs(deltaP)}{10})^2 \quad (3.7)$$

has a heavy penalty for large pitch variations. The transition score is not scaled in amplitude (like the transition score in equation 3.4), so that when the pitch score

is very small (due to the weaker periodic sound being masked), the transition score will increase in importance (to increase contextual information).

The weaker pitch track is constrained to start and stop at known pitch endpoints (obtained when the dominant pitch period is assigned to this sound stream). This second dynamic programming algorithm fills in the values of the pitch track of each sound source when the dominant pitch value was not assigned to this sound stream.

The result of this iterative dynamic programming algorithm is a pitch period value at every time frame (every 10 msec) for each of two sound streams. The dominant pitch period is computed first, and is then assigned to the sound stream with the closest average pitch period. A second dynamic programming algorithm is used to fill in the missing pitch values of each sound stream, (i.e., when the dominant pitch period is not assigned to this sound stream).

The value of a pitch period in each time frame does not mean that there are two steady-state periodic sounds present. These pitch values will be used in the next stage to determine whether there is one sound present or two sounds present. These values are also used to determine whether each sound source is periodic or nonperiodic.

3.6.2 Hypothesis Determination

The speech database of 40 single-speaker digit strings (of seven continuous digits) was handlabeled based on a finite state model. Every 10 msec, a label was assigned to that segment of the speech database. The finite state model currently consists of seven states: silence, periodic, nonperiodic, onset, offset, increasing-periodicity, decreasing-periodicity. Three of the states are 'steady states', while the other four states are 'transitional states'. A speech sound can remain in a steady-state sound from one time frame to the next, but it can only remain in a transitional-state for 10 msec, after which it must transition to one of the steady states. A state transition diagram along with the transitional probabilities between states (computed from the database) is shown in figure 3.15.

The model for two sound sources consists of a separated state transition

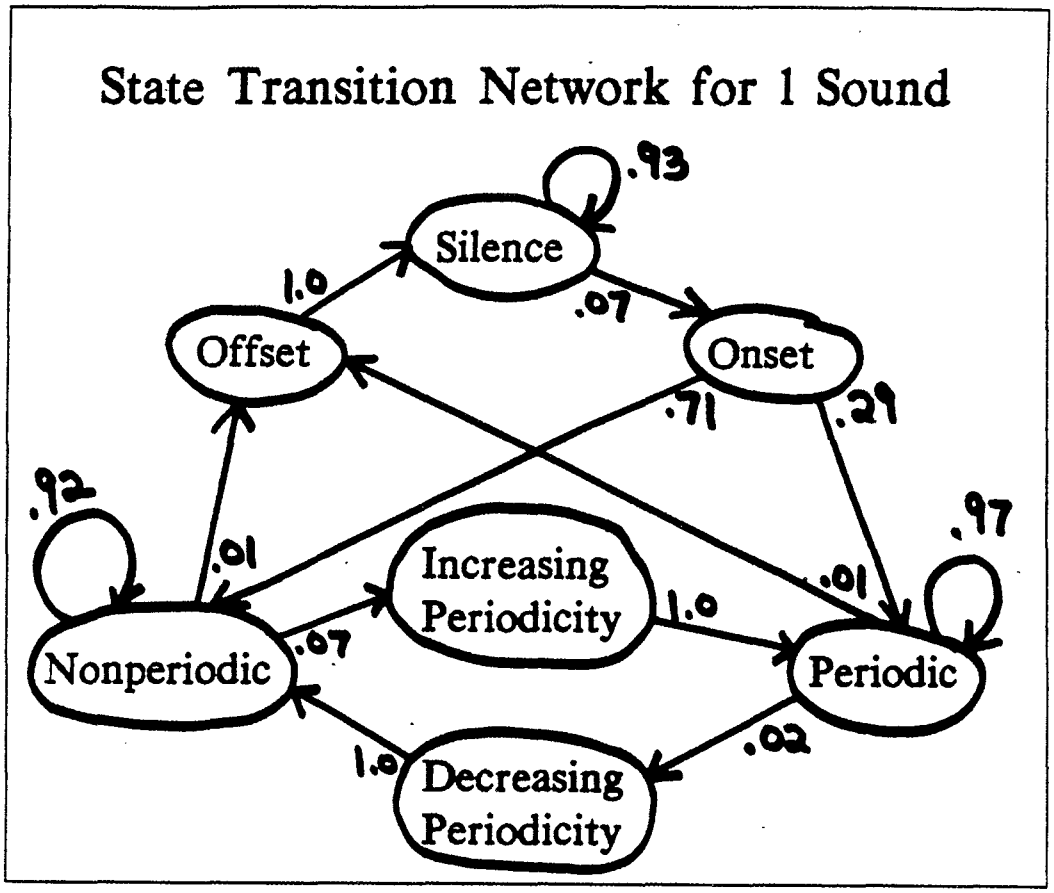

Figure 3.15: State transition network for a model of one sound source.

network for each sound stream. The only constraint imposed on the two-sound model is that only one sound can make a transition to or from a transitional state (onset, offset, increasing-periodicity, decreasing-periodicity) at a time. One sound source model must remain in the same state while the other sound source model transitions between the different steady states. The resulting model consists of 33 states and 69 state transitions.

Instead of making independent decisions at each point in time, the state transitions constraints of the Markov model are used to maintain decision continuity across time. The Viterbi algorithm is used to decide both how many sound sources are present, and what the characteristics of each sound source are (periodic, nonperiodic, ...). The system could decide that only one sound source was

present by labeling one of the sound streams as silence.

At every time frame, the hypothesis cost of every possible state

$$HypCost(State_i, T) \stackrel{\text{def}}{=} \min_{allowable\ transitions} \left\{ \begin{array}{l} HypCost(State_j, T-1) + \\ \quad TransCost(State_j \rightarrow State_i) \end{array} \right\} \tag{3.8}$$

is computed. The hypothesis cost of each of the 33 states is computed from the hypothesis cost of the previous state at the previous time and the cost of the transition (from the previous state to the current state). The transition cost

$$TransCost(State_j \rightarrow State_i) \stackrel{\text{def}}{=} \begin{array}{l} TransDataCost(State_j \rightarrow State_i) + \\ \quad TransAPrioriCost(State_j \rightarrow State_i) \end{array} \tag{3.9}$$

consists of two components: a cost which depends on the data and a cost which depends on the *a priori* state transition probabilities. The *a priori* state transition cost

$$\begin{array}{l} TransAPrioriCost(State_j \rightarrow State_i) \stackrel{\text{def}}{=} \\ \qquad -\log Probability(State_j \rightarrow State_i \mid State_j) \end{array} \tag{3.10}$$

is computed from the state transition probabilities computed from the handmarked database. The data transition cost

$$\begin{array}{l} TransDataCost(State_j \rightarrow State_i) \stackrel{\text{def}}{=} \\ \qquad Avg_{freq} \left\{ \begin{array}{l} PeriodicityCost(f, State_j \rightarrow State_i) + \\ \quad AmplitudeCost(f, State_j \rightarrow State_i) + \\ \quad\quad AmpChngCost(f, State_j \rightarrow State_i) \end{array} \right\} \end{array} \tag{3.11}$$

is the average data cost across all the different frequency channels. The data cost in each frequency channel for a state transition consists of three components: a periodicity cost, an amplitude cost, and an amplitude change cost. The local periodicity cost

$$PeriodicityCost(f, State_j \rightarrow State_i) \stackrel{\text{def}}{=}$$
$$-\log Probability(f, Periodicity\ Data \mid State_j \rightarrow State_i) \tag{3.12}$$

is computed from the probability of observing the periodicity information for each of the state transitions. Different periodicity information is used depending on the state transition we are considering.

$$Probability(f, Periodicity\ Data \mid State_j \rightarrow State_i) \stackrel{\text{def}}{=}$$
$$\begin{cases} Prob(MCoin(f, \phi = Pitch_1)) & \text{if one sound present} \\ Prob(MCoin(f, \phi = Pitch_1)) & \text{if sound one is periodic, second sound is nonperiodic} \\ Prob(\max(MCoin(f, P_1), MCoin(f, P_2))) & \text{if both sounds are periodic} \end{cases} \tag{3.13}$$

Each of the probabilities in equation 3.13 is obtained by histogramming the periodicity information for each state transition. When the state being considered is only one person speaking (the other speaker's model remains in the silent state), the probability of observing this value of the modified coincidence function at that person's pitch period is computed. When the state being considered is two simultaneous periodic sounds, the probability of observing the maximum of the modified coincidence function is computed. The probability histograms of these values were computed for each state transition. The pitch periods used in computing these probability histograms were the pitch tracks that were computed on each of the isolated sounds. The probability of the amplitude and amplitude change data

$$Probability(f, Amplitude\ Data \mid State_j \rightarrow State_i) \stackrel{\text{def}}{=}$$
$$Probability(Coin(f, \phi = 0) \mid State_j \rightarrow State_i) \tag{3.14}$$

$$Probability(f, Amplitude\ Change\ Data \mid State_j \rightarrow State_i) \stackrel{\text{def}}{=} Probability(\frac{Coin(f,\phi=0,T)-Coin(f,\phi=0,T-1)}{Coin(f,\phi=0,T)+Coin(f,\phi=0,T-1)} \mid State_j \rightarrow State_i) \quad (3.15)$$

are computed from probability histograms for each of the different possible state transitions.

At the end of the incoming sound signal, the best path is determined by backtracking from the state with the minimum cost at the last time frame. The resulting state path determines when periodic and nonperiodic group objects start and stop, and which sound stream they belong to.

If two simultaneous sounds are present at one time, the Markov model implicitly assigns each of the different group objects to different sound streams. The only way that a nonperiodic segment could always be correctly assigned to a sound stream is if this segment overlapped in time with a periodic segment from the other speaker.

3.6.3 Spectral Amplitude Estimation

The previous two sections have described the way the system determines both the number of sounds present and the characteristics of each sound source. The next step is the estimation of each sound source's spectral amplitude, given that we know the types of sounds present.

When the system has determined that there are no sounds present (both speakers are silent), the spectral amplitude estimate of each sound source is 0. When it has determined that there is only one sound present (one sound source is silent, while the other sound is in one of the other six possible states), then the spectral amplitude estimate for this sound source is equal to the observed spectrum (the other spectral amplitude is set to 0). When there are two sounds present, the system uses the algorithms described later in this section to compute an estimate of the spectral amplitude of each of the two sounds present.

The difficulties in estimating the spectrum of each sound source (when there are two sound sources present) are listed below:

- When there are two periodic sounds and some harmonic components are approximately integer multiples of both fundamental frequencies, it is impossible to determine an amplitude estimate at these frequencies for each sound source only using information at the current time. When the pitch periods vary over time, estimates at neighboring time slices can help to resolve the uncertainty in the estimates of these harmonic components.

- Since the periodic signal typically undergoes changes from one pitch period to the next pitch period, it can be viewed as consisting of two components: a part that repeats exactly from one pitch period to the next, and a part that has changed from the previous pitch period. When one sound is periodic and the other sound is nonperiodic, there is a difficulty in estimating the spectral amplitude, since it is impossible to determine what part of the sum signal's observed nonperioidicty is due to the nonperiodic sound source, and what is due to the periodic sound source.

- Since the frequency responses of neighboring frequency channels overlap, independent estimates of the spectral amplitude in each frequency channel may produce a spectral estimate which is not physically realizable. There are constraints on both what spectral amplitudes are possible to synthesize and what spectral amplitudes are likely to be produced by the human voice.

One would ideally like to generate the maximum likelihood spectral estimate for each sound source given the periodicity information present and the spectral continuity requirements of speech sounds. However, a joint estimation of the spectral amplitude of each sound source over a frequency time region is not feasible, since it would require the joint estimation of thousands of variables. The spectral estimation algorithms that are described in this section use a two step approach listed below.

1. Compute an initial spectral estimate of each sound source using only local periodicity information.

2. Iteratively compute a spectral estimate for each sound source which locally minimizes a cost function (maximizing the probability of each local spectral estimate). At each iteration, the cost of a local spectral estimate is computed based on both the observed periodicity information and the current spectral estimates of the neighboring frequency-time regions. The spectral estimate with the lowest cost is used as the current spectral estimate for this frequency-time region at this iteration level.

