PITCH AND SPECTRAL ANALYSIS OF SPEECH BASED ON AN AUDITORY SYNCHRONY MODEL

by

Stephanie Seneff

B.S., Massachusetts Institute of Technology
(1968)
M.S., E.E Massachusetts Institute of Technology
(1980)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January, 1985

Signature of Author . . . . . .
Department of Electrical Engineering and Computer Science
January 16, 1985

Certified by . . . . . .
Kenneth N. Stevens
Thesis Supervisor

Accepted by . . . . . .
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

PITCH AND SPECTRAL ANALYSIS OF SPEECH BASED ON AN AUDITORY SYNCHRONY MODEL

by

STEPHANIE SENEFF

Submitted to the Department of Electrical Engineering
and Computer Science on January 22, 1985 in partial fulfillment
of the requirements for the Degree of Doctor of Philosophy

ABSTRACT

There has been a substantial interest in the last few decades in the problem of training computers to recognize human speech. In spite of the concentrated efforts of conscientious teams of researchers, however, the solution remains elusive, unless the task is kept so restricted as to be uninteresting. These discouraging results may be due in part to the fact that researchers in the past paid little attention to models for human processing of auditory signals to guide in the design of speech front-end processing strategies. The picture is rapidly changing at the present time, although we have not yet realized any direct benefits from the available models.

Voiced speech sounds are characterized in the spectral domain by prominent peaks at specific frequencies that correspond to certain resonances in the vocal tract. The frequencies of these "formants" convey most of the information necessary to identify the phonetic content. The peripheral level of the auditory system performs a frequency analysis, but also compresses the dynamic range of input stimuli. The net effect is to reduce the prominence of spectral peaks, relative to those obtained through standard Fourier analysis. Recent research on the response of a large population of auditory nerve fibers in the cat's ear to speech-like stimuli [Sachs and Young, 1979, 1980] has demonstrated that mean rate response alone does not in general convey adequate information to show clearly the frequencies of the formants. However, a significant amount of information is retained in the patterns of firing which is lost when a simple count of number of spikes per unit time is derived. Sachs and Young, and others, have suggested that a form of processing that measures the synchrony in the response to certain periodicites may be able to accentuate peaks in the spectrum. This thesis concerns the development of a specific strategy for such synchrony detection, and its application to the two separate tasks of spectral analysis and estimation of the fundamental frequency of voicing.

The approach of the thesis is to process the incoming speech signal through a system which models what is known about peripheral auditory processing, and then to apply a synchrony measure to accentuate the spectral attributes that are known to be important for the identification of the phonetic content of speech. The design of the synchrony measure is motivated in large part by a preconceived notion of what represents a "good" result. The main criteria were that peaks in the original speech spectrum should be preserved, but amplitude information, particularly general

spectral tilt factors or overall gain, should be deemphasized. A spectral representation that is smooth in frequency and time, without sacrificing resolution, is considered desirable.

The peripheral model includes a bank of critical-band filters, followed by a dynamic range compression scheme and a nonlinear half-wave rectifier. The speech is never reduced to a spike sequence; rather, a continuous waveform describing the probability of firing of a nerve fiber is obtained. The model derives outputs for a set of 30 filters, covering the frequency range from 230 to 2700 Hz, thus representing the region of importance for sonorant segments of speech. The outputs of the peripheral model are then further processed through a hypothesized "central" processor, which consists of a spectral analyzer and a parallel pitch estimator. These two independent systems both make use of a "Generalized Synchrony Detector" [GSD], which detects specific prominent periodicities present in the input waveform.

The GSD algorithm, the major novel idea of the thesis, is a ratio of the envelope amplitude of a sum of two inputs to the envelope amplitude of a difference of the same two inputs, where the second input is a delayed version of the first. Due to the ratio, the algorithm achieves a normalization with respect to signal level, an important aspect that reduces the effect of overall signal level and eliminates fluctuations in amplitude over time due to random placements of the window relative to glottal pulses. For spectral analysis, the output of each peripheral channel is processed through a GSD with the delay equal to the "center period" of the peripheral filter; thus the response is strong when the center frequency component is dominant in the original waveform. A plot of the outputs of the GSD's as a function of center frequency yields a "pseudo spectrum", which exhibits generally sharp peaks at the formant frequencies.

The fundamental frequency is extracted from a waveform which is obtained by summing the outputs of the peripheral model across the "place" dimension. This waveform exhibits strong periodicity at the fundamental, but the periodicities at formant frequencies have in general been reduced, relative to the original waveform, mainly as a consequence of the dynamic range compression of the peripheral model. The GSD algorithm is applied to the pitch waveform for the range of periods appropriate for the fundamental frequency. The pitch estimate is obtained from the first prominent peak in the resulting "pseudo autocorrelation", a plot of the GSD outputs as a function of the delay period. The method can extract the pitch of an isolated tone, as well as the fundamental frequency of a sequence of higher harmonics.

The performance of the model was evaluated by processing a number of synthetic and natural speech tokens through the system. In some cases, results were compared with available perceptual data. A comparison was also made between the GSD method for detecting synchrony and several alternative methods, as applied to spectral analysis. It remains to be demonstrated whether in fact the representation for the speech spectrum obtained through this model results in an improved performance in a speech recognition system, or whether a pitch detector based on the pseudo autocorrelation is of superior quality to existing pitch detectors.

THESIS SUPERVISOR: Kenneth N. Stevens
TITLE: Professor of Electrical Engineering

Acknowledgements

The ideas for this thesis were inspired mainly by a series of articles by Murray Sachs and Eric Young on the responses of a large population of auditory nerve fibers to speech-like stimuli. I greatly admire their pioneering work in this area, and my debt to them is boundless.

Sincere thanks go to my faculty advisor, Professor Ken Stevens, for the time he spent guiding me in the research and in the writing of the document. I am also indepted to my four thesis committee members, Professors Jon Allen, Steve Colburn, Tom Knight, and Camp Searle, for carefully reading early versions of the manuscript, and suggesting improvements and/or further directions for research. Special thanks are due Professor Searle, who spent considerable time brainstorming with me on diverse topics, all related to auditory processing.

I thank Professor Victor Zue for providing the Symbolics LISP machine facilities, with assorted software appropriate for computer speech processing, which is a marvelous environment for doing research of this sort. In this regard, I also thank Dave Shipman, Scott Cyphers, and others responsible for the software development of the SPIRE package, which allowed ease of programming and rapid turn-around time of ideas.

Finally, and especially, I thank my family, Victor, Lily, Michael, Gregory, Timmy, and Cory, for graciously putting up with the hectic life-style we have led over the last six years. I give special thanks to my oldest son, Michael, for figuring out how to draw "MacTemplates" for my figures on the Macintosh, and for solving individual crises in that vein as they came up. My mother-in-law, Lily Zue, has unerringly managed the home front during the day, giving her love to the kids, and providing us with home-cooked meals. Most of all, I thank my husband, Victor Zue, for his endless patience and persistent support through the ups and downs of graduate life, without which none of this would have been possible.

Contents

Chapter 1

Introduction

The acoustic theory of speech production has been a very useful theoretical framework in which speech analysis systems, such as Linear Prediction [Makhoul, 1975; Markel, 1976], have been developed. This theory has also been successfully applied to the design of speech-synthesis algorithms and devices. By analogy, it ought to be possible to make use of our knowledge of the human auditory system to motivate the design of the front end of a speech recognition system. Unfortunately, the human auditory system in general, and the perception of speech in particular, is not nearly as well understood as is the production system.

This lack of understanding is not due to a lack of interest; on the contrary the literature on the physiology and psychoacoustics of the auditory system is vast. The problem is that the system is very complex; several nonlinearities exist even at the peripheral level, and much of what is observed from psychoacoustical studies is probably taking place at very high levels (such as in the auditory cortex). Physiological measures have for the most part been done on stimuli that bear little resemblance to the complex speech signal. It is difficult to extrapolate from the results of such experiments to a conclusion about how to process speech on a computer. It is probably for these reasons that speech researchers have been deterred from designing a speech analysis system that is motivated by the human auditory response mechanism.