The spectral amplitude estimation algorithm first computes the ratio of the spectral amplitude of the two sound sources, and then the estimate of each sound source is computed from this estimate of the spectral amplitude ratio. Since the initial stage of the cochlear filterbank is a linear process (before compression and half wave rectification), the filterbank output of the sum of the two sounds is equivalent to the sum of the filterbank outputs of the isolated sounds. The sum of the two filterbank outputs might produce an output which is smaller than the original filterbank output depending on the phase relations between the filterbank output of the two sound sources. The expected value of the average filterbank output

$$\begin{aligned}\overline{AS} &\approx \tfrac{1}{2\pi}\int_{\theta=0}^{\theta=2\pi}\sqrt{A1^2+A2^2-2*A1*A2*\cos(\theta)}\,d\theta \\ &= \sqrt{A1^2+A2^2}\left[1-\tfrac{1}{4}\left(\tfrac{A1*A2}{A1^2+A2^2}\right)^2-\tfrac{15}{64}\left(\tfrac{A1*A2}{A1^2+A2^2}\right)^4-\ldots\right]\end{aligned} \tag{3.16}$$

can be approximated by the expected value of the sum of two narrowband filter filters with a random phase between them. This formula is then used to estimate the actual spectral amplitudes of each sound source, given: 1. an estimate of the ratio of spectral amplitudes between the two sound sources, and 2. the observed spectral amplitude of the sum of the two sounds. These equations are listed below.

$$\begin{aligned}\widehat{A2} &= \frac{AS}{\sqrt{1+\hat{R}^2}}\left[1-\tfrac{1}{4}\left(\tfrac{\hat{R}}{1+\hat{R}^2}\right)^2-\tfrac{15}{64}\left(\tfrac{\hat{R}}{1+\hat{R}^2}\right)^4-\ldots\right] \\ \widehat{A1} &= \widehat{A2}*\hat{R}\end{aligned} \tag{3.17}$$

The initial computation of the amplitude ratio $\hat{R}$ is obtained using the periodicity information in the current frequency-time region. The way that this estimate of the amplitude ratio is computed for two periodic sounds is described below.

1. Those time frames in the database of 38 dual-speaker digit strings when both speakers voices have been labeled as periodic are collected into a new database consisting of two simultaneous periodic sounds. Pitch tracks for each of the two speakers were computed on the single-speaker digit strings.

2. In each frequency channel, a histogram was computed of the values of the AS smoothed and normalized coincidence (ASNC) function (see the top right picture in figure 3.13) at the pitch periods. Based on this histogram, the values of the ASNC at the pitch period was divided into five equally probable regions. The value of the ASNC at the pitch period was then transformed into a bin value between 1 and 5.

3. In each frequency channel, the amplitude ratio of the two periodic sound sources was computed. The actual amplitude ratio was obtained by computing the ratio of the amplitude values of the isolated periodic sounds.

4. A histogram of the amplitude ratio of the two periodic sounds was then computed. There were 125 $(= 5 * 5 * 5)$ different histograms; a different histogram was computed for each possible bin combination (BP1,BP2,BPD). BP1 is the bin value of the ASNC at the first pitch period (see step 2), BP2 is the bin value of the ASNC at the second pitch period, BPD is the bin value of the ASNC at the difference in pitch periods (abs(P1-P2)). These 125 histograms contained the distribution of the amplitude ratio based on the coincidence information at these three locations. These histograms were computed with the data from all the different frequency channels.

5. The width of the amplitude ratio histograms reflects the uncertainty in the amplitude estimate based on the coincidence representation at these three

points. For each amplitude ratio histogram, an initial estimate of the amplitude ratio of the two sounds is computed.

6. The initial amplitude ratio estimate for separating two periodic sounds is computed by determining into which of the 125 bins the current frequency-time region's periodicity information falls.

A picture of several of the amplitude ratio histograms is shown in figure 3.16.

The way that the amplitude ratio estimate is computed for a periodic and a nonperiodic sound is very similar to the above procedure. Instead of having 125 different bins which averaged the information across all the different frequency regions, 5 histograms were computed in each of the 85 different frequency channels. The value of the ASNC at the pitch period of the periodic sound was used to divide the amplitude ratio information into 5 different categories in each frequency channel.

Once the initial amplitude ratio estimates have been computed, an initial spectral estimate for each sound source is computed (see equation 3.17). The initial spectral estimate of each sound source can fluctuate greatly across both frequency and time. Each frequency channel's amplitude estimate is based only on the local periodicity information in that frequency-time region.

After the initial spectral estimate is computed, an iterative algorithm for using local spectral continuity constraints is used. At each iteration, in each frequency channel, at every time frame, the amplitude ratio is varied in 100 steps between 0 and ∞. The new estimate of the amplitude ratio for this frequency-time region is then chosen as the amplitude ratio which is the best match between: (1) maintaining spectral continuity with neighboring frequency-time regions and (2) a highly probable amplitude ratio based on the periodicity information (computed from the amplitude ratio histograms).

The local spectral continuity information is obtained through probability distributions of amplitude change for a single sound source. For both periodic and nonperiodic segments of speech, histograms were computed for the following two

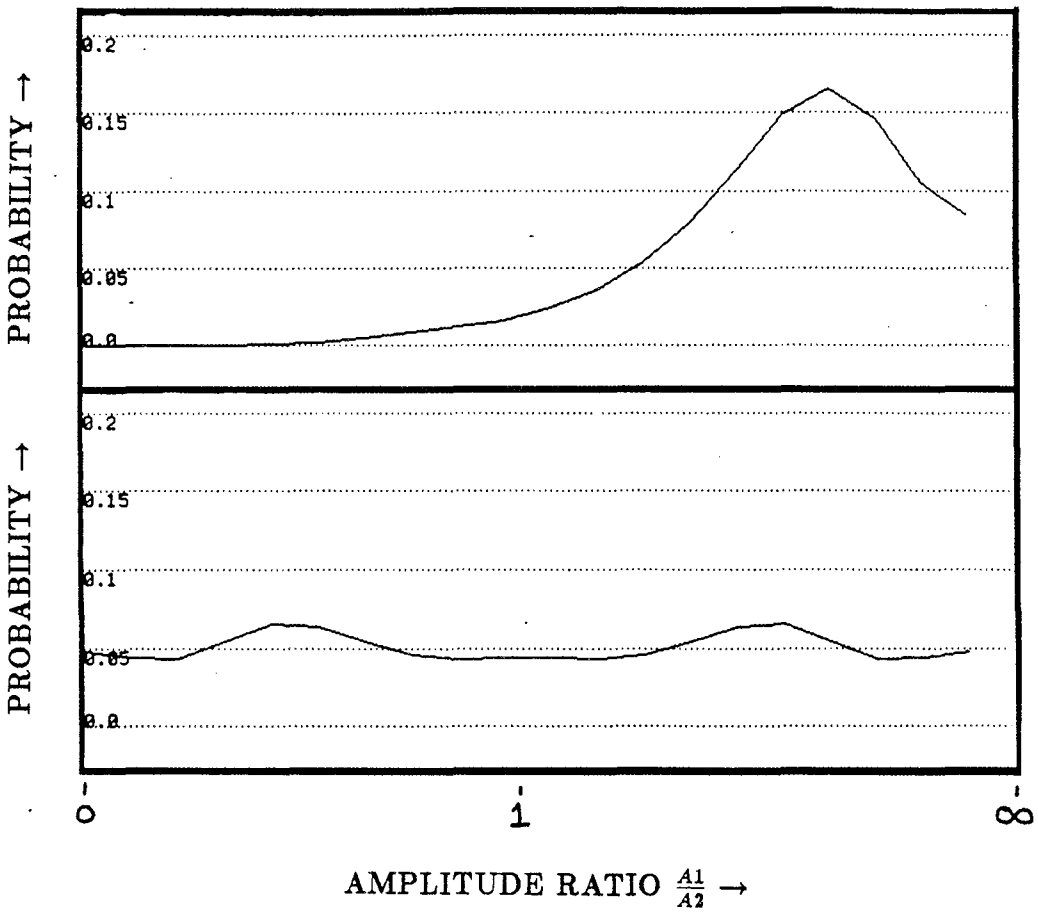

Figure 3.16: Top: Amplitude ratio histogram when the coincidence function is large at the first pitch period, small at the second pitch period and small at the pitch difference (P1-P2). Most of the time, sound one is larger in amplitude than sound two.
Bottom: Amplitude ratio histogram when the coincidence function is large at the the pitch difference (P1-P2), and small at the first and second pitch period. It is likely that the events came from sound one or sound two, but it is not clear which sound is larger.

amplitude change quantities (equations 3.18 and 3.19) in each frequency channel. The distributions of the change in amplitude obtained for periodic speech segments is shown in figure 3.17.

$$\frac{Coin(freq, \phi = 0, T) - Coin(freq, \phi = 0, T-1)}{Coin(freq, \phi = 0, T) + Coin(freq, \phi = 0, T-1)} \tag{3.18}$$

$$\frac{Coin(freq, \phi = 0, T) - Coin(freq - 1, \phi = 0, T)}{Coin(freq, \phi = 0, T) + Coin(freq - 1, \phi = 0, T)} \tag{3.19}$$

These histograms (figures 3.16 and 3.17) contain the important information which is then used to compute an iterative estimate of each sound source. The details of the spectral estimation procedure are now described.

If the Markov model has determined that there is only one sound present, the spectral estimate of that sound source is equal to the observed spectrum. This spectral estimate remains fixed and does not change over the iterative spectral estimation procedure. If the Markov model has determined that there are two sounds present, the spectral estimate of each sound source is computed using a different estimation procedure depending on the types of sounds present. An initial estimate of the amplitude ratio

$$R(F,T) \stackrel{\text{def}}{=} \frac{A1(F,T)}{A2(F,T)} \tag{3.20}$$

is estimated using only local information. The spectral amplitude estimates are obtained from the amplitude ratio using equation 3.17. When there are two periodic sounds present, the initial estimate of the amplitude ratio is determined from the amplitude ratio histograms (which histogram is used will depend on the value of the coincidence representation at P1, P2, and Pd). When there is a periodic and a nonperiodic sound present, the initial estimate of the amplitude ratio is determined from a different set of amplitude ratio histograms (which histogram to use is based on the value of the coincidence representation at the pitch period of the periodic sound source). When one of the sound sources is in an onset state, the initial estimate is determined from amplitude ratio histograms (which histogram to use is based on the value of the amplitude percent change, computed using equation 3.18). When there are two nonperiodic sounds present, the spectrum is split evenly between the two sound sources. For all of the other possible combinations of two sound sources (when one sound source is in the offset, increasing-periodicity, or decreasing-periodicity transitional states) the amplitude ratio estimate at that time frame is a linear interpolation between the amplitude ratio at neighboring

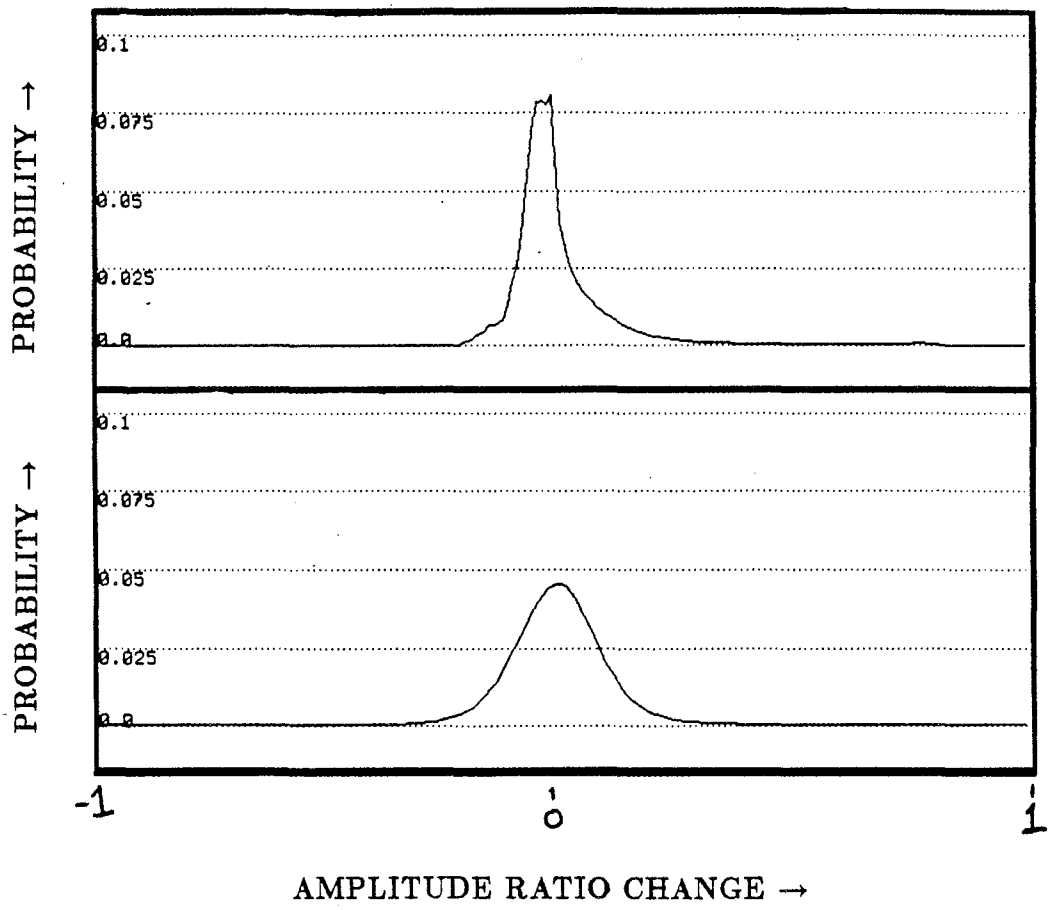

Figure 3.17: Top: The probability distribution of the amplitude change from equation 3.19 is shown for a single frequency channel. Bottom: The probability distribution of the amplitude change from equation 3.18 is shown for a single frequency channel.

time frames.

After the computation of an initial spectral estimate for each sound source, the system computes an iterative spectral estimate for each sound source. The amplitude ratio

$$\hat{R}(F,T,Iter) = \arg \min_{R(F,T)=0}^{\infty} [PerCost(R(F,T)) + SpecContCost(R(F,T), Iter)] \tag{3.21}$$

is computed when there are two periodic sound sources present, or when there is a

periodic and a nonperiodic sound source present. At each iteration, the amplitude ratio R(F,T) is varied between 0 and ∞ in 100 increments for each value of F (all frequency channels) and T (every time frame), and the amplitude ratio estimate is equal to the value of the amplitude ratio which minimizes equation 3.21. The cost at each frequency-time location is the sum of a periodicity cost and a spectral continuity cost. The periodicity cost

$$PerCost(R(F,T)) \stackrel{\text{def}}{=} -\log Probability[R(F,T) \mid SmoothNormCoin(F,T,P1,P2)] \tag{3.22}$$

is based on the probability that one would observe this amplitude ratio given the value of the smoothed and normalized coincidence representation at the pitch values. The spectral continuity cost

$$SpecContCost(F,T,Iter,R) \stackrel{\text{def}}{=}$$

$$\frac{1}{8} * \left\{ \begin{array}{r} AmpChngCost[possA1(F,T) \mid \widehat{A1}(F-1,T,Iter-1)]+ \\ AmpChngCost[possA1(F,T) \mid \widehat{A1}(F+1,T,Iter-1)]+ \\ AmpChngCost[possA1(F,T) \mid \widehat{A1}(F,T-1,Iter-1)]+ \\ AmpChngCost[possA1(F,T) \mid \widehat{A1}(F,T+1,Iter-1)]+ \\ AmpChngCost[possA2(F,T) \mid \widehat{A2}(F-1,T,Iter-1)]+ \\ AmpChngCost[possA2(F,T) \mid \widehat{A2}(F+1,T,Iter-1)]+ \\ AmpChngCost[possA2(F,T) \mid \widehat{A2}(F,T-1,Iter-1)]+ \\ AmpChngCost[possA2(F,T) \mid \widehat{A2}(F,T+1,Iter-1)] \end{array} \right\} \tag{3.23}$$

consists of eight terms. There are four amplitude change costs for each sound source since each frequency-time channel has four neighbors (the same frequency channel at the previous and next time frames, and the two neighboring frequency channels at the same time frame). Each amplitude change cost

$$AmpChngCost[possA1(F,T) \mid A1(F-1,T,Iter-1)] \stackrel{\text{def}}{=}$$

$$-logProbability\left[\frac{possA1(F,T)-A1(F-1,T,Iter-1)}{possA1(F,T)+A1(F-1,T,Iter-1)}\right] \tag{3.24}$$

is based on the probability that one would observe this amplitude change. When the amplitude ratio is varied between 0 and ∞, the spectral amplitude estimate that would result from using this estimate is computed from the sum-splitting formula discussed earlier in this section (equation 3.17).

$$possA2 = \frac{AS}{\sqrt{1+R^2}}\left[1 - \frac{1}{4}\left(\frac{R}{1+R^2}\right)^2 \frac{15}{64}\left(\frac{R}{1+R^2}\right)^4 - \ldots\right] \tag{3.25}$$

$$possA1 = possA2 * R \tag{3.26}$$

This iterative algorithm computes an estimate of the spectral amplitude of each sound source using local spectral continuity constraints. It attempts to compute the maximum likelihood spectral estimate for each sound source given the periodicity information present and the spectral continuity requirements of speech sounds. The iterative algorithm presented is not guaranteed to converge to the absolute minimum in the total cost function (because of bimodal probability distributions, see figure 3.16), and may reach only a local minimum. The spectral estimate obtained after 25 iterations is used as the final spectral estimate for each of the two sound sources.

3.6.4 Resynthesis

The output of the separation system is an estimate of the cochlear spectrum of each of the two sound sources present. Since the Kopec-Bush recognition system uses LPC spectral estimates, it was simpler to interface to the recognition system by resynthesizing a separated waveform than to convert the cochlear spectrum to LPC codebook entries. One advantage of resynthesizing the separated output is that one can listen to the separated output.

The cochlear model's filterbank output of the sum of the two speech sounds (before compression and half wave rectification) is used as the basis of the resynthesis process. Each frequency-time region is multiplied by the percent of the sound $\left(\frac{\widehat{A1}}{AS}\right)$ that belongs to this sound source, and a frequency gain (to compensate for the spectral tilt from the cochlear model). This separated cochlear output is then time reversed, and passed through a backwards original cascade filterbank

(reverse the arrows in figure 3.1). When the output of the backwards filterbank is time reversed again, the resulting waveform is the resynthesized output. The reason for time reversing the different waveforms is to compensate for the time delay imposed by the cochlear model in the low frequency regions.

The resynthesis of the separated output using this method has one important disadvantage. An example which exhibits this difficulty is the case of a sine wave of 100 Hz added to a sine wave of 110 Hz. The resulting signal is a sine wave of 105 Hz amplitude modulated at 10 Hz. If the separation system worked perfectly, it would correctly estimate that the amplitude of the 100 Hz sine wave was equal to the amplitude of the 110 Hz sine wave. However when the original sum is multiplied by the constant amplitude fraction $\left(\frac{A1}{AS}\right)$, the resynthesized output sounds like a softer version of the original sum waveform, and not like the original 100 Hz sine waveform.

3.7 Summary of Computational Model

This chapter has described a computational model which was developed to separate two simultaneous talkers. The computer model is based on the theory of sound separation developed in chapter two. The construction of this model has helped clarify what decisions need to be made by the auditory system when it separates sounds.

The computer model of sound separation consists of several stages of processing. The input to the separation algorithms is the computational model of the cochlea developed by Lyon. The timing information contained in each frequency channel is then converted into an event representation, which is used to compute local periodicity information called the coincidence function.

Two different separation systems were developed. The current separation system uses a handlabeled database of connected digit strings to compute the probability distributions used for both the Markov model and the iterative spectral estimation algorithms.

The current separation system consists of four sequential steps: (1) an iterative dynamic programming pitch tracking algorithm to determine the pitch period

of each sound source, (2) a Markov model to determine the number of sounds present (i.e., one or two) and the type of each sound source (i.e., periodic or non-periodic), (3) an iterative algorithm to estimate the amplitude (in each frequency channel) of each sound source present, using both periodic information and spectral continuity constraints, and (4) an algorithm to resynthesize a waveform of the separated output.

Chapter 4

Evaluation

Whereas chapter two presented a theory of how the auditory system separates sounds, the precise details of how the auditory system separates sounds are not known. Chapter three described a computer model that separates two simultaneous speakers based on this theory of auditory sound separation. The computational model used only some of the acoustic information that the auditory system is known to use (periodicity, onsets, spectral continuity), and did not use any higher level knowledge to aid in the separation of the two speakers.

The computational model is a 'functional model' since it hypothesizes both the output of the auditory system and a mechanism to compute these quantities. It is not claimed that these computer algorithms are the same ones used by the auditory system to separate sounds. What is claimed is (1) that the auditory system performs similar operations and tries to compute similar quantities (the fundamental frequency of each speaker over time, when each person starts and stops talking, and a spectral estimate of each talker) and (2) that the auditory system uses similar representations (such as the coincidence function) when separating sounds.

The aim of the computational model of sound separation is to improve the ability of computers to recognize sounds in a noisy environment. Different algorithms in the computer model which have specific subgoals (e.g. to track the pitch period of each of two simultaneous speakers) are evaluated on how well they perform the tasks they were designed to achieve. The performance evaluation reflects

how well these algorithms are able to achieve the different tasks, and does not directly reflect the ability of the theory to explain auditory sound separation.

This section discusses the results of the following algorithms: (1) pitch tracker for two simultaneous speakers, (2) Markov model for determining how many people are speaking and the characteristics of each speaker, and (3) iterative spectral estimation algorithms for computing a spectral estimate of each sound source. The ability of a recognition system [Kopec and Bush 1985] to determine what each of the two speakers has said is also presented.

No performance evaluation of the cochlear model or the coincidence representation is presented. These algorithms are evaluated indirectly based on the performance of the different systems that use this information. To evaluate the effect of changing parameters of these algorithms (such as changing the time constant for computing periodicity information in the coincidence representation) would require training and testing the complete system for each value of the parameter that is being varied. The computational requirements for this type of evaluation are not feasible at this time.

4.1 Experimental Results of Computer Model

The Markov model and the spectral estimation algorithms both rely on probability distributions for computing how many sounds are present and the spectral estimate of each sound source. In order to compute these probability distributions, a database of simultaneous speech sounds was constructed. The database consists of 39 single-speaker digit strings (of seven continuous digits) spoken by two males and two females. It also consists of 38 examples of dual-speaker digit strings (obtained by adding the single-speaker waveforms of a male and a female speaking different digit strings). The system was trained and tested on the same database of speech sounds, and was limited in size for computational reasons (eg, limits on disk storage, computer time necessary to process the data). Since the system was trained and tested on the same database, the model's performance might decrease when tested on a new database of simultaneous sounds.

The advantage of using single-speaker recordings over using recordings of two

simultaneous speakers is that the accuracy of the spectral estimation algorithms can be computed. By using the sum of two single-speaker recordings as the input to the separation system, the original spectrum of each speaker is known. The spectral estimate of each of the two speakers (produced by the computer model) can then be compared with the spectrum of the original waveforms to evaluate the spectral estimation procedures.

Recordings of speakers made during the presence of other interfering sounds show that people increase their speech level as the background noise level increases in amplitude [Pearsons et. al. 1976]. Their comparison of speech levels to background noise levels show that "... people maintain about a 5 - 8 dB speech to noise ratio when conversing outside their homes and a 9 - 14 dB speech to noise ratio when talking inside their homes. " If the speakers had been recorded in the presence of other talkers, each of the speakers might have changed how they spoke to compensate for the interfering sounds (by changing their speech level, and the clarity of pronunciation).

The speech database of isolated sounds was handmarked according to the Markov model presented in section 3.6.2. Every 10 msec was assigned one of the seven possible labels (silence, periodic, nonperiodic, onset, offset, increasing transitional periodicity, or decreasing transitional periodicity). This labeling was used for both training and evaluation of the separation system.

4.1.1 Pitch Tracker Accuracy

The pitch algorithm is evaluated by comparing the pitch tracks computed on two simultaneous sounds with those computed on the isolated pitch tracks (it was not compared with a handmarked pitch track). The pitch period used for the isolated pitch track is equal to the location of the maximum of the average coincidence representation (see bottom picture in figure 3.14) of the isolated sound. The pitch tracks for each of the two simultaneous sounds were computed using the iterative dynamic programming algorithm (described in section 3.6.1).

The results presented below are categorized by the types of sounds present. The two important cases are (1) when one sound is periodic and the other sound

SNR (in dB)	# Frames in SNR Interval	Pitch Errors (% of frames with error magnitude = x samples) 0	1	2	3	4	5	> 5
$SNR \leq -5$	28	17.9	14.3	7.1	3.6	3.6	3.6	50.0
$-5 < SNR \leq 0$	107	36.4	24.3	6.5	4.7	1.9	3.7	22.4
$0 < SNR \leq 5$	487	51.3	22.6	6.4	3.5	2.3	.4	13.6
$5 < SNR \leq 10$	872	66.3	20.1	2.4	1.7	.2	.5	8.8
$10 < SNR \leq 15$	971	70.3	17.8	2.7	1.3	.4	.2	7.2
$15 < SNR \leq 20$	769	71.1	18.1	1.8	1.7	.9	.5	5.9
$20 < SNR \leq 25$	545	77.6	14.5	1.8	.4	.4	.7	4.6
$25 < SNR \leq 30$	385	74.0	16.6	1.6	.8	.5	.3	6.2
$30 < SNR$	408	83.1	7.1	.7	.5	.0	.5	8.1
Average	4572	68.9	17.5	2.6	1.6	.7	.5	8.3

Table 4.1: Distribution of pitch errors of the periodic sound when one sound source is periodic and the other sound source is nonperiodic [as a function of the periodic to nonperiodic SNR (signal to noise ratio)].

is nonperiodic, and (2) when both sounds are periodic. When neither sound was labeled as periodic, the pitch tracks have no meaning.