However, in recent years our knowledge about the ear has become more specific, and researchers have begun to actually measure physiological response to speech-like stimuli [Sachs and Young, 1979, 1980; Delgutte, 1980; Miller and Sachs, 1981]. From detailed spectral analysis of firing rates of nerve fibers at different places along the basilar membrane to vowel-like stimuli, researchers have been able to make conjectures about further processing that might take place at a more central level of the auditory system in order to turn firing patterns into something from which such information as formant frequencies could be recovered. The Sachs and Young data in particular reopened the debate about synchronized response (i.e., detecting quasi-periodic patterns in the temporal characteristics of auditory nerve firings) versus a simple rate response mechanism. Their data argue convincingly for some measure of synchrony taking place at some later stage in the system, in order to preserve formant peaks at high signal levels. Their data represent a significant departure from earlier arguments about rate versus synchrony, because the speech signal is much more complex than tone bursts, yet relatively well understood in terms of the salient features such as formant frequencies.

To date most of the engineering efforts to obtain a model for the speech spectrum have not made use of human auditory perception, except in the crude sense that a frequency analysis is performed. Channel vocoders obtain a spectral model by extracting the envelopes of the outputs of a bank of bandpass filters. Linear prediction models are based on the acoustic theory of speech production,

where the poles of the model correspond to the resonances of the vocal tract. DFT-based spectral analysis is similar to channel vocoder methods, since each Fourier coefficient can be viewed as the output of a bandpass filter. Cepstral analysis is basically an extension of the DFT method into a nonlinear domain. The log spectrum is passed through a smoothing filter to remove the individual harmonics of the fundamental frequency. Motivation for smoothing the log spectrum comes from mathematical arguments about converting a convolution between the excitation and the vocal tract frequency-shaping filter to a sum.

Two notable exceptions to this generality are the speech front end processors designed by Searle et al. [1979, 1980] and Lyon [1982]. These researchers have developed "critical band" linear filter banks with the major feature that the filters increase in bandwidth, and hence in temporal resolution, with increasing frequency. In both systems the filtering process is followed by an envelope detector and a nonlinear compression scheme. Searle et al. choose a simple log for the compression, whereas Lyon includes a complex automatic-gain-control mechanism with multiple feedback loops. Neither of these models, however, includes any processing beyond the compression to make use of details in the waveshape of each filter output to improve frequency resolution.

Another crucial aspect of speech processing is the pitch extraction problem. The perception of pitch is a topic which has fascinated psychoacousticians for many decades. Many experiments have been performed on an assorted collection of synthesized sounds from which definitive statements can be made at least about how the ear does **not** perceive pitch. The temporal versus spatial debate crops up here, and there exists strong supportive evidence in both camps. Emerging pitch perception theories include models by Goldstein [1973], Wightman [1973], and Terhardt et al. [1982].

Pitch detection as an engineering problem is also a well-explored subject, and there are at least a dozen published pitch extraction algorithms in the literature [see, for example, the review article by Rabiner et al., 1976]. These also tend to fall into one of two categories, processing of the waveform (temporal) versus processing in the spectral domain (i.e., place). Both approaches have their strong points and vulnerabilities. For example, with waveform methods phase plays a critical role, whereas with spectral methods the fixed size of the time window for spectral analysis is a limitation.

This thesis concerns a computer implementation of a speech analysis system that is motivated by available knowledge about the human auditory system. The initial-stage processing of the system borrows extensively from what is known about the peripheral auditory system, with the goal of obtaining outputs that resemble histogram data on nerve fiber responses in the cochlea. The final spectral estimates are obtained after a second-stage processing that involves a proposed synchrony measure which utilizes the known phase-locking property of the nerve fibers to enhance spectral prominences due to vocal tract resonances.

For the development of this second-stage processing, it is not possible to make use of any known physiological evidence for processing that may take place beyond the 8th cranial nerve, except in a very general sense. Instead, the design criteria were defined in terms of what is judged to be a "good" representation for the spectrum of speech sounds. It was assumed that peaks in the log

spectrum of the signal should be retained in the representation, at least if they are not too close together, and that no "spurious" peaks should be introduced at frequencies where no peak exists in the log spectrum. However, it was not felt necessary to preserve the amplitudes of the peaks; in fact, it may be preferable to normalize out certain features such as a general spectral tilt factor or the overall signal level. It is desirable that the spectral representation be smoothly varying from frame to frame in time. Vertical striations such as occur in a wide-band spectrogram are not considered acceptable, because, although the eye is capable of tracking the underlying formant movement, the computer can not deal easily with such rapidly fluctuating spectral ranges.

Ideally, each formant in the representation should show up as a distinct peak, such that it would be possible to enumerate the formants without error. In practice, such a criterion may not be realistic. For example, the first formant of nasalized vowels is often represented by a pole-zero-pole complex, that results in a peak-splitting. The location of the underlying formant is then in the valley between the two peaks. Furthermore, there are certain sounds, such as /ɜ̃/, in which the second and third formants are very close in frequency. In such a case, it may be more appropriate to look for consistency in the representation, rather than a reliable separation of the two formants into distinct peaks on the spectrum. If the representation sometimes but not always resolves the two peaks, for different instances of the same phoneme, then it would be more difficult to devise a mechanism for identifying the phoneme than if the representation consistently merged the two formants into a single peak.

It is important also that the spectral representation be relatively robust against background noise that may be present with the signal. Any mechanism that falls apart in the presence of noise is not suitable, even if in the absence of noise it produces excellent results. Similarly, the central processing strategy should be relatively insensitive to variations in the nonlinearities that are introduced in the peripheral model. If, for example, the final spectral output changes drastically when a piece-wise linear half-wave rectifier in the peripheral model is replaced with a hyperbolic tangent half-wave rectifier, then the system is probably too sensitive to the detailed wave-shape of the input.

After examining a number of different alternatives for the second-stage processing, we arrived at a mechanism for synchrony detection that, when incorporated into a model for spectral analysis, seemed to obey most of the requirements enumerated above. The resulting "Generalized Synchrony Detector" [GSD] responds strongly only to signals which are periodic with the specified control period. In the spectral processing system, each peripheral level channel output is processed through a GSD which is tuned to the center frequency of the corresponding peripheral filter. Input stimuli that have a prominent spectral component at the center frequency of a particular peripheral channel yield a corresponding strong response at the output of the subsequent GSD. The set of GSD outputs, plotted along the place dimension, form a "pseudo spectrum," which generally has much sharper peaks at the formant frequencies than would be available from the mean rate response measured at the peripheral level output.

The GSD algorithm consists of the ratio of the estimated amplitude of a sum waveform to the estimated amplitude of a difference waveform. The inputs to the sum and difference computation are

the GSD input signal and a delayed version of the input signal, with the delay period corresponding to the frequency to which the GSD is tuned. When the input to the GSD is perfectly periodic with the delay period, the amplitude of the difference waveform is zero. Hence, the ratio can become infinitely large during perfect synchrony. To constrain the response to be within reasonable limits, a final soft-limiter is applied to the output of the ratio. In addition, a threshold is subtracted from the numerator prior to the divide, to prevent a response to signals that are too weak. The value of this threshold is very close to the level of spontaneous response in the peripheral model.

The implemented system consists of a set of 30 channels spanning the frequency region from about 230 Hz to 2700 Hz, which is adequate for most vowel-like sounds. The initial stage processing includes a bank of critical band filters, follwed by a dynamic range compression scheme and a half-wave rectifier. The 30 channel outputs are then each processed through a GSD tuned to the corresponding peripheral filter center frequency. The outputs of the GSD's are downsampled to a 10ms update rate, to produce a two-dimensional array of amplitudes at each time-frequency coordinate. Most of the data that will illustrate system performance consist of either pseudo spectal cross sections at specific points in time, or pseudo spectrograms, produced as a plot of time vs frequency, with intensity indicated by gray levels.

After having developed a strategy for obtaining a spectral representation for speech derived from the peripheral level channel outputs through a specific synchrony measure, we examined to what extent the presence of the peripheral model was necessary in order for the subsequent synchrony model to still function adequately. From an engineering standpoint it is important to consider which aspects of the system are absolutely essential to performance, and which are included mainly because of our knowledge about peripheral processing. It remains unclear, for example, whether features such as dynamic range compression and half-wave rectification are essential for the overall performance of the complete auditory system, or merely represent limitations of neural mechanisms.