Table 4.1 shows how the accuracy of the pitch tracking algorithm varies as a function of the signal to noise ratio (of the periodic sound source to the nonperiodic sound source). The frames in the database of two simultaneous digit strings were categorized both by the types of sounds present and the local signal to noise ratio between the two sounds. The signal to noise ratio

$$SNR = 10 \log_{10} \frac{E_1}{E_2} \tag{4.27}$$

is computed from the local energy ratio between the two sound sources. The energy of each sound source (in equation 4.1) was computed by smoothing the local energy (energy in a 10 msec window) using an exponential window with a 15 msec time constant. The pitch error was computed by comparing the iterative dynamic programming algorithm pitch track with the pitch period computed on the isolated sound sources (computed from the location of the maximum of the

average coincidence function: see bottom picture in figure 3.14).

The table shows that as the energy of the periodic sound increases relative to the nonperiodic sound, the accuracy of the pitch period increases. The accuracy of the periodic pitch period decreases rapidly as the signal to noise ratio decreases below 5 dB. The accuracy of the pitch tracking algorithm (in table 4.1) is different from computing the pitch track of a single sound source in the presence of a steady state nonperiodic noise (computing the best single pitch track, rather than determining the pitch of each of two sound sources).

When one sound source was periodic and the other sound source was nonperiodic, the pitch period of the periodic sound was not determined from the dominant pitch track (but from the pitch track of the weaker periodic sound) in 8.8% of the frames. This can be attributed to one of two factors: (1) the periodic sound had just started and the peak in the coincidence function was not yet large enough to be considered as the actual pitch period, or (2) the nonperiodic sound had just switched from being periodic to being nonperiodic, and the dominant pitch track which followed this sound source's pitch period had not yet switched over to the other periodic sound.

Of the frames when the pitch error of the periodic sound was greater than 5 samples (at a sample rate of 16 KHz), (1) 38.1% of the time the pitch period of the periodic sound was not obtained from the dominant pitch track, and (2) 27% of the time the pitch period of the periodic sound was approximately half of the isolated pitch track; these 'errors' can be attributed to the pitch period doubling of the isolated pitch track and are not really errors.

Table 4.2 shows the accuracy of the dominant pitch track, while table 4.3 shows the accuracy of the weaker pitch track. These tables shows that as the energy of the dominant periodic sound increases relative to the weaker periodic sound, the accuracy of the dominant pitch period increases while the accuracy of the weaker periodic sound decreases. The accuracy of the dominant pitch period decreases when the signal to noise ratio drops below -5 dB. The accuracy of the weaker periodic sound is much lower than the accuracy of the dominant periodic sound.

SNR (in dB)	# Frames in SNR Interval	Pitch Errors (% of frames with error magnitude = x samples) 0	1	2	3	4	5	> 5
$SNR \leq -5$	79	57.0	13.9	8.9	1.3	2.5	.0	16.5
$-5 < SNR \leq 0$	893	62.5	12.2	1.3	1.3	1.3	1.0	20.3
$0 < SNR \leq 5$	2121	68.9	14.9	2.3	.8	.6	.5	12.0
$5 < SNR \leq 10$	756	79.6	13.2	1.1	.8	.8	.4	4.1
$10 < SNR$	139	77.7	14.4	.7	.7	.7	.7	5.0
Average	3988	69.6	13.9	1.9	1.0	.9	.6	12.2

Table 4.2: Distribution of pitch errors of the dominant periodic sound when both sound sources are periodic [as a function of the dominant-periodic to weaker-periodic SNR (signal to noise ratio)].

SNR (in dB)	# Frames in SNR Interval	Pitch Errors (% of frames with error magnitude = x samples) 0	1	2	3	4	5	> 5
$SNR \leq -5$	79	43.0	34.2	10.1	1.3	.0	.0	11.4
$-5 < SNR \leq 0$	893	34.8	28.3	9.7	5.5	2.6	1.8	17.2
$0 < SNR \leq 5$	2121	33.1	21.8	9.2	5.1	4.3	2.8	23.7
$5 < SNR \leq 10$	756	19.8	16.3	9.3	6.7	5.2	4.9	37.8
$10 < SNR$	139	10.8	11.5	7.2	10.8	5.0	2.2	52.5
Average	3988	30.4	22.1	9.3	5.6	4.0	2.9	25.7

Table 4.3: Distribution of pitch errors of the weaker periodic sound when both sound sources are periodic [as a function of the dominant-periodic to weaker-periodic SNR (signal to noise ratio)].

When two periodic sounds are present, the errors can be divided into three main categories: (1) 3.2% of the time the dominant pitch period was correctly determined but was assigned to the wrong sound stream (the assignment of the dominant pitch period is based on which sound stream's average pitch period is closer), (2) 2.6% of the time the dominant pitch track is not assigned to either sound stream (this occurs when the difference between the dominant pitch period and the average pitch period for either sound stream is greater than 60% of the average pitch

period of that sound stream), and (3) 7.0% of the time there was a large error in the dominant pitch period which could be attributed to a pitch doubling error [defined as $ABS(DominantPP - ActualPP_i) > MAX(5, ABS[.5(DominantPP - ActualPP_j)])$]. These pitch doubling errors account for 57.8% of the times when the pitch error of the dominant pitch track is greater than 5 samples.

The average pitch period of each talker (computed from the isolated pitch tracks) was used by the pitch tracking system to assign the dominant pitch period to the sound stream with the closest average pitch period. The system was tested using the digit strings from a male and female talker. In assigning the dominant pitch period to one of the two talkers, the average pitch difference between the male and female talkers is very helpful to the system. If two male talkers with the same average pitch period were speaking at the same time, the pitch algorithm would probably have made many errors in assigning the dominant pitch period to the wrong sound stream. A more sophisticated method of assigning pitch periods to sound streams is needed.

The literature contains only one algorithm for determining the pitch period of each of two simultaneous talkers [Parsons 1976]; but as far as the author knows, this section represents the first quantitative evaluation of a dual-speaker pitch tracking algorithm. Future research on the dual-speaker pitch tracker may deal with optimizing pitch period accuracy by varying some of the many parameters in the iterative dynamic programming pitch tracking algorithm (such as changing the pitch transition cost in the dynamic programming algorithm, the time constant used in computing the coincidence function, the addition of a dominant frequency weighting so that some frequency regions are more important for the determination of the pitch period, etc.).

4.1.2 Hypothesis Determination Accuracy

The Markov model was used to determine both how many sounds are present and the characteristics of each sound. The system determined the best path through the state transition network for labeling each of the two sound sources. The state transition network is used by the Markov model to maintain continu-

ity in assigning the characteristics of each sound. The model was evaluated by comparing the labels assigned to each of the two simultaneous sounds with the manually-labeled database.

One constraint imposed on the two sound model was that only one sound source could be in a transitional state in any time frame. In the two sound database, only .3% of the frames were handmarked with both sound sources being simultaneously in a transitional state.

The probability distributions used by the Markov model were derived from both the one sound and two sound database, using the pitch periods computed from the isolated pitch tracks. Because the Markov model was constructed before the iterative dynamic programming pitch track was developed, the system was tested using both the isolated pitch tracks and the pitch tracks obtained from the iterative dynamic programming algorithm. The computational model was developed in this way (first the spectral estimation procedure was developed, then the Markov model and finally the pitch tracker) to use the assumptions about the types of sounds present for the spectral estimation procedure. After the spectral estimation algorithm had been developed, the Markov model and pitch tracker were developed to compute the desired quantities.

Table 4.4 shows the overall accuracy of the two sound source Markov model. The first case uses the pitch tracks which were computed on the isolated digit strings (before they were added together). The second case uses the iterative dynamic programming pitch tracks (described in section 3.6.1).

The Markov model correctly identified 73% of the frames (the labels on both the male and female speakers were correct) when the system was tested on the same database of simultaneous sounds that it was trained on. When the Markov model used the iterative dynamic programming pitch tracks, the accuracy dropped to 26% of the frames when both sounds were correctly identified. The Markov model performed quite well (98.3%) in correctly labeling at least one of the two sound sources.

When the Markov model used the iterative dynamic programming (IDP) pitch tracks, 74% of the time it did not meet the requirement of labeling both

	Percent of States Correctly Identified			
	Female	Male	Male and Female	Male or Female
Isolated Pitch Track	84.8%	86.7%	73.3%	98.3%
IDP Pitch Track	46.9%	50.4%	26.0%	71.3%

Table 4.4: Overall accuracy of the Markov model when two simultaneous sounds are present.
The results are shown for the case when the isolated pitch tracks are used, and when the iterative dynamic programming pitch tracks are used. The results show the accuracy of the female talkers, the male talkers, when both male and female were identified correctly, and when either of the two were identified correctly.

sounds correctly. There are two main categories of errors in the labeling process: (1) 18.3% are due to the system reversing the labels of the two speaker's (the female speaker was assigned the correct label of the male speaker, and the male was assigned the female speaker's label), and (2) 54.4% are due to the system labeling one of the speakers as silent when that speaker was not silent.

The percentages in table 4.4 reflect the accuracy of the Markov model's labels on the 10101 frames in the two sound database (a total of 101 seconds of two simultaneous talkers, at a frame rate of 100 per second). The accuracy of the Markov model is computed for each of the different types of sounds present. Tables 4.5 through 4.9 show the accuracy of the Markov model for each of the major categories of simultaneous sounds (two simultaneous periodic sounds, periodic and nonperiodic sound, two nonperiodic sounds, one periodic sound, and one nonperiodic sound).

Table 4.5 shows the accuracy of the Markov model labeling when both simultaneous sounds were handlabeled as periodic. The Markov model correctly labeled the two sounds as periodic 93.6% of the time when the isolated pitch tracks were used. When the iterative dynamic programming pitch tracks were used, the system correctly labeled only 32.2% of the frames. Most of the errors are due to the system deciding that there was only one periodic sound present. This performance can be attributed to the fact that the Markov model was trained using the isolated

Label Assigned to Two Periodic Sounds	Isolated Pitch Track	IDP Pitch Track
Periodic, Periodic	93.6%	32.2%
Periodic, Silence	1.2%	46.4%
Periodic, Nonperiodic	3.0%	18.6%
Other	2.2%	3.0%

Table 4.5: Accuracy of the Markov model labeling when two periodic sounds are present.

pitch tracks; when the system was tested using the iterative dynamic programming pitch tracks, the system frequently failed to recognize the pitch period of the weaker periodic sound.

Table 4.6 shows the accuracy of the Markov model labeling when one sound was periodic and the other sound source was nonperiodic. The system correctly labeled 55.4% of the frames when the isolated pitch tracks were used, but only 11.2% of the frames when the IDP pitch tracks were used. When the IDP pitch tracks were used, the system reversed the labeling of the two sound sources 26.2% of the time, and 52% of the time decided that there was only one periodic sound present. The Markov model did not make these errors when it used the isolated pitch tracks since the pitch track of the nonperiodic sound would be at some random location with respect to the periodic sound.

Table 4.7 shows the accuracy of the Markov model labeling when both sounds were nonperiodic. The Markov model correctly labeled both sounds as nonperiodic 88.9% of the time when the isolated pitch tracks were used. When the Markov model used the IDP pitch tracks, 42.4% of the time the system labeled the sounds as one nonperiodic sound, and 34.3% as one nonperiodic and one periodic sound. The reason for this performance difference is that the system performed better when it was tested with the same pitch values as the training data than when it was tested using the IDP pitch periods.

Table 4.8 shows the accuracy of the Markov model labeling when one sound source was periodic and the other sound source was silent. When the Markov

Label Assigned to Periodic and Nonperiodic Sound	Isolated Pitch Track	IDP Pitch Track
Periodic, Nonperiodic	55.4%	11.2%
Nonperiodic, Periodic	.2%	26.2%
Periodic, Silence	6.7%	25.2%
Silence, Periodic	.0%	26.8%
Silence, Nonperiodic	1.0%	.8%
Periodic, Periodic	27.8%	7.0%
Other	8.9%	2.6%

Table 4.6: Accuracy of the Markov model labeling when one periodic sound and one nonperiodic sound is present.

Label Assigned to Two Nonperiodic Sounds	Isolated Pitch Track	IDP Pitch Track
Nonperiodic, Nonperiodic	88.9%	5.6%
Nonperiodic, Silence	6.3%	42.4%
Nonperiodic, Periodic	.0%	34.3%
Other	4.7%	17.6%

Table 4.7: Accuracy of the Markov model labeling when two nonperiodic sounds are present.

model was tested using the IDP pitch tracks, 76.8% of the time it would correctly

Label Assigned to Periodic and Silent Sound	Isolated Pitch Track	IDP Pitch Track
Periodic, Silent	62.4%	36.5%
Silent, Periodic	.0%	40.3%
Nonperiodic, Nonperiodic	24.6%	3.3%
Other	13.0%	19.9%

Table 4.8: Accuracy of the Markov model labeling when one periodic sound is present.