The GSD algorithm not only has applications in the area of spectral representation, but also can be used, with slight modifications of parameter values, for determining the fundamental frequency of voicing. For pitch processing, we decided to combine all of the outputs of the peripheral level analysis across the place dimension to construct a single waveform from which to extract appropriate periodicities. This waveform is then fanned out to a series of GSD's covering the range of human pitch, and the results can be plotted as a function of the delay period to form a "pseudo autocorrelation", which exhibits peaks at multiples of the fundamental period. It is then the task of a pitch detector to estimate the fundamental frequency by determining the first prominent peak in this pseudo autocorrelation.

The first several chapters of this thesis are concerned with a review of relevant psychophysical and physiological data on the human auditory system. First, a brief summary is given of the peripheral auditory system as it is presently understood. There follows a chapter describing a variety of different models that have been proposed for various aspects of the auditory system. The next two chapters concern pitch, the first on the perception of pitch and the second on previous computer pitch detection algorithms. These are included to set the stage for the pitch estimator developed in the thesis. The final review chapter describes the nature of the speech signal, and the

limitations of currently available methods for representing the speech spectrum.

The second half of the thesis concerns the system that has been implemented on the LISP machine/FPS facility at MIT for pitch and spectral analysis. Chapter 7 gives an overview of and motivation for some of the key concepts that were developed and used in the thesis. In Chapter 8 the complete reference system is described in detail, such that the interested reader could reimplement it. Chapter 9 gives several illustrations of the outputs of the spectral model, for simple synthetic speech stimuli, and natural speech tokens. Chapter 10 attempts to relate the pseudo spectrum to psychophysics, by comparing the performance of pseudo spectral processing of a number of synthetic speech tokens with existing perceptual results on the same or related data. Chapter 11 examines which aspects of the spectral model, both peripheral and central, are most important, and what are the various effects of simplifying or even eliminating certain steps in the processing. It also examines several other possible methods for detecting synchrony, and compares them to the reference system. Chapter 12 describes the pitch detector in detail, including the voiced/ unvoiced decision, again so that the interested reader could implement it. The results of processing the pitch detector on a set of 19 short words spoken by preschoolers are given in detail, in order to illustrate system performance on a particularly difficult data set. The last chapter discusses further possibilities for additional research in the area of auditory modelling for speech processing. In particular, it addresses the issue of how to extract from the pseudo spectra the relevant information for phonetic identification, and emphasizes the need for a demonstration of applications in the area of computer speech recognition.

Chapter 2

The Human Auditory System: A Brief Review

In this chapter a general review is presented of the current understanding of some aspects of the human auditory system. The review is necessarily limited in scope, and biased towards those aspects of the system that are most relevant to the thesis. The main purpose of the chapter is to provide motivation for the detailed design structure of the implemented computer model.

2.1 Basic Structure and Function of the Peripheral Auditory System

The human peripheral auditory system consists of the outer, middle, and inner ears, as schematized in Figure 2.1. Sound travels past the pinna and down the auditory canal of the outer ear and impinges on the eardrum, or tympanic membrane, causing it to vibrate. The middle ear, containing the three small bones, the malleus, incus, and stapes, acts as an impedance-matching device to ensure efficient transfer of sound from the air to the cochlear fluid in the inner ear.

The smallest and last bone of the middle ear, the stapes, makes contact with the oval window at the base of the cochlea in the inner ear. The cochlea is a small coiled tube of approximately 2 1/2 turns, with a length of 3 1/2 cm. It has bony rigid walls and is filled with incompressible fluids. It is divided along its length by two flexible membranes, Reissner's membrane and the basilar membrane. The basilar membrane grows in thickness from the base of the cochlea to the apex at the inner end of the coil. The membrane is about 100 times as stiff at the base as it is at the apex. Different portions of the membrane respond preferentially to different frequencies, with response to high frequencies being dominant at the stiff or basal end, and response to low frequencies exhibiting amplitude maxima at the apex. Thus the basilar membrane performs a frequency analysis of the incoming signal, where the frequency scale is mapped into the place dimension of the membrane.

Attached to the basilar membrane is the organ of Corti, a complex organ in which are contained the inner and outer hair cells, separated by an arch known as the tunnel of Corti. The outer hair cells outnumber the inner ones by a factor of seven; however, the role of the outer cells remains obscure, although their role in feedback is somewhat understood. There are about 3500 inner hair cells, each with about 40 hairs. Action potentials are generated on a given auditory nerve fiber as a consequence of a bending motion of the hairs at the top of the hair cells, presumably in response to a shearing motion created between the basilar membrane and the tectorial membrane, lying above the hairs. The nerve fibers originating at the hair cells feed into the 8th cranial auditory nerve. The 8th nerve fibers terminate in the cochlear nucleus, the first of a series of synapses in the brain stem that constitute the central auditory system.

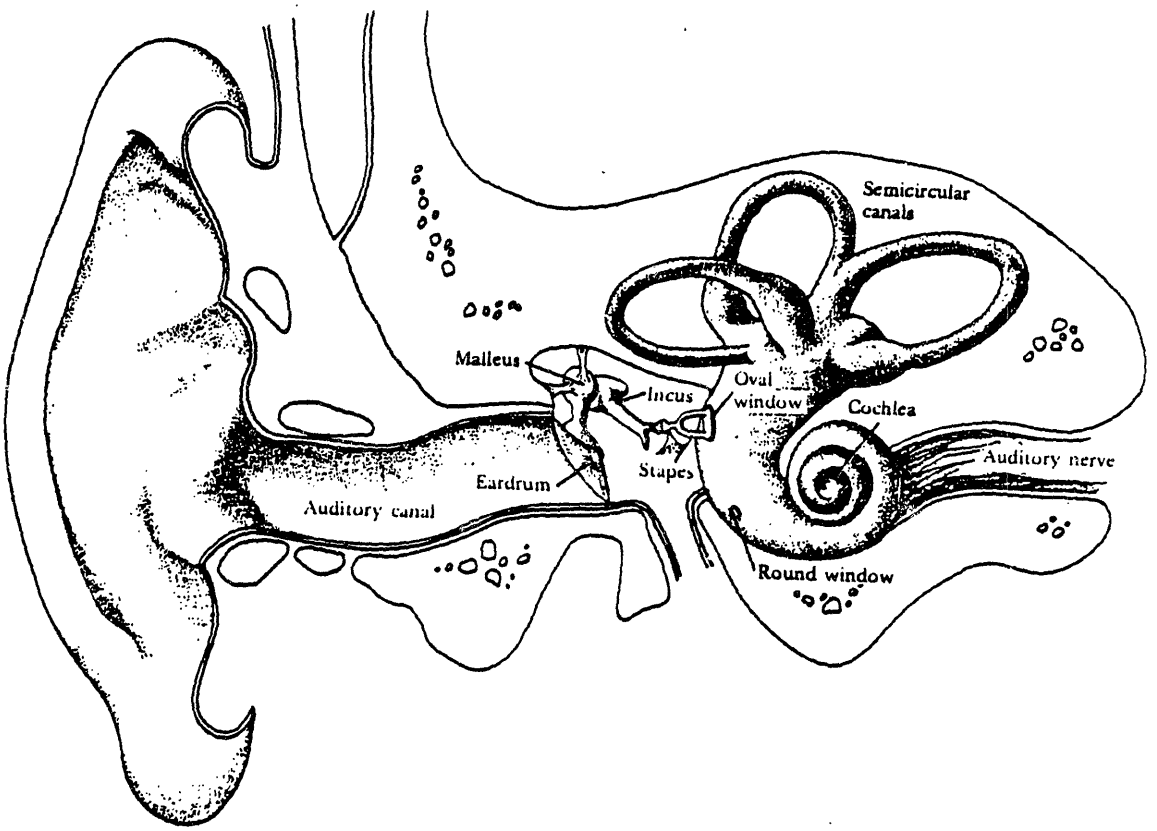

Figure 2.1: Illustration of the structure of the peripheral auditory system, showing the outer, middle, and inner ear. [from Moore, 1982, p.14].

The tuning characteristics of the nerve fibers can be determined by a variety of different experimental techniques. The experimental results are that each fiber is tuned to a center frequency which bears a monotonic relationship with the distance along the basilar membrane from the apex of the cochlea. The tonotopic organization of the nerve fibers such that place has meaning in terms of frequency appears to be preserved at much higher levels of the auditory system, and thus this place information is considered to be important.