Label Assigned to Nonperiodic and Silent Sound	Isolated Pitch Track	IDP Pitch Track
Nonperiodic, Silence	49.7%	25.0%
Silence, Nonperiodic	8.5%	35.1%
Nonperiodic, Nonperiodic	31.3%	2.5%
Nonperiodic, Periodic	.3%	11.4%
Other	10.1%	25.9%

Table 4.9: Accuracy of the Markov model labeling when one nonperiodic sound is present.

determine that there was only one sound present, but it made a large number of errors in deciding which of the two sound streams was periodic.

Table 4.9 shows the accuracy of the Markov model labeling when one sound source was nonperiodic and the other sound source was silent. Using the IDP pitch tracks, 60.1% of the time the Markov model correctly determined that there was only one nonperiodic sound present, but it made errors in determining which of the two sounds was nonperiodic.

This algorithm represents the first system that can determine both how many sounds are present and recognize the characteristics of each sound source. The limited performance of the Markov model is due to both the difficulty of determining the correct label for each of two simultaneous sounds and to the simplicity of the algorithm used. Future research on determining how many sounds are present and the characteristics of each sound can focus on: (1) training the system using iterative dynamic programming pitch tracks, (2) conditioning the probability of a particular type of sound on more data points than just the value of the smoothed coincidence function at the pitch period, and (3) implementing some of the information sources for assigning group objects to sound streams presented in chapter two.

4.1.3 Spectral Estimation Accuracy

Most speech recognition systems rely on spectral estimates of the speech utterance to classify what words a talker has spoken. In order for a recognition system to correctly classify the speech of a talker, it is essential that the system use an accurate estimate of the spectrum of that talker. The output of the spectral estimation procedure is an estimate of the average cochlear spectrum for each of the two sounds present. This spectral estimate can be compared with the spectrum of each of the original sounds to evaluate how accurately the algorithm has estimated the spectrum of each sound source.

The cochlear spectrum consists of the average output of each of the 85 frequency channels of the cochlear model. Each cochlear spectrum is normalized in amplitude by the following equation:

$$Norm(Spec_i) = \frac{1}{\sqrt{\sum_{j=1}^{85} s_i(j)^2}} Spec_i \tag{4.28}$$

$$\%Spectral\ Distance = \frac{EuclidDist[Norm(OrigSpec_i), Norm(SepSpec_i)]}{EuclidDist[Norm(OrigSpec_i), Norm(SumSpec)]} \times 100\% \tag{4.29}$$

The accuracy of the spectral estimation algorithm is computed by comparing the euclidean distance between the original and separated cochlear spectrum with the euclidean distance between the original and the cochlear spectrum of the simultaneous sounds. If the separated cochlear spectrum is very close to the original cochlear spectrum, this fraction will be small. If the percent spectral distance (equation 4.3) is less than 100%, this means that the separation program has improved the spectral estimate (over the case of no separation). If the percent spectral distance (equation 4.3) is greater than 100%, this means that the spectrum of the sum of the two simultaneous sounds is a better estimate of the original spectrum than the estimate obtained from the separation program.

Table 4.10 shows how the spectral estimation procedure improves the spectral estimate of the weaker sound (from a cumulative spectral distance of 856 down to 669) but does not improve the spectral estimate of the stronger sound

	Stronger Sound	Weaker Sound
Both Sounds Correctly Labeled	$\frac{322}{296} = 108.8\%$	$\frac{669}{856} = 78.2\%$

Table 4.10: Percent spectral distance of the spectral estimation procedure when both sounds are correctly labeled.

	Correctly Labeled Sound	Incorrectly Labeled Sound
One of Two Sounds Correctly Labeled	$\frac{1016}{1018} = 99.7\%$	$\frac{1751}{1213} = 144.3\%$

Table 4.11: Percent spectral distance of the spectral estimation procedure when one of the two sounds are correctly labeled.

source. Table 4.11 and 4.12 show that when the system has incorrectly labeled the sounds, the spectral estimation procedure results in a deterioration of the spectral estimate of each sound source.

Table 4.13 shows how the spectral estimation procedure improves the spectral estimate when two periodic sounds are present and the correct control information is used. Both the initial spectral and the iterative spectral estimation procedure provide spectral estimates better than the original unseparated spectral estimate. The last line in table 4.13 is the accuracy of the separated estimate if the actual spectral ratio (equation 3.20) is known. Table 4.14 shows similar improvements for a periodic and nonperiodic sound source with known control information. The first line in table 4.14 is the spectral improvement that would result if a Wiener filter had been used to separate the periodic and nonperiodic sound sources (a Wiener filter can be used since the average spectrum of a periodic sound is different from a nonperiodic sound).

Figure 4.1 shows the spectral distance improvement as a function of the number of iterations of the spectral estimation algorithm. Initially, the estimation procedure improves the spectral estimate of each sound source with each iteration. However, the spectral distance reaches a minimum after approximately 10 iterations and begins a gradual increase in the spectral distance. The iterative

	Stronger Sound	Weaker Sound
Neither Sound Correctly Labeled	$\frac{1164}{358} = 324.9\%$	$\frac{1499}{1270} = 118.0\%$

Table 4.12: Percent spectral distance of the spectral estimation procedure when neither sounds is correctly labeled.

	Percent Spectral Distance
Initial Separation Estimate	78.5%
Separation Estimate after 10 Iterations	69.3%
Separation using actual ratio between sounds	38.8%

Table 4.13: Percent spectral distance for two periodic sounds using isolated pitch tracks and correct state labels.

algorithm is not guaranteed to converge to the global minimum in the spectral distance contour. The bimodal shape of some of the probability distributions (see figure 3.16) is one reason why the algorithm is not guaranteed to converge.

This system is the first algorithm to compute a spectral estimate of each sound source using spectral continuity constraints. It has the ability to compute a spectral estimate when two periodic sounds are present, or a periodic and a nonperiodic sound is present. The spectral estimation procedure does not work well when errors are made in either the determination of the how many sounds are present or in the types of sounds present.

	Percent Spectral Distance	
	Periodic Sound	Nonperiodic Sound
Wiener Filter	72.3%	59.6%
Initial Separation Estimate	65.4%	50.5%
Separation Estimate after 10 Iterations	55.5%	41.5%
Separation using actual ratio between sounds	28.2%	17.7%

Table 4.14: Percent spectral distance for a simultaneous periodic and nonperiodic sound using isolated pitch track and correct state labels.

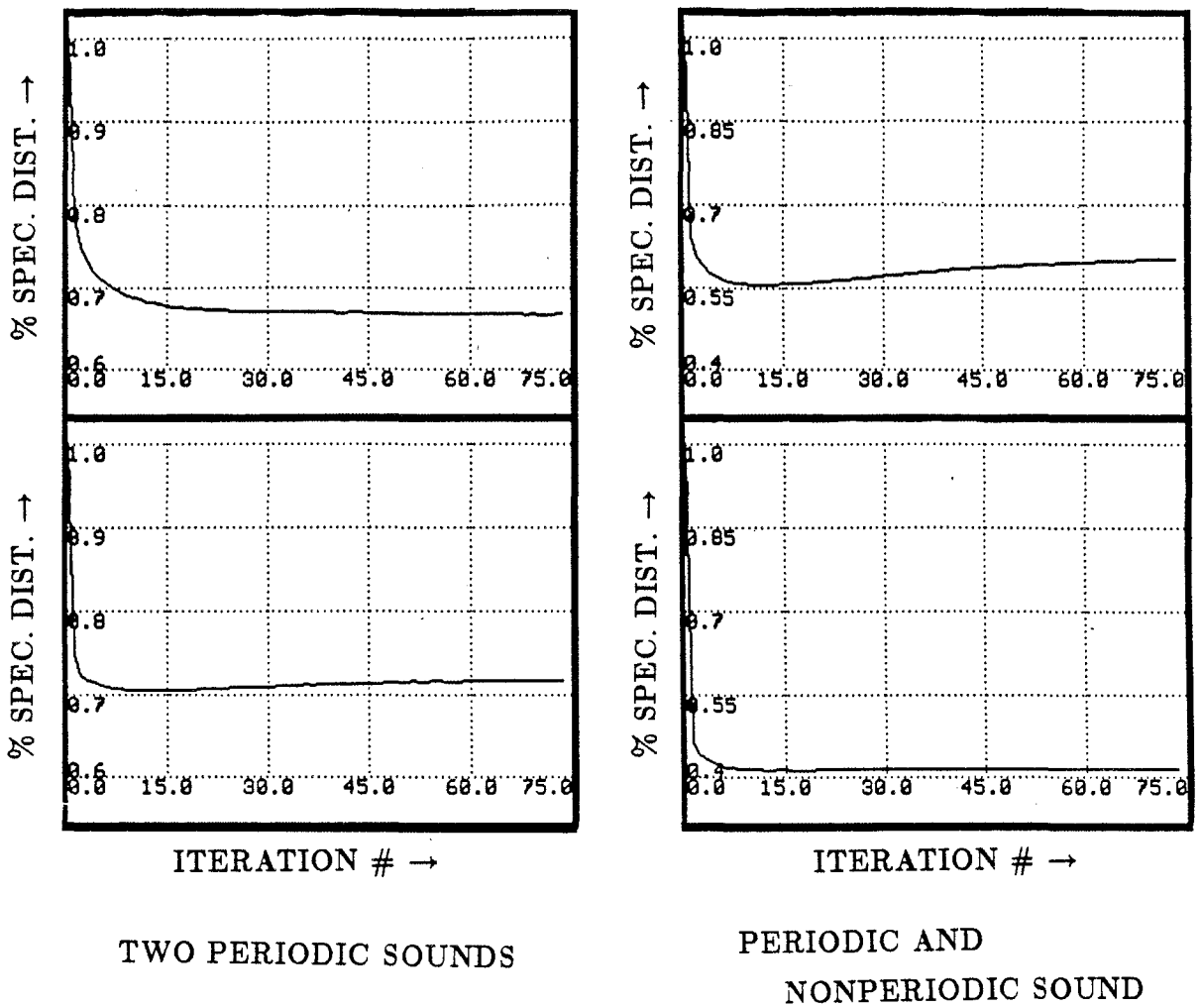

Figure 4.1: Spectral distance improvement as a function of the number of iterations.
Left Side: Spectral distance improvement when both sounds are periodic (top is louder periodic sound and bottom is weaker periodic sound). Right Side: Spectral distance improvement when one sound is periodic and one is nonperiodic (top is periodic sound and bottom is nonperiodic sound).

4.1.4 Recognition Accuracy

One goal of the sound separation system is to improve the ability of computers to recognize speech in the presence of interfering sounds. The Kopec-Bush [85] recognizer which was used to test the separated output, uses a network based approach to speaker-independent continuous digit recognition. The recognition system is designed for isolated sounds, and computes the best path through the

	Correct	Substitutions	Insertions	Deletions
Isolated Sounds	87.0%	10.5%	2.5%	1.4%
Resynthesized Isolated Sounds	86.1%	10.6%	3.3%	.0%
Two Simultaneous Sounds				
Male	44.4%	36.4%	19.2%	3.8%
Female	34.2%	39.6%	26.2%	3.5%
Separated Simultaneous Sounds				
Male	57.0%	39.6%	17.8%	18.7%
Female	31.7%	52.6%	15.7%	20.9%

Table 4.15: Accuracy of the recognition system.

network using the gain-optimized Itakura-Saito distance measure between vector quantized LPC spectral templates and the LPC spectrum of the incoming sound.

The accuracy of the recognized output was computed by comparing the string of digits that each person actually said with the output digit string of the recognition system. Every possible way of matching the recognition string of digits with the actual digit string is computed to find the match that yields the highest percentage of correct digits. The errors made by the recognizer are categorized as substitutions (recognizer output is one digit while the person said a different digit), insertions (recognizer has an extra digit), and deletions (recognizer missed one of the original digits). The percentage of correct digits (equation 4.4) reflects the accuracy of the recognition system.

$$\%\ Correct\ Digits = \frac{\#Digits\ Correctly\ Recognized}{\#Digits\ Hypothesized\ by\ Recognizer} \times 100\% \qquad (4.30)$$

Table 4.15 shows the results of the recognition system evaluation. The recognizer was tested on the original database of isolated sounds and the resynthesized version of each of the isolated digit strings. The results show that the analysis and resynthesis of a digit waveform did not noticeably decrease the performance of the recognition system. The system was also tested on the two simultaneous talkers without the aid of any separation system. The output of the recognition system

was compared with each of the actual digit strings of both the male and female talkers. The recognition system was also tested on the individual resynthesized waveforms of both the male and the female talkers.

The separation system improved the recognizer's ability to recognize the male voice from 44.4% to 57.0%. It did not improve the recognition performance of the female voice. The increase in deletions by the recognition system when tested on the separated output (from 3% to 20%) reflects the fact that the separation system often missed the detection of the weaker sound source (40% of the frames were labeled as one sound being present by the separation system when there were two sounds present; since the system decided that there was only one sound present, one of the sound streams was silent during these intervals).

These results represent the first quantitative evaluation of a computer recognition system attempting to recognize each of two simultaneous talkers.