The final result when a sound has been transmitted from the external environment to an auditory nerve fiber is a spike train of action potentials with a non-homogeneous Poisson-like distribution [Kiang et. al., 1965, p. 100], which captures the relevant time-frequency-amplitude characteristics of the signal in a probabilistic fashion. A mean firing rate, expressed in spikes per second, is related in a complex way to the amplitude of the incoming signal. Even in the absence of a stimulus, most nerve fibers fire randomly with a spontaneous rate, that is a characteristic of the fiber. As far as can be deduced, adjacent nerve fibers appear to fire independently of one another [Johnson, 1970], and hence their responses could theoretically be combined to obtain an improved statistical sampling of the output response characteristics.

At stimulus onset, the response rate is significantly higher than it is after the signal has been present for a long time. This decrease in response rate is referred to as adaptation [Smith and Zwislocki, 1975]. There is a very rapid initial decay in rate immediately after onset, followed by a slower decay to a steady state level. The average steady state rate response to stimuli at CF resembles an exponential function of signal amplitude for low stimulus levels, but tends to reach a maximal saturation response level for high stimulus levels.

In addition to the mean rate response characteristic, nerve fibers can also be characterized by the detailed temporal patterns in the response as they relate to the response of the basilar membrane. A response is considered to be "synchronized" to the hair cell input stimulus if there can be detected in the probabilistic response pattern some similarity to the wave shape of the stimulus. It has been observed that, for low frequency periodic stimuli, nerve fibers tend to fire in a phase-locked fashion; that is, the intervals between firings tend to cluster near multiples of the stimulus period. The phase-locked component tends to decrease as the frequency of the stimulus is increased, and is almost nonexistent for high stimulus frequencies [e.g., above 5 kHz, although the extent of loss of synchrony is surely species-specific].

A powerful method for determining a detailed probabilistic response wave shape is to compute a "period histogram" of the response to a periodic signal such as a sine wave. The method involves placing time markers at intervals equal to the period of the stimulus, and dividing each interval into N bins. Whenever a spike occurs within a given bin, a counter is incremented for that bin. In this fashion, after several periods have been analyzed, the N bin counters become N sequential samples of the probabilistic response wave shape.

The detailed temporal patterns in the neural response are a potential source of more specific information about the frequencies present in the input signal. Fourier analysis of a period histogram, if the input is a sine wave and the filter center frequency is sufficiently low, will exhibit peaks at multiples of the frequency of the sine wave. The higher harmonics are introduced as a consequence

of rectification and nonlinear response characteristics, issues that are addressed in more detail below. High frequency fibers tend to retain synchrony to the **envelope** of a stimulus, thus preserving information about the fundamental period of a stimulus consisting of several harmonics near the center frequency. Thus a component at the fundamental frequency is present in the period histogram, even when components at the frequencies of the individual harmonics may be absent due to a loss of phase-locking capabilities at these high frequencies. Such envelope synchrony should be useful for pitch perception [Delgutte, 1980; Delgutte and Kiang, 1984].

There is an ongoing debate about whether, and, if so, how, such information may be utilized at later stages of processing. While it is generally assumed that a comparison of the detailed signal arriving from each ear is essential to such tasks as localization [Colburn, 1973], it is not at all clear how much of the preserved detail is necessary for tasks such as pitch perception or formant perception. Unfortunately, the higher auditory system is not well understood at the present time, although auditory physiologists are beginning to characterize responses at several of the brain stem nuclei [Moushegian et al., 1972; Smith et. al., 1978].

In the next section, a more detailed account of the response characteristics of the peripheral auditory system will be discussed, including a summary of the relevant research efforts in the fields of auditory neurophysiology and psychophysics.

2.2 Review of Studies Designed to Determine Peripheral Response Characteristics

In this section a review of the response characteristics of the auditory system at the peripheral level, i.e., up to the point of auditory nerve fiber responses along the basilar membrane in the cochlea, will be given. An implicit division of the response characteristics into a sequence consisting of a linear filter followed by a nonlinear time-dependent amplitude compression, including half-wave rectification, is assumed to simplify the discussion.

The first two subsections will focus on a characterization of the assumed-linear filter characteristics, which have been derived from three main, somewhat independent, methodologies. The different methods yield similar but not identical filter response shapes, but the differences between results are diminishing as experimental techniques improve.

The most direct approach is to measure the vibrations of the basilar membrane, for example, using the Mossbauer technique [Johnstone and Boyle, 1967], or by laser interferometry [Khanna and Leonard, 1982]. Alternatively, the responses of nerve fibers originating at distinct places along the membrane can be recorded, and from an analysis of these responses a "neurophysiological tuning curve" can be derived. A final important source of information on the filter characteristics, one which deals directly with human subjects, is masking experiments in the field of psychological acoustics, although filter characteristics can not be measured in as careful detail as through the other methods.

Psychophysical tuning curves developed from masking experiments and physiological tuning curves measured from nerve fibers or from the basilar membrane thus yield three distinct forms of

a "critical band filter bank" which must be combined to obtain a specification for a computer simulation. The usual course of action is to assume a linear model for the basilar membrane frequency analysis, although this is known to be incorrect, particularly at high amplitudes. Nonlinearities are then assumed to be introduced predominantly in the response characteristic of the nerve fiber (in conjunction with the hair cell). Phenomena such as half-wave rectification, saturation, and adaptation, which will be discussed later in this chapter, enter into the nonlinear aspects of a model, and account for measured physiological response patterns and observations on masking in psychoacoustics.

2.2.1 Psychophysical Tuning Curves

The term "critical band" was first coined by Fletcher [1940] in order to quantify the results of a psychophysical masking experiment involving the detection of a pure tone in bandpass noise. The tone was introduced at a low signal level at the center frequency of the noise band, and the bandwidth of the noise was varied. For each bandwidth, the level of the noise needed to just mask the signal was determined. The observation was that beyond a certain "critical bandwidth" an increase in bandwidth made little difference on the spectrum level, i.e., the level per Hz of bandwidth, necessary to mask the signal. The critical bandwidth could be obtained as a function of frequency by performing the same experiment for center frequencies spanning the frequency scale.

Since Fletcher's experiment, more refined techniques have been developed to obtain more detailed characterizations of the filter shapes. One approach is to fix the frequency of the signal and vary the center frequency of the noise band. Sometimes a pure sinusoid is used for the masker, as well as the signal, although beating phenomena may complicate the results. If the level of the signal is kept low, then only fibers with center frequencies near the frequency of the signal will respond to it. The assumption is then made that the level of the masker necessary to mask the signal is related directly to the response level to the masker frequency of the fibers tuned to the frequency of the signal. If this assumption is valid, then a complete "psychophysical tuning curve" (PTC) can be obtained. Such an approach was used for example by Vogten [1974] who obtained filter shapes remarkably similar to neurophysiological tuning curves [compare Figures 2.2 and 2.3].

One concern with the above approach is that the listener is likely to pay attention to the filter which yields the best signal-to-noise ratio rather than a filter that is centered on the signal. Such a strategy is known as off-frequency listening, and results in a sharper tip of the PTC than probably exists in any single auditory filter [Moore, 1982, p. 90]. A procedure to minimize this effect is described by Patterson [1976], which is to distribute the noise power over the entire frequency band **except** a narrow band centered at the frequency of the test tone signal. As before, the level of the noise necessary to mask the signal is measured. When the width of the notch is increased by an amount δf, the corresponding decrease in the threshold of detection of the tone defines the amount of noise passed by the auditory filter in the narrow bands at either edge of the notch. In this way a filter shape can be obtained, although any asymmetries between the high and low bands will not