4.2 Overview

This chapter presented an evaluation of a system which attempts to separate two simultaneous talkers. The evaluation consisted of (1) the iterative dynamic programming pitch tracker's ability to correctly determine the pitch period of each sound source, (2) the determination by the Markov model of how many sounds are present and what the characteristics of each sound source is, (3) the accuracy of the spectral estimate of each of the two sounds present, and (4) the results of the recognition system which attempted to recognize each of the two simultaneous talkers. The results obtained suggest that there is a need for improvement in each of these algorithms in order to achieve a high level of separation performance.

There are many different algorithmic modifications which can be tested to try to improve the system's performance. The next chapter will outline some of the major changes which this author viewes as important improvements in the computational model of auditory sound separation.

Chapter 5

Future Directions

The computational model presented in this thesis represents an important step in allowing computers to determine both how many sounds are present and what each speaker is saying. The separation algorithms are based on the theory of auditory sound separation (presented in chapter two). Currently, the algorithms use only some of the available monaural acoustic information sources, and do not use any of the higher level processes described in the theory of auditory sound separation.

Not only can one use additional information to improve the current separation system, but one can also modify the separation algorithms that are currently used. This chapter will describe changes that are possible in the current separation algorithms and some of the different information cues that can be added to help separate two simultaneous sounds.

5.1 Modifications in the model

5.1.1 Improved Two Talker Pitch Tracking

There are many possible modifications that one can make to the pitch tracking algorithms described in section 3.6.1. While the pitch tracker was able to follow the pitch period of the dominant periodic sound fairly well, the pitch error on the weaker periodic sound was fairly large. By modifying some of the parameters of the pitch tracking algorithm (and subsequently testing each modification with

respect to performance of the pitch tracker), a better two talker pitch tracker can be constructed. These parameters include: the method of assigning the dominant pitch period to one of the two sound streams, the transition cost used in computing both the dominant and secondary pitch period, the function used to compute the score of the weaker pitch period, the signal used to smooth each row of the coincidence representation, and the different parameters used in computing the coincidence representation.

Even if one could precisely determine both pitch periods, one still has to determine which sound stream these pitch periods belong to. When both talkers have the same average pitch period, determining which pitch period belongs to which talker is a difficult problem which has no straightforward solution.

5.1.2 Improved Spectral Estimation

The spectral estimation algorithms described in section 3.6.3 used an iterative algorithm to combine both periodicity information and spectral continuity constraints.

The periodicity information consisted of probability distributions of the amplitude ratio $\left(\frac{A1}{AS}\right)$ conditioned on the value of the coincidence function at different locations. If the two sounds consisted of one speaker's voice being periodic while the other speaker's voice was nonperiodic, the amplitude ratio was conditioned on the value of the smoothed and normalized coincidence function (see top right picture in figure 3.13) at the pitch period of the periodic speaker. The pitch period was obtained from the iterative dynamic programming pitch tracker, and the decision whether the sound was periodic or nonperiodic was obtained from the output of the Markov model. If both sound sources were periodic, the amplitude ratio was conditioned on the value of the coincidence function at P_1, P_2, and $P_{diff} = Abs(P_1 - P_2)$.

Although the values of the coincidence function at every location were used in the determination of the pitch period for each of the two speakers, only the values of the coincidence function at a few select points (related to the pitch period) were used when computing the amplitude ratio of the two sounds based on

periodicity information. The width of the amplitude ratio histograms reflects the uncertainty in the amplitude ratio information based on the value of the coincidence function at the pitch period. The distributions might have narrowed in width (and therefore uncertainty) if the amplitude ratio were conditioned on the value of the coincidence function at additional delay locations. However, in order to condition the amplitude ratio on more information, more probability distributions are needed. A larger database would be required to accurately compute all of these new probability distributions.

The spectral estimation algorithms used depend on accurate pitch information. Information about the accuracy of the pitch tracks could be incorporated into the spectral estimation algorithms in several different ways. By increasing the width of the AC smoothing filters (see figure 3.12), the value of the smoothed coincidence function around the pitch period will not vary as much, and will therefore be less sensitive to small pitch errors. An increase in the smoothing function's width may broaden the probability distributions (which are conditioned on the periodicity information) and therefore increase the uncertainty of the amplitude ratio. Therefore, one would like to use as narrow a smoothing filter as is possible. One possible solution would be to first predict the accuracy of the pitch track, and to use a smoothing algorithm whose width depended on the predicted accuracy of each pitch track. When the pitch periods obtained from the iterative dynamic programming pitch tracker cannot be determined very accurately, spectral estimation algorithms would use a wider smoothing function than when the pitch periods could be determined with more precision. (Note: in the case of speech enhancement, this would correspond to varying the width of the comb filter based on the accuracy of the pitch period)

5.1.3 Assignment of Group Objects to Sound Streams

The current separation system assigned group objects to sound streams using a very simple mechanism. If the Markov model determined that there was a periodic group object present that overlapped in time with a nonperiodic group object, the nonperiodic group object was assigned to the other sound stream from

the periodic group object. The periodic group object was assigned to a sound stream based on which sound stream had a closer average pitch period to the pitch period of the group object.

If the nonperiodic group object did not overlap in time with any other group object, the separation system would not know which sound stream to assign this nonperiodic group object to. The system did make errors in assigning group objects to the wrong sound stream.

The theory of auditory sound separation described many different sources of information that could be used by the auditory system in determining which group object belonged to which sound stream (see section 2.4.5.3). The current computational model implemented only one of these information cues. The addition of these other information sources could be used to improve the assignment of group objects to sound streams.

The use of spectral continuity between group objects is one of the acoustic information sources that can possibly improve the assignment of group objects to sound streams. A change by the vocal cords may alter the characteristics of speech from periodic to nonperiodic, but the spectral transition between the periodic and nonperiodic speech segments will reflect the continuity in the articulatory domain. This spectral continuity between the spectrum at the onset of one group object with the spectrum at the offset of a previous group object can be used by a separation system in the assignment of group objects to sound streams.

5.1.4 Addition of a 'MASKED' Hypothesis

One of the major sources of error in the current separation system was the determination of how many sounds are present. In order to determine how many sounds are present, the system must determine how many hypotheses are needed to explain the data at the current time.

All mistakes in the determination of how many group objects are present and when group objects start and stop lead to errors which are compounded by the spectral estimation procedure. When the system misses the detection of a second group object, all the energy is mistakenly assigned to the only group object

present. When the system detects a group object which is not present, energy will be assigned to this group object even though there is no sound source present. Even when a group object is correctly determined to be present, mistakes about when this group object starts and stops lead to similar types of errors. For example, if the system starts a group later than the group object actually starts, then additional energy will be assigned to the other sound source and will not be assigned to the sound source which is actually present.

As the second sound source gets weaker and weaker in amplitude (with respect to the first sound source), it becomes more and more difficult to determine when group objects from this sound source are present, and when these group objects start and stop. At some amplitude level, it may be impossible to determine what is happening to the weaker sound source. Only during silent intervals of the stronger sound source can the system gather accurate information about the second sound source.

Since it can be very hard to determine the properties of the weaker sound source, the addition of an extra **masked** state in the state transition diagram (see figure 3.15) might be appropriate. When the system believes that a second sound source is present, but it can not determine what the characteristics of that second sound source are, the masked hypothesis would be the appropriate label for the state of this sound source. The addition of a masked state would not allow for this sound to be separated from the stronger interfering sound, but it would allow a recognizer to know that the separation system believed that the weaker second sound source is present, even though the system can not determine what it is. In order for this masked state to serve a useful function, the recognition system must have some way of dealing with a sound when it is in the masked state as we have reason to believe the auditory system does.

5.2 Additional Information Sources for Sound Separation

5.2.1 Binaural Information Processing

This thesis has focused on how monaural information is used to separate two simultaneous sounds, even though we know that the auditory system uses both monaural and binaural information to separate sounds. Binaural information such as the timing and intensity differences between the cochlear output of the two ears are used by the auditory system to help it focus on the sound coming from a particular direction [see Lyon 1983 and Lindemann 1983 for computer models of this process]. In addition to these binaural information cues, the auditory system can also combine the results of monaural sound separation processing (performed on the cochlear output of each ear) to help it separate the incoming sounds.

The experimental results of Cutting [1976] indicate that the results of each ear's monaural processing are combined at several different levels (sound localization, fusion of local acoustic features, fusion of linguistic features) to form a new representation at that level. The interaction between the monaural and binaural system only serves to make a difficult problem (understanding how the auditory system separates sounds) even harder. We do not yet know the details of how the monaural and binaural processes interact to separate sounds. Even though we do not know how the auditory system combines monaural and binaural information for sound separation, one can construct a separation system that uses binaural information to help separate sounds.

5.2.2 Higher Level Processing

The phenomenon of 'auditory induction' and 'phonemic restoration' [Warren 1971, 1972, 1974] demonstrates that the auditory system uses its knowledge about sounds to help it separate them from noise. The contextual information contained in a model of the sound that we are listening to can also help a computer separate one sound from another.

If the background noise is a repetitive sound such as a typewriter or the

ringing of a telephone, a computer model of the incoming sound can be used by a separation system to help assign incoming neural events with the appropriate model. Even though it is not clear how these sound models can help in the detection of different sound sources (the lowering of the SNR of this sound's detection threshold), they can be used by the spectral estimation algorithms to compute an estimate of each of the two simultaneous sound sources.

The addition of higher level processing can also help in the assignment of group objects to sound streams. In the /three/ - /seven/ example discussed in section 2.3.1.6, after each of the group objects has a phonetic label, linguistic information (about the likelihood of different phoneme sequences) can be used to assign the /s/ of the digit /seven/ to the female sound stream. The addition of these higher level processes can help to correct the errors when the group objects were correctly identified but assigned to the wrong sound stream.

5.2.3 Interface with a Recognition System

The output of the current separation system was evaluated using a recognition system designed for a single speaker. It is possible that the recognition performance would have increased if the system had been evaluated using a recognition system specifically designed to deal with more than one simultaneous sound.

The simplest modification to a recognition system would be the use of a different distance metric, one which was developed for dealing with two simultaneous sounds [Bridle et. al. 1984]. This spectral matching technique uses not only the reference spectrum and the spectral estimate of the desired sound, but also the spectral estimate of the noise spectrum.

Another possible modification would be to use information about the accuracy of the spectral estimate of each sound source in the spectral distance metric. The output of the separation system is not only a spectral estimate of each sound source, but also includes an estimate of the accuracy of the spectral estimate. The spectral accuracy information is easily determined from the width of the different amplitude ratio probability distributions [see figure 3.16]. By knowing both the spectral estimate and the accuracy of the spectral estimate for each sound source,

one can more accurately compute the probability that this spectrum denotes a specific phonetic category.

More complicated recognition systems which could interpret the 'masked' state described in section 5.1.4 are also possible. When the weaker sound is too weak to be detected accurately, instead of deciding that this sound source was silent, the separation system would decide that the sound source was 'masked' by the other sound. The recognizer would interpret the masked state differently from a silent state [see figures 2.2 and 2.3 for an illustration of the difference between the silence and masked state]. The addition of a masked state in a recognition system would help to reduce the number of words that are missed by the current recognition system (caused when the weaker sound source was not detected).

5.3 Future Psychoacoustic Experiments

There are many details of human auditory processing which remain unknown. Our understanding of the computations performed by the auditory system is limited primarily to the peripheral auditory system. Details of the computations performed at levels beyond the cochlea are sparse. Most of the information that is known (psychoacoustic experiments) about the auditory system deals with the perception of different sounds by a human listener.

The field of pitch perception illustrates our lack of a detailed understanding of how the human auditory system computes periodicity information. The psychophysical literature contains hundreds of different experiments on the perception of pitch by the auditory system. There are currently four main theories on how the auditory system computes pitch. These are the theories of Goldstein, Wightman, Terhardt, and Licklider. The first three of these theories (Goldstein, Wightman, Terhardt) are mathematical models, which attempt to compute the same pitch value that is computed by the auditory system. Since the auditory system could have computed that pitch value in a different way from the above theories, they do not necessarily reflect the computations that are performed by the auditory system. The theory which comes the closest to being a model of the computations performed in auditory pitch perception is Licklider's theory. How-

ever, Licklider's theory describes only part of the computations that might be performed by the auditory system (the rest are left for a neural network). Over thirty years of experimental research on pitch perception has still not determined the precise operations that are actually performed by the auditory system during the computation of periodicity information.

It is extremely difficult to determine how the auditory system computes what it does. It will probably be many years before the detailed computations performed by the auditory system will be well understood. In the meantime, we can refine the theory of auditory sound separation, hypothesize different computational mechanisms, compute the results obtained from these algorithms and compare them to the results obtained from the auditory system. The refinement of a computational model and subsequent comparison with the results from auditory psychoacoustic experiments can help to determine what algorithms the auditory system might use to interpret the information that it hears.