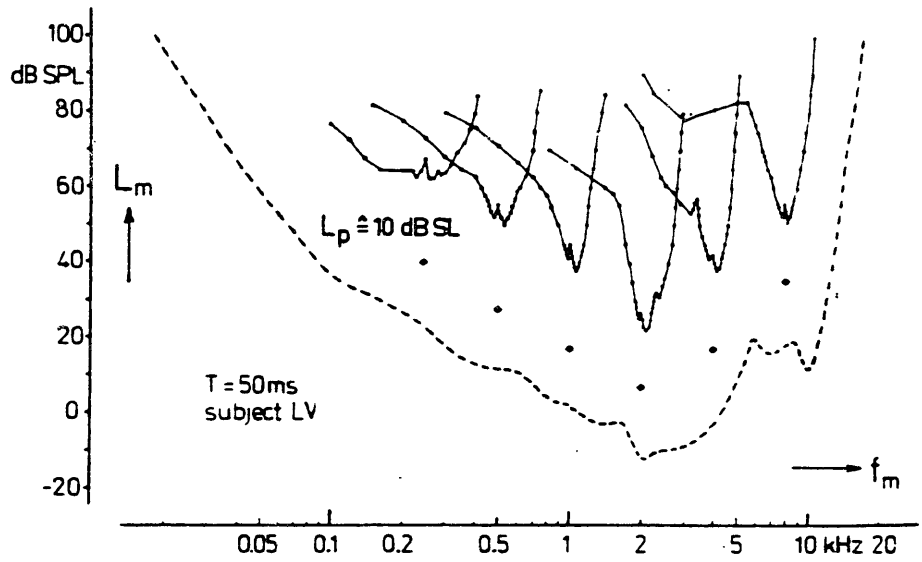

Figure 2.2: Psychophysical tuning curves (PTC's) determined in simultaneous masking, using sinusoidal signals at 10 dB SL. For each curve the solid diamond below it indicates the frequency and level of the signal. The masker was a sinusoid which had a fixed starting phase relationship to the brief, 50 ms signal. The masker level required for threshold is plotted as a function of masker frequency (logarithmic scale). The dashed line shows the absolute threshold for the signal. [from Vogten, 1974].

be detected.

From filter shapes that are derived from these types of experiments, a critical bandwidth can be computed as a function of frequency by using some arbitrary definition such as the distance between 3 dB points on the filter curve. Zwicker [1961] has invented a critical band frequency scale, with the units in "Barks", in memory of Barkhausen, who created the unit of loudness level [Zwicker, 1961]. A difference of one Bark corresponds to the width of one critical band over the whole frequency scale. This Bark scale provides a reference framework for the design of a critical band filter bank. Matching 3 dB points of each filter to the 1 Bark interval then becomes a well-defined criterion as an essential step in the filter design.

2.2.2 Neurophysiological and Mechanical Tuning Curves

An independent method for obtaining a characterization of the auditory filter shapes is to directly measure the auditory nerve fiber responses to simple stimuli. A method used by Kiang et. al. [1965] is to stimulate a nerve fiber with pure tones which map out the frequency region of interest and measure the level of each tone necessary to cause the fiber to fire at a rate that is noticeably above spontaneous rate. An example of sample tuning curves obtained for cats using this method is reproduced here as Figure 2.3.

The methodology has been refined through the years, and in particular it has become evident that the response characteristics are extremely sensitive to the physiological state of the animal. For example, hypoxia (lack of oxygen) has been shown to have a deleterious effect on the response, by reducing the sharpness of tuning of the fiber. Evans and Wilson [1973] have demonstrated that under ideal conditions the high frequency slope of the filter can be extremely steep, on the order of several hundred dB/oct.

Another ingenious method for obtaining tuning curves is to use the property that a cross-correlation of a filter's response with a white-noise input will yield the impulse response of the filter. Such a method was used originally by deBoer [1967], and described fully by deBoer and Jongh [1978]. The method assumes a model of a linear filter followed by a static nonlinearity (rectification process) and a stochastic pulse generator. If the input signal is Gaussian white noise then the linear filter characteristics can be determined by correlating the pulse train with the input. The resulting filter shapes bear a remarkable resemblance to shapes obtained using more traditional methods. More importantly, the authors were able to demonstrate a stability in the filter characteristics over a wide range of signal amplitudes. The results suggest that a model which assumes a linear filter as an initial stage is valid, except at very high signal levels, where the relative bandwidth becomes broader and the tip shifts toward a lower frequency.

Direct measurements of mechanical vibration amplitudes at appropriate locations along the basilar membrane provide yet a third method for obtaining filter shapes. The method has the advantage that it is directly measuring the response before the signal has been converted to a probabilistic Poisson-like process, further complicated by such nonlinear phenomena as adaptation and half-wave rectification. On the other hand, it is difficult to achieve accuracy in the measurement,

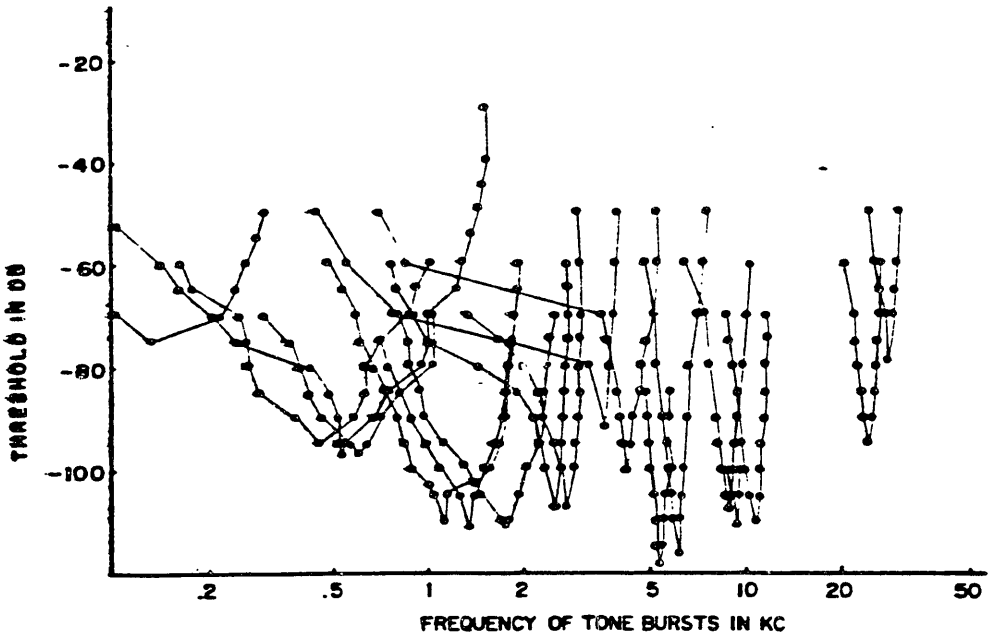

Figure 2.3: Physiological Tuning Curves obtained by combining the data from two cats. [from Kiang et al. 1965].

and thus it is unclear how to interpret the results. There was for a long time a discrepancy between the filter shapes at the level of basilar membrane vibration and the shapes derived from tuning curve data. However, as techniques for measuring basilar membrane motion improve, this gap is slowly diminishing.

Von Bekesy [1960] was the first to attempt to measure motions of the basilar membrane. He was restricted to the apex or third turn of the cochlea (i.e., the low frequency region), and his optical techniques required extremely high signal levels which probably exceeded the linear response region. There was a large discrepancy between the filter fall-off characteristics obtained from his basilar membrane measurements and those obtained from measurements of nerve fiber responses. It was not until many years after his pioneering work that more refined techniques were developed which yielded more sharply tuned response characteristics, although a small discrepancy still remains between basilar membrane response and the nerve fiber responses.

One of these new methods is the Mossbauer technique used by Johnstone and Boyle [1967], Johnstone, Taylor and Boyle [1970], and Rhode [1971]. The method depends upon the Doppler shift effect; a source of gamma rays is placed on the membrane and an absorber is located near it. The movement of the membrane in response to a sound source introduces a modulation in the gamma radiation which can be accurately measured. The technique is capable of detecting a relative velocity as small as 1mm/sec [Green, 1976, p.150]. The function that is usually plotted is the ratio of the displacement of the basilar membrane to the displacement of the stapes, representing the transfer function of the cochlea. Johnstone and Boyle found a linear response characteristic for signal levels up to 95 dB SPL. However, Rhode reported significant nonlinear response characteristics as a function of amplitude near the resonance frequency for levels as low as 75 dB SPL.

Johnstone, Taylor and Boyle measured slopes of 13 dB/oct and 105 dB/oct on the low- and high- frequency sides respectively. This is to be contrasted with slope measurements in the range of hundreds of dB per octave on the high frequency side for neurophysiological tuning curves [Evans and Wilson, 1973]. One possible explanation is that the hair cell may not be responsive to the **amplitude** of vibration of the basilar membrane but rather to the velocity or the acceleration. Evans and Wilson propose a "second" filter of physiological nature that acts before the conversion of mechanical energy to neural spikes. They suggest that this second filter fails to operate under certain stress conditions, such as oxygen deprivation, during which the response of the fiber emulates more closely the response of the basilar membrane. Until there is a better understanding of the exact mechanism of hair cell stimulation, this issue will remain unresolved.

2.2.3 Phase and Overall Gain

In addition to the shape of the tuning curve, two sources of information necessary to fully characterize the response characteristics of the linear filter that represents the first stage of the peripheral auditory system are the phase and the overall gain. Again, relevant information has been obtained from different experimental procedures.

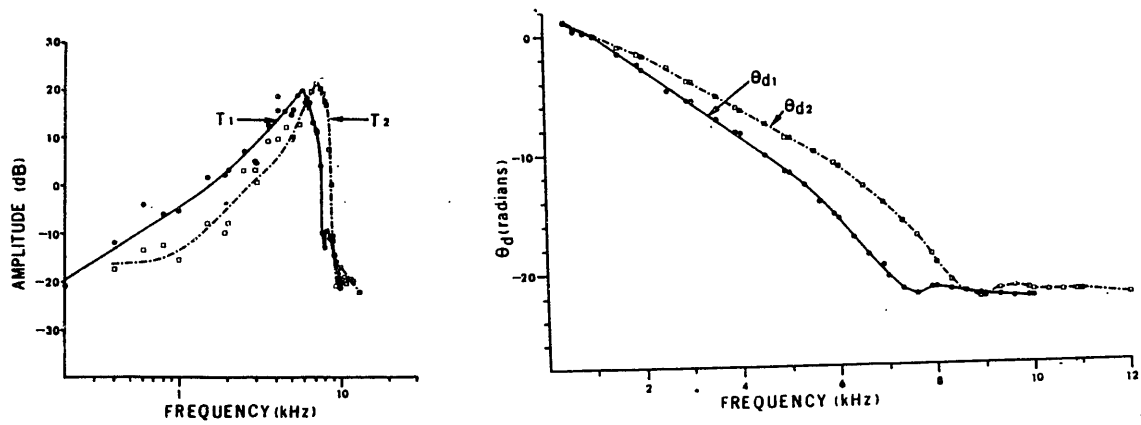

Figure 2.4: Magnitude and phase response characteristics for two positions along the basilar membrane, measured simultaneously using the Mossbauer technique [from Rhode, 1971].

There exist very little data on the phase response of the basilar membrane. Phase was measured by von Bekesy at a few places on the low frequency end, in conjunction with his measurements of amplitude of vibration. W. Rhode has since obtained additional phase data using the Mossbauer technique. Von Bekesy's measurement technique removed any linear phase components, whereas Rhode measured the phase difference between the basilar membrane and the malleus; thus a strong linear phase component was present due to the delays in the filters. Von Bekesy assumed a zero phase characteristic at the resonance frequency, and observed that phase increased to about π as frequencies were decreased below resonance, and decreased to negative values as stimulus frequencies were increased above resonance. Rhode's results are for much higher resonance frequencies, in the range of 8 to 10 kHz. He found that the phase response was linear over a wide frequency region below CF, but that the slope of the phase response steepened as frequency approached CF, as shown in Figure 2.4. Above resonance the phase difference approached a constant value. The slope of the linear phase portion of the phase curve was steeper with increasing distance from the stapes, reflecting the longer delays for the low frequency filters. If the linear phase component were removed, the phase data would resemble the data obtained by von Bekesy. In particular, there is about a π difference in phase between a linear extension of the linear part of the curve and the measured value for phase at the resonance frequency.

Jont Allen [1983] obtained phase data for a wide range of stimulus frequencies and filter center frequencies by measuring nerve fiber responses. To plot his data, he subtracted out a large estimated

linear phase component due to the delays in the system. He found, as a general rule, that phase changed slowly with increasing stimulus frequency for a given nerve fiber, until the frequency approached the characteristic frequency of the fiber, at which point the phase dropped rapidly. For the higher frequency units, he was able to demonstrate, by normalizing phase with the cochlear microphonic, that his data were consistent with Rhode's data.

Overall gain should be related to the level of the tip of the tuning curve, as obtained for a collection of fibers clustered at a particular place on the basilar membrane. In other words, the amplitude of the stimulus necessary to just evoke a response at a particular fiber's characteristic frequency should be inversely proportional to the gain of a simulated filter at that frequency. With reference to the data collected by Kiang (Figure 2.3), it is apparent that the signal amplitude necessary to evoke a response increases with decreasing characteristic frequency for fibers tuned below about 1 kHz. These data are also in line with loudness perception at low intensities; low frequency tones must be higher in level to evoke an equivalent loudness response to higher frequency tones [Fletcher and Munson, 1933]. The amplitude at the tip of the tuning curve also increases as frequency is increased above about 1 kHz. This effect may be offset somewhat, however, by a broad outer ear resonance near 3 kHz [Shaw, 1974].

2.2.4 Responses of Auditory Nerve Fibers to Tones

Neurophysiological tuning curves derived from nerve fiber responses to tone stimuli are determined by detecting the level necessary to just evoke a rate response above the spontaneous rate. More complete characterizations of the response patterns to tone stimuli can yield information about the response characteristics as a function of tone level, the detailed steady-state shape of the probabilistic response curve, and the changes in rate response that take place over time. Experiments that use pure tones as stimuli will not be able to answer more complex questions about masking phenomena, and therefore researchers have also attempted to characterize the responses to two-tone complexes, which show a clear nonlinear inhibitory effect, both when presented simultaneously and when presented in sequence.

A thorough study of steady state response patterns to isolated tones is described by Johnson [1971, 1980]. He measured rate response and computed period histograms of the response of each nerve fiber as a function of stimulus frequency and amplitude. He also computed a measure referred to as the synchronization index [Anderson, 1973], which basically measures the component in the spectrum of the histogram at the frequency of the tone, and ranges from 0 to 1.

Johnson's period histograms reveal several interesting phenomena. At low signal levels, as shown in Figure 2.5., the response approximates a sine wave shape for the positive half of the cycle. However, as amplitude increases, the response shape tends to become peakier, and is skewed such that the peak is early in the cycle, particularly for low frequency tones. Furthermore, if the tone is sufficiently low in frequency and high in intensity, a peak-splitting phenomenon occurs. This phenomenon is conjectured to be the result of a prominent second harmonic introduced by nonlinearities in the cochlea.

Smith and Zwislocki [1975] used tone pedestals as stimuli and measured rate responses of guinea pig auditory nerve fibers as a function of time in order to address the issue of short-term adaptation. The stimuli consisted of sudden-onset tone bursts whose amplitudes, I, were doubled at a time τ = 150 ms after initial onset. A PST histogram of the response was computed, and a difference between the response just before and just after the amplitude doubling constituted a "steady state incremental response":

$$IR = R_\tau^+ - R_\tau^-$$

This incremental response (IR) was then compared with an "onset incremental response", defined as the difference between the response to an onset tone at level 2I and one at level I. They made two important observations from the data: 1) The steady state and onset IR's were equal, and 2) The ratio of the response, R_0 at onset to the response R_τ^- at steady state was approximately equal to 2.5 regardless of the onset intensity level, I. Their conclusion is that adaptation is basically a linear process, and does not result from a multiplicative gain term.

The overall rate response to a tone stimulus tended to fall off with an exponential decay as time progressed, with a time constant of about 40 ms. However, the decay was much more rapid at the moment of stimulus onset. A detailed analysis revealed that an exponential decay with a time constant of only 5 ms characterized this initial rapid drop fairly accurately.

The most logical next step in stimulus complexity is to examine the response characteristics to a two-tone complex, in order to assess the amount of nonlinearity in the system. Such studies have been done, and the results indicate that nonlinearities are prevalent [Sachs and Kiang, 1968]. For example, the overall rate response can **decrease** if a second tone near the edge of the response curve of a fiber is added to a tone at the characteristic frequency, even though the energy in the input stimulus has increased. This phenomenon is referred to as "two-tone suppression". Psychophysical observations on masking of a high-frequency tone by a low-frequency tone are in line with the results [Fletcher, 1953], although it is very difficult to relate the masking results to anything specific in the auditory system.