5.4 Summary

The field of sound separation is a new and exciting area of auditory research. It represents an opportunity for researchers to apply their knowledge about the auditory system to develop computational models of auditory sound separation. It is also an opportunity to test our models of auditory processing against the auditory system in order to modify and improve them.

The development of computer models of auditory processing requires a great deal of computational resources. The computation time required to run these algorithms is currently the limiting factor in the time taken to develop and test new separation algorithms. Since many of the algorithms are simple and repetitive (the same operation takes place in each frequency channel every 10 msec), new SIMD computer architectures [Lyon 1984] represent a promising approach to providing the computation power needed at a reasonable cost. The availability of sufficient computer power will allow researchers to develop and test computer simulations of auditory processing.

The addition of top-down information from a higher level processing system,

and the interface with a recognition system designed to handle multiple simultaneous sounds, are important steps which are needed to form a complete model of auditory sound separation.

It is the author's view that an analogy can be made between the future of sound separation and the field of speech recognition. Initially, there is a great deal of room for increasing the performance of sound separation systems. After a certain period of time, all the simple improvements will have been made, and to attain performance levels which are close to the auditory system will require long years of hard effort and scientific study.

Appendix 1

In section 3.3, two advantages of using coincidence formula number four were discussed. The first advantage is that the coincidence of two different events must always be less than the coincidence of one of those events with itself. The second advantage is that this formula enhances the modulation depth of the resulting coincidence representation. This appendix will prove these two results.

Result # 1. Show that:

$$Coincidence(event_a, event_a) \geq Coincidence(event_a, event_b) \quad \forall event_a, event_b \tag{1.1}$$

The coincidence function is defined as follows:

$$Coincidence(event_a, event_b) = avg(area_a, area_b) \times \left(\frac{min(area_a, area_b)}{ma\text{-}(area_a, area_b)}\right)^2 \tag{1.2}$$

When $event_a$ is equal to $event_b$,

$$Coincidence(event_a, event_a) = area_a \tag{1.3}$$

Substituting the right hand side of equations 1.2 and 1.3 into equation 1.1, we must show that:

$$area_a \geq avg(area_a, area_b) \times \left(\frac{min(area_a, area_b)}{max(area_a, area_b)}\right)^2 \quad \forall area_a, area_b \tag{1.4}$$

This can be rewritten as:

$$a \geq \frac{a+b}{2} \times \left(\frac{min(a,b)}{max(a,b)}\right)^2 \quad \forall a, b \tag{1.5}$$

For a fixed value of a, b must be greater than a, equal to a, or less than a. Each of these three cases is examined below.

1. $b > a$

   Equation 1.5 can be rewritten as:

$$a \geq (\frac{a}{2} + \frac{b}{2}) \left(\frac{a^2}{b^2}\right) \quad \forall a, b \tag{1.6}$$

   Which can be rewritten as:

$$a \geq a \frac{\left(\frac{a^2}{b^2} + \frac{a}{b}\right)}{2} \quad \forall a, b \tag{1.7}$$

   Since $b > a$, both fractions on the right hand side are less than 1.0, and their sum is less than 2.0 Therefore, the inequality holds in this case.

2. $b = a$

   This case reduces to $a \geq a$, which is true.

3. $b < a$

   Equation 1.5 can be rewritten as:

$$a \geq \frac{a+b}{2} \left(\frac{b^2}{a^2}\right) \quad \forall a, b \tag{1.8}$$

   Which can be rewritten as:

$$a \geq a \left(\frac{b^2}{a^2}\right) - \frac{a-b}{2} \left(\frac{b^2}{a^2}\right) \quad \forall a, b \tag{1.9}$$

   Since $b < a$, the first term on the right hand side is less than a. The second term on the right hand side is a positive quantity, and therefore further reduces the right hand side. Therefore, the inequality holds in this case.

   Since equation 1.1 holds for all three cases, it is true for all a and b.

Result # 2. Show that coincidence formula version four enhances the coincidence representation for an amplitude modulated signal, while the other three versions do not.

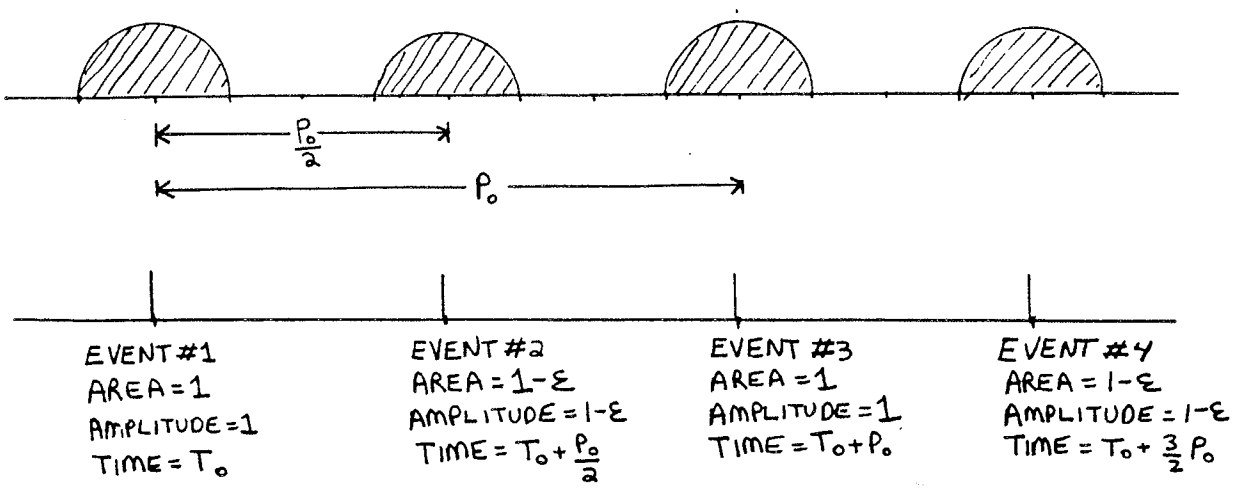

Figure 1.1: Top: Output of a single frequency channel of the cochlear model over time. Bottom: The event representation and the parameters of each event. Note the slight amplitude modulation at the pitch period Po.

For simplicity, let us consider the output of a single cochlear filter whose frequency location is close to the second harmonic of a periodic signal. The output of this frequency channel of the cochlear model is slightly amplitude modulated, since this channel does not completely filter out the other harmonic components. An example of this channel's output is shown in figure one.

The value of the coincidence representation at a delay equal to the pitch period (column two in table one) is equal to $C(event_1, event_3) + C(event_2, event_4)$. The value of the coincidence representation at a delay equal to half the pitch period (column three in table one) is equal to $C(event_1, event_2) + C(event_2, event_3)$. The amplitude modulation of the original cochlear output is equal to $\frac{1-\epsilon}{1} = 1 - \epsilon$. The modulation depth of the original cochelar output is equal to $1 - \frac{1-\epsilon}{1} = \epsilon$. The amplitude modulation of the coincidence representation is equal to the value at half the pitch period (column three) divided by the value at the pitch period (column two). The modulation depth of the coincidence representation is equal to one minus the fraction of column three divided by column two.

Version	Value of Coincidence Representation at delay = Pitch Period	Value of Coincidence Representation at delay = half Pitch Period	Modulation Depth of Coincidence Representation
1	$1+(1-\epsilon)^2$	$(1-\epsilon)+(1-\epsilon)$	$\approx \frac{1}{2}\epsilon^2$
2	$1+(1-\epsilon)$	$\sqrt{1-\epsilon}+\sqrt{1-\epsilon}$	$\approx \frac{1}{8}\epsilon^2$
3	$1+(1-\epsilon)$	$2\left(\frac{1+(1-\epsilon)}{2}\right)\left(\frac{1-\epsilon}{1}\right)$	ϵ
4	$1+(1-\epsilon)$	$2\left(\frac{1+(1-\epsilon)}{2}\right)\left(\frac{1-\epsilon^2}{1}\right)$	$2\epsilon-\epsilon^2$

Table 1.1: How the Coincidence Representation of figure one varies as a function of which formula is used to compute the coincidence of two events.

Since ϵ is small, the output of the cochlear model is slightly amplitude modulated. When the coincidence representation is computed on this amplitude modulated signal, can we determine what the pitch period is from the coincidence representation? If we used the first or second version of the coincidence formula, the value of the coincidence function at half the pitch period would be virtually identical to the value at the pitch period (since ϵ^2 is very small). Therefore, we would have great difficulty distinguishing between P_0 and $\frac{1}{2}P_0$ as the correct pitch period. The amplitude modulation present in the cochlear model's output is preserved in the coincidence representation using formula three, and is enhanced when formula four is used. This enhanced modulation makes the determination of P_0 as the correct pitch period easier.

References

Allik, J., Mihkla. M., Ross, J. (1984) "Comment on 'Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception' ", Acoustical Society of America, 75(6), 1855-1857.

Boll, S.F. (1979) "A spectral subtraction algorithm for suppression of acoustic noise in speech", IEEE ICASSP 200-203.

Berouti, M., Schwartz, R., Makhoul, J. (1979) "Enhancement of speech corrupted by acoustic noise", IEEE ICASSP 208-211.

Bregman, A.S., and Dannenbring, G.L. (1973) "The effect of continuity on auditory stream segregation", Perception and Psychophysics, 13(2), 308-312.

Bregman, A.S., and Rudnicky, A.I. (1975) "Auditory segregation: stream or streams?", Journal of Experimental Psychology: Human Perception and Performance, 1(3), 263-267.

Bregman, A.S. (1978a) "Auditory streaming is cumulative", Journal of Experimental Psychology: Human Perception and Performance, 4(3), 380-387.

Bregman, A.S., and Pinker, S. (1978b) "Auditory streaming and the perception of timbre", Canadian Journal of Psychology, 32(1), 19-31.

Bregman, A.S. (1978c) "Auditory streaming: competition among alternative organizations", Perception and Psychophysics, 23(5), 391-398.

Bregman, A.S. (1978d) "The formation of auditory streams", in Attention and Performance VII, edited by J. Requin. Lawrence Erlbaum Associates.

Bregman, A.S. (1981) "Asking the 'What For' Question in Auditory Perception" in Perceptual Organization, edited by M. Kubovy and J.R. Pomerantz. Lawrence Erlbaum Associates.

Bridle, J.S., Ponting, K.M., Brown, M.D., and Borrett, A.W. (1984) "A noise compensating spectrum distance measure applied to automatic speech recognition", Proc. Inst. Acoust. Autum Meeting, Windermere.

Broadbent, D.E. and Ladefoged, P. (1957) "On the fusion of sounds reaching different sense organs", Journal of the Acoustical Society of America, 29(6), 708-710.

Broadbent, D.E. (1967) "Word-frequency effect and response bias", Psychological Review, 74(1), 1-15.

Broadbent, D.E. (1977) "The hidden preattentive process", American Psychologist, 109-118.

Brokx, J.P.L., Nooteboom, S.G., and Cohen, A. (1979) "Pitch differences and the intelligibility of speech masked by speech", IPO Annual Progress Report 14

Brokx, J.P.L., and Nooteboom, S.G. (1982) "Intonation and the perceptual separation of simultaneous voices", Journal of Phonetics, 10, 23-36.

Cherry, E.C. (1953) "Some experiments on the recognition of speech, with one and two ears", Journal of the Acoustical Society of America, 25(5), 975-979.

Chistovich, L.A., Fyodorova, N.A., Lissenko, D.M., and Zhukova, M.G. (1975) "Auditory segmentation of acoustic flow and its possible role in speech processing", in Auditory Analysis and Perception of Speech, edited by Fant, G. and Tatham, M.A.A.

Chow, K.L. (1951) "Numerical estimates of the auditory central nervous system of the rhesus monkey", Journal Comparative Neurol., 95, 159-175.

Cooper, W.E. (1979) "Speech perception and production", Ablex Publishing Corporation.

Cutting, J.E. (1976) "Auditory and linguistic processes in speech perception: inferences from six fusions in dichotic listening", Psychological Review, 83(2), 114-140.

Dannenbring, G.L. (1976) "Perceived auditory continuity with alternately rising and falling frequency transitions", Canadian Journal of Psychology 30(2), 99-114.

Dannenbring, G.L., and Bregman, A.S. (1978) "Streaming vs. fusion of sinusoidal components of complex tones", Perception and Psychophysics, 24(4), 369-376.

Darwin, C.J. and Bethell-Fox, C.E. (1977) "Pitch continuity and speech source attribution", Journal of Experimental Psychology: Human Perception and Performance, 3, 665-672.

Darwin, C.J. (1978) "The perception of speech", in Handbook of Perception IV, edited by E.C. Carterette and M.P. Friedman. Academic Press.

Darwin, C.J. (1981) "Perceptual grouping of speech components differing in fundamental frequency and onset-time", Quarterly Journal of Experimental Psychology, 33A, 185-207.

Darwin, C.J. (1984a) "Grouping frequency components of vowels: when in a harmonic not a harmonic?", Quarterly Journal of Experimental Psychology 36A, 193-208.