Rose et al. [1974] examined the detailed waveshape of responses to two-tone stimuli, and derived a formula to describe the response characteristics. In order to estimate the level of the response to each tone independently, they matched histograms of responses with half-wave rectified sums of two sine waves at the tonal frequencies, with appropriate amplitude and phase. They found that a tone T_1 near the edge of a filter's response region could reduce the response to a tone T_2 at the characteristic frequency even when T_1 was at a level as much as 30 dB below threshold. As the amplitude of T_1 was increased, the rate continued to decline until the amplitude reached threshold. Further increases resulted in an increased rate, but the dominant frequency in the response pattern shifted from T_2 to T_1.

A theoretical paper by Hall [1980] on two-tone suppression relates the phenomenon to the second filter issue. He proposes that the hair cells may be sensitive to the first or second spatial derivative of the basilar membrane motion, resulting in a sharpening of the Q of the response. He views two-tone

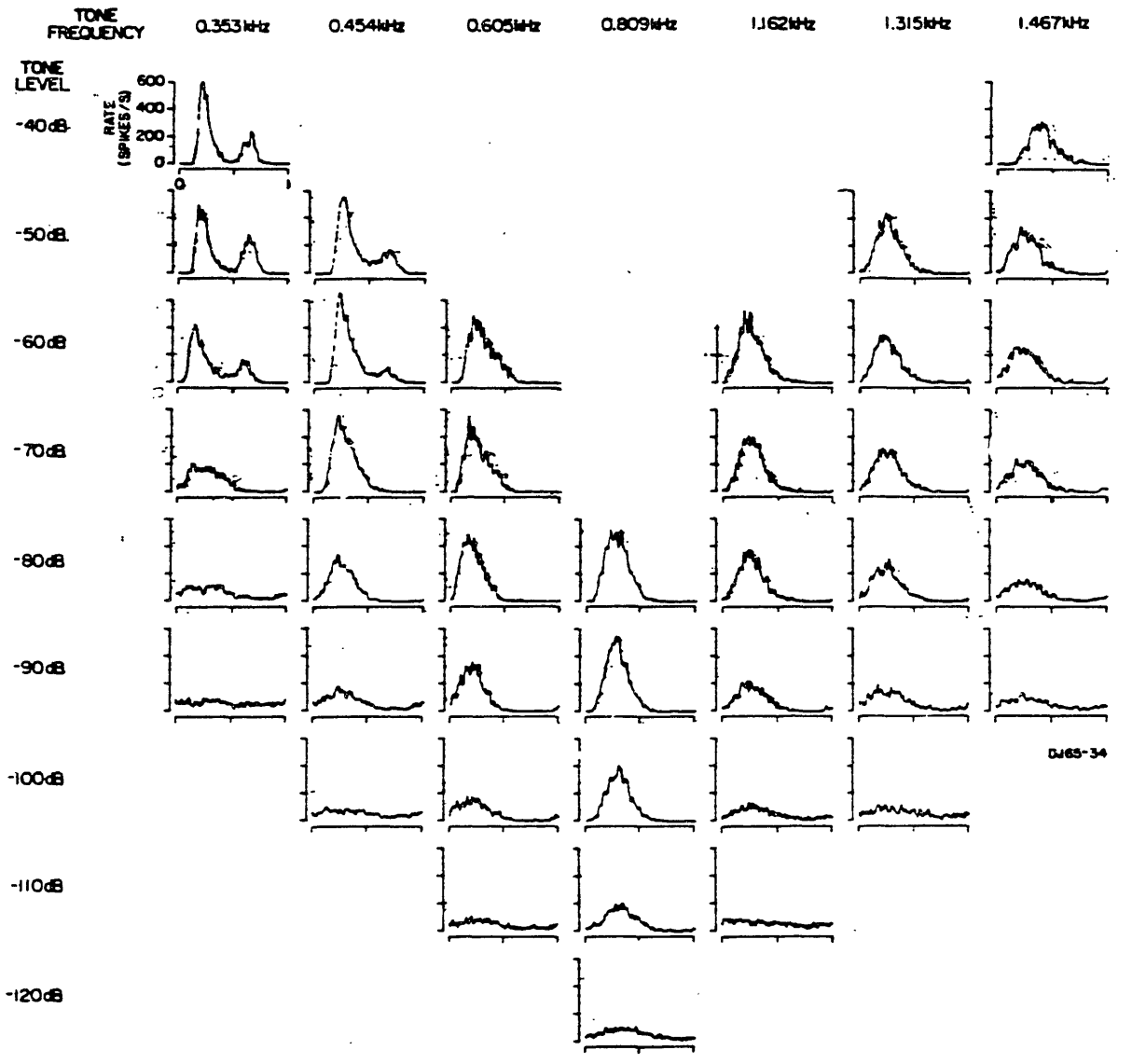

Figure 2.5: Period histograms computed from the discharge pattern of unit DJ65-34 in response to single tones. Each column contains histograms of the response to a tone that decreases in level from top to bottom. Tone levels are specified in decibels relative to the peak voltage level driving the condenser earphone (corresponding to 115 dB SPL in this frequency range). Each histogram represents 15 sec of recording. The vertical scale of each histogram is in spikes/sec, and the horizontal scale covers one period of the stimulus. The reference phases of the histograms in each column are the same, but are arbitrarily chosen to represent the histograms in approximately sinusoidal phase. The spontaneous discharge rate of this unit was 60 spikes/sec and its CF is 0.8 kHz. [from Johnson, 1980].

suppression as a mechanical effect that takes place at the broadly tuned level of basilar membrane displacement. The second filter then filters out the response to the suppressing tone, leaving the overall response noticeably reduced. He suggests an experimental procedure that would map out "iso-excitation contours" and "iso-suppression contours" which should be analogous to threshold tuning curves. If his conjecture is true, then the iso-excitation contours should be considerably steeper than the iso-suppression contours.

Harris and Dallos [1979] studied the response patterns to sequences of tones in order to address the issues of forward masking, as opposed to simultaneous masking. They analyzed responses of chinchilla auditory nerve fibers to tone-burst stimuli consisting of a masking stimulus followed, after a short silence interval, by a probe stimulus. They found that the recovery of probe response as a function of the time interval between masker offset and tone onset followed an exponential time course. By varying the masker frequency, they obtained forward masking tuning curves from iso-forward masking contours near threshold of masking, which approximated quite closely the fiber's neurophysiological tuning curve. Furthermore, they observed a monotonic relationship between the rate response induced by the masker and the reduction in probe response magnitude. These results are in contrast with simultaneous two-tone suppression effects, and thus the two phenomena are quite distinct from each other. The authors propose that the reduction in stimulus response caused by the presence of the preceding tone is a phenomenon directly related to recovery fron short-term adaptation.

2.2.5 Responses to Speech-like Stimuli

Only recently have researchers begun to examine the nerve fiber response characteristics to complex stimuli that more closely resemble natural speech. Successful efforts in this area involve experiments that analyze the responses of a large population of nerve fibers to the same stimulus. Noteworthy are the work by Sachs and Young [1979, 1980] on response of the cat's ear to the steady state synthetic vowel /ɛ/, by Miller and Sachs [1981] on responses to the synthetic CV stimuli /bɑ/ and /dɑ/, and the work by Delgutte [1980] on the response to transitions as well as steady states, and to fricative-like stimuli as well as vowel-like stimuli. These researchers observed response patterns that were consistent with results obtained from previous studies; for example, such effects as two-tone suppression are evident in the responses to the resonance frequencies of the formants.