Darwin, C.J. (1984b) "Perceiveing vowels in the presence of another sound: constraints on formant perception", Acoustical Society of America, 76(6) 1636-1647.

DeBoer, E. (1975) "On the residue and auditory pitch perception" in Handbook of Sensory Physiology V, edited by W. Keidel.

Deutsch, J.A. and Deutsch, D. (1963) "Attention: Some theoretical considerations", Psychological Review, 70, 80-90.

Duifhuis, H., Willems, L.F., Sluyter, R.J. (1982) "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception", Acoustical Society of America, 71(6), 1568-1580.

Evans, E.F. (1981) "The dynamic range problem: place and time coding at the level of cochlear nerve and nucleus", in Neuronal Mechanisms of Hearing, edited by J. Syka and L. Aitkin. Plenum Press.

Fay, R.R. (1972) "Perception of amplitude-modulated auditory signals by the goldfish", Journal of the Acoustical Society of America, 52(2) 660-666.

Furukawa, T., Matsuura, S. (1981) "Adaptation and dynamic response occuring at hair cell-afferent fiber synapse", in Neuronal Mechanisms of Hearing, edited by J. Syka and L. Aitkin. Plenum Press.

Goldstein, J.L. (1966) "An investigation of monaural phase perception", University of Rodchester Ph.D. Thesis

Goldstein, J.L. (1973) "An optimum processor theory for the central formation of the pitch of complex tones", Journal of the Acoustical Society of America, 54, 1496-1516.

Goldstein, J.L., Srulovicz, P. (1977) "Auditory-nerve spike intervals as an adequate basis for aural frequency measurement", in Psychophysics and Physiology of Hearing, edited by E.F. Evans and J.P. Wilson. Academic Press.

Grenier, Y., Bry, K., LeRoux, J., Sulpis, M. (1981) "Autoregressive models for noisy speech signals", IEEE ICASSP, 1093-1096.

Hall, J.W., Haggard, M.P., and Fernandes, M.A. (1984) "Detection in noise by spectro-temporal pattern analysis", Acoustical Society of America 76(1) 50-56.

Hanson, B.A., Wong, D.Y., Juang, B.H. (1983) "Speech enhancement with harmonic synthesis", IEEE ICASSP 1122-1125.

Hawkins, H.L., and Presson, J.C. (1977) "Masking and preperceptual selectivity in auditory recognition", in Attention and Performance VI, edited by S. Dornic. Lawrence Erlbaum Associates.

Houstma, A.J.M, Wicke, R.W., Ordubadi, A. (1980) "Pitch of amplitude-modulated low-pass noise and predictions by temporal and spectral theories", Journal of the Acoustical Society of America, 67(4), 1312-1322.

Howes, D. (1957) "On the relation between the intelligibility and frequency of occurrence of english words", Journal of the Acoustical Society of America, 29(2), 296-305.

Javel, E. (1980) "Coding of AM tones in the chinchilla auditory nerve: implications for the pitch of complex tones", Acoustical Society of America, 68(1), 133-146.

Johnson, D.H. (1980) "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones", Acoustical Society of America, 68(4), 1115-1122.

Kiang, N.Y.S., Moxon, E.C. (1974) "Tails of tuning curves of auditory-nerve fibers", Acoustical Socitey of America, 55(3), 620-630.

Kopec, G.E., and Bush, M.A. (1985) "Network-based isolated digit recognition using vector quantization", to be published in IEEE Transactions on Acoustics, Speech, and Signal Processing, August 1985.

Langner, G. (1981) "Neuronal mechanisms for pitch analysis in the time domain", Experimental Brain Research, 44, 450-454.

Lesser, V.R., Fennell, R.D., Erman, L.D., and Reddy, D.R. (1975) "Organization of the Hearsay II Speech Understanding System", IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23, 11-24.

Licklider, J.C.R. (1951) "A duplex theory of pitch perception", Experientia, 7, 128-133.

Licklider, J.C.R. (1959) "Three Auditory Theories", in Psychology: A Study of Science, edited by S. Koch McGraw Hill.

Lim, J.S., Oppenheim, A.V., Braida, L.D., (1978a) "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition", IEEE Transactions Acoustics, Speech, Signal Processing 26(4) 354-358.

Lim, J.S. (1978b) "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise", IEEE Transactions Acoustics, Speech, Signal Processing 26(5) 471-472.

Lim, J.S., Oppenheim, A.V. (1979) "Enhancement and bandwidth compression of noisy speech", Proceedings of the IEEE, 67(12) 1586-1604.

Lim, J.S. (1983) "Speech Enhancement", Prentice Hall.

Lindemann, W. (1983) "Extension to a binaural cross-correlation model by means of a lateral inhibition mechanism", Paper QQ5 presented at the 106th meeting of the Acoustical Society of America at San Diego, CA.

Lyon, R.F. (1982) "A computational model of filtering, detection, and compression in the cochlea", IEEE ICASSP, 1282-1285.

Lyon, R.F. (1983) "A computational model of binaural localization and separation", IEEE ICASSP, 1148-1151.

Lyon, R.F. (1984) "MSSP: A bit-serial multiprocessor for signal processing", VLSI Signal Processing, (Workshop Proceedings.), IEEE Press.

Massaro, D.W., and Cohen, M.M. (1983) "Evaluation and integration of visual and auditory information in speech perception", Journal of Experimental Psychology: Human Perception and Performance, 9(5), 753-771.

Mathes, R.C., and Miller, R.L. (1947) "Phase effects in monaural perception", Journal of the Acoustical Society of America, 19(5), 780-797.

McAdams, S. (1984) "The auditory image: a metaphor for musical and psychological research on auditory organization", to be published in Cognitive Processes in the Perception of Art.

McGurk, H., and MacDonald, J. "Hearing lips and seeing voices", Nature, 264, 746-748.

Miller, G.A., Heise, G.A., and Lichten, W. (1951) "The intelligibility of speech as a function of the context of the test materials", Journal of Experimental Psychology, 41, 329-335.

Moller, A.R. (1979) "Coding of complex sounds in the auditory nervous system", in Hearing and Speech Mechanisms, edited by O. Creutzfeldt, H. Scheich, Chr. Schreiner. Springer-Verlag.

Moller, A.R. (1981) "Coding of complex sounds in the auditory nervous system", in Neuronal Mechanisms of Hearing, edited by J. Syka and L. Aitkin. Plenum Press.

Moore, B.C.J. (1977) "Effects of relative phase on the components on the pitch of three-component complex tones", in Psychophysics and Physiology of Hearing, edited by E.F. Evans and J.P. Wilson. Academic Press.

Moray, N. (1970) "Attention: Selective processes in vision and hearing", Academic Press.

Nawab, H., Oppenheim, A.V., Lim, J.S. (1981) "Improved spectral subtraction for signal restoration", IEEE ICASSP 1105-1108.

Neisser, U. (1967) "Cognitive psychology", Appleton-Century-Crofts.

Parsons, T.W. (1976) "Separation of speech from interfering speech by means of harmonic selection", Journal of the Acoustical Society of America, 60, 911-918.

Patterson, R.D., Johnson-Davies, D. (1977) "Detection of a change in the pitch of AM noise", in Psychophysics and Physiology of Hearing, edited by E.F. Evans and J.P. Wilson. Academic Press.

Pearsons, K.S., Bennett, R.L., and Fidell, S. (1976) "Speech levels in various environments", BBN Report No. 3281.

Peterson, T.L., Boll, S.F. (1981) "Acoustic noise suppression in the context of a perceptual model", IEEE ICASSP 1086-1088.

Rasch, R.A. (1978) "The perception of simultaneous notes such as in polyphonic music", Acustica, 40(1), 21-33.

Rose, G.J., and Capranica, R.R. (1984) "Processing amplitude-modulated sounds by the auditory midbrain of two species of toads: matched temporal filters", Journal Comparative Physiology A, 154, 211-219.

Rose, J.E., Hind, J.E., Anderson, D.J., and Brugge, J.F. (1971) "Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey", Journal of Neurophysiology, 34, 685-699.

Rubenstein, H., and Pollack, I. (1963) "Word predictability and intelligibility", Journal of Verbal Learning and Behavior, 2, 147-158.

Sachs, M.B., and Abbas, P.J. (1974) "Rate versus level function for auditory-nerve fibers in cats: tone-burst stimuli", Acoustical Scoiety of America, 56(6), 1835-1847.

Sachs, M.B., and Young, E.D. (1980) "Effects of nonlinearities on speech encoding in the auditory nerve", Acoustical Society of America, 68(3), 858-875.

Sanders, D.A. (1977) "Auditory Perception of Speech", Prentice Hall

Scheffers, M.T.M. (1979) "The role of pitch in the perceptual separation of simultaneous vowels", IPO Annual Progress Report, 14, 51-54. Eindhoven, Netherlands.

Scheffers, M.T.M. (1982) "The role of pitch in the perceptual separation of simultaneous vowels II", IPO Annual Progress Report, 17, 41-45. Eindhoven, Netherlands.

Scheffers, M.T.M. (1983) "Simulation of auditory analysis of pitch: An elaboration of the DWS pitch meter", Acoustical Society of America, 74(6), 1716-1725.

Schneider, W., and Shiffrin, R.M. (1977) "Controlled and automatic human information processing: I. detection, search, and attention", Psychological Review, 84(1), 1-54.

Schroeder, M.R. (1968) "Period histogram and product spectrum: new methods for fundamental frequency measurement", Acoustical Society of America, 43(4), 829-834.

Shiffrin, R.M., and Schneider, W. (1977) "Controlled and automatic human information processing: II. perceptual learning, automatic attending, and a general theory", Psychological Review, 84(2), 127-190.

Steiger, H., and Bregman, A.S. (1981) "Capturing frequency components of glided tones: frequency separation, orientation, and alignment", Perception and Psychophysics, 30(5), 425-435.

Treisman, A.M. (1960) "Contextual cues in selective listening", Quarterly Jounal of Experimental Psychology, 12, 242-248.

Treisman, A.M. (1964) "The effect of irrelevant material on the efficiency of selective listening", Journal of Psychology, 77(4), 533-546.

Treisman, A.M. (1964) "Monitoring and storage of irrelevant messages in selective attention", Journal of Verbal Learning and Behavior, 3, 449-459.

Ullman, S. (1979) "The interpretation of visual motion", MIT Press.

Underwood, G. (1974) "Moray vs. the rest: the effects of extended shadowing practice", Quarterly Journal of Experimenal Psychology, 26, 368-372.

VanNorden, L.P.A.S. (1971) "Rhythmic fission as a function of tone rate", IPO Annual Progress Report, 6.

VanNorden, L.P.A.S. (1974) "Temporal coherence in random tone sequences", IPO Annual Progress Report, 9.

VanNorden, L.P.A.S. (1975) "Temporal coherence and the perception of temporal position in tone sequences", IPO Annual Progress Report, 10.

Voight, H.F., Sachs, M.B., Young, E.D. (1981) "Effects of masking noise on the representation of vowel spectra in the auditory nerve", in Neuronal Mechanisms of Hearing, edited by J. Syka and L. Aitkin. Plenum Press.

Von Wright, J.M., Anderson, K., and Stenman, U. (1975) "Generalization of conditioned GSR's in dichotic listening", in Attention and Performance V, edited by P.M.A. Rabbitt and S. Dornic. Academic Press.

Warren, R.M., and Obusek, C.J. (1971) "Speech perception and phonemic restorations", Perception and Psychophysics, 9(3B), 358-362.

Warren, R.M., Obusek, C.J., and Ackroff, J.M. (1972) "Auditory induction: perceptual synthesis of absent sounds", Science, 176, 1149-1151.

Warren, R.M., and Sherman, G.L. (1974) "Phonemic restorations based on subsequent context", Perception and Psychophysics, 16(1), 150-156.

Warren, W.H.Jr., Vebrugge, R.R. "Auditory perception of breaking and bouncing events: a case study in ecological acoustics", Journal of Experimental Psychology: Human Perception and Performance, 10(5), 704-712.

Weintraub, D.J. and Walker, E.L. (1966) "Perception", Wadsworth Pub.

Whitfield, I.C. (1978) "The neural code", in Handbook of Perception IV, edited by E.C. Carterette and M.P. Friedman. Academic Press.

Wightman, F.L. (1973) "The pattern-transformation model of pitch", Journal of the Acoustical Society of America, 54, 407-416.

Willems, L.F. (1983) "DWS pitch detection algorithm extended to the time domain", IPO Annual Progress Report, 18, 14-19. Eindhoven, Netherlands.

Witkin, A.P. and Tenenbaum, J.M. (1983) "What is perceptual organization for?", Proceedings of the Eight International Joint Conference on Artificial Intelligence, 1023-1026.

Young, E.D. and Sachs, M.B. (1979) "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers", Acoustical Society of America, 66(5), 1381-1403.

Zwicker, E. (1961) "Subdivision of the audible frequency range into critical bands", Acoustical Society of America, 33(2), 248.