Sachs and Young were particularly interested in addressing the issue of whether rate alone is sufficient for vowel identification, or whether some form of synchrony measure is needed at a higher stage in the auditory system to recover formant peaks. They studied a large population of fibers, and collected both data on mean rate response and histograms of response patterns to synthetic vowel stimuli presented at several stimulus levels. They found that as a consequence of rate saturation and two-tone suppression phenomena, the formant information was almost completely obliterated from the rate response at the higher amplitudes, as shown in Figure 2.6, for the vowel /ɛ/. In fact, at the highest stimulus levels, the rate response to frequencies intermediate between

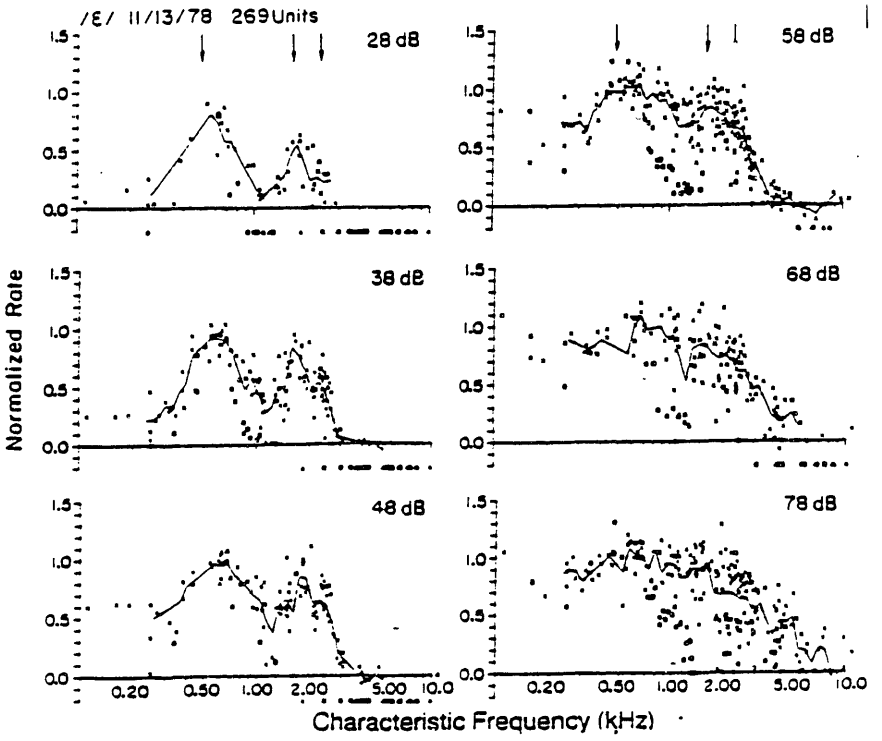

Figure 2.6: Normalized rate vs CF for 260 units studies on 11/13/78 with synthesized /ɛ/ as the stimulus. Positions of the formant frequencies are shown by the arrows. [from Sachs and Young, 1979].

F_1 and F_2 was actually **higher** than the response at F_2. This result could be explained by two-tone suppression, since the response at the intermediate frequency would be dominated by the first formant resonance, which was present at a level near saturation, and the response to the second formant would be reduced below the saturation level by the presence of the first formant at the edge of the tuning curve, below threshold. One would predict that an examination of the frequency distribution of the histogram of the response at the intermediate frequency would reveal an overwhelming component at the first formant, whereas the filter tuned to the second formant would be responding in nearly complete synchrony to the second formant, since the first formant would be below threshold.

The above discussion suggests that a mechanism that examines the synchrony in the patterns of the responses would yield a better separation of the two formant peaks, and a response function that

would be more stable as a function of stimulus amplitude. Sachs and Young developed a model for measuring synchronized rate response which did yield a better resolution of the formants and a more consistent performance with amplitude variations. The measurement, labelled "Average Localized Synchronized Rate" (ALSR), detected the concentration of energy near a particular frequency in the responses of a collection of fibers tuned to that frequency. A discussion of this model will be deferred until Chapter 3, in the context of other similar models and simulations.

The above model was applied by Miller and Sachs [1983] to the detection of formant motion near vowel onset in two synthetic speech stimuli, /b*a*/ and /d*a*/. They were able to demonstrate that ALSR computations from the fiber responses maintained the correct peaks at the formant frequencies during the transitions. Rate response alone performed significantly better during transition and onset conditions than during steady state conditions, although still not as well as the synchronized response measurement.

Delgutte also studied nerve fibers of cats' ears, and his stimuli consisted of tone burst sequences, simple synthetic CV's such as "m*a*" and "b*a*", speech-like noise bursts with slow or rapid onset, corresponding to /š/ or /č/, and steady state single-formant vowels. His results will be discussed here in considerable detail, since they are the best data available as auditory response patterns which could be used as a guide in developing computer models to process actual speech data.

Rate versus level response patterns to tone bursts and noise-bursts exhibited similar characteristics, and the results were consistent with those reported by Smith and Zwislocki. The rate responses to tone bursts, reproduced here as Figure 2.7 show a sharp initial response with a rapid decay down to an "elbow" at about 15-20 ms after onset. Following the elbow, the rate continues to decay with a much slower time constant (around 30-45 ms) to a steady state value. The mean rate before the elbow exhibits a much wider range with signal amplitude than the mean rate after the elbow.

Sudden-onset noise bursts show a similar sharp peak at onset. When a **gradually** rising noise burst is used as stimulus, so as to model a /š/, for example, as contrasted with /č/, a peak in the response still appears shortly after onset, and long before the stimulus has reached the steady-state level. This peak, however, is not nearly as high or as sharp as the peak that results from a rapid-onset /č/-like stimulus.

Two related experiments were designed to test "post adaptation" effects, which are the reduction in rate response to a stimulus when it has been preceded by a long adapting signal. The first experiment consisted of measuring the rate response to a fixed-amplitude short test tone following a long adapting tone of variable amplitude, with a 60 ms silence interval between the two tones. The response to the test tone fell off in a systematic fashion as the level of the adapting tone was increased (Figure 2.8) and was reduced by as much as a factor of 6, when the adapting tone was at its highest level, relative to the rate in the absence of the adapting tone.

The other experiment involved a comparison of the response to a synthetic "m*a*" with the response to a synthetic "b*a*" for five different fibers. Since the /m/ stimulated low frequency fibers much more than the /b/, the response of these fibers at vowel onset in the "m*a*" did not exhibit the sharp peak onset that was characteristic of the vowel in "b*a*", because of the adapting influence

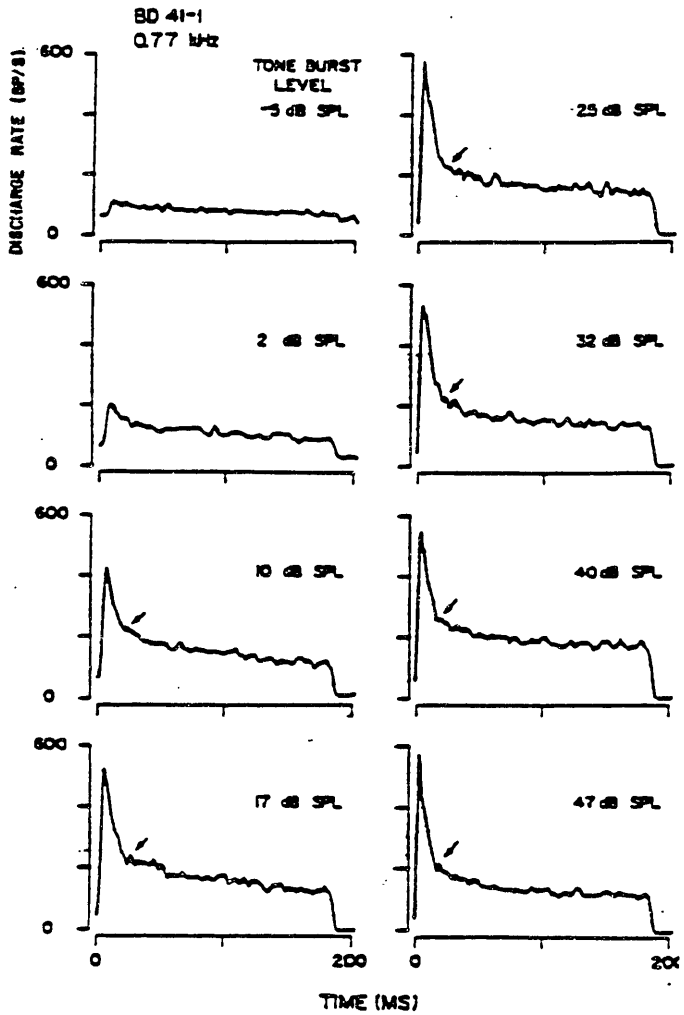

Figure 2.7: Response patterns of an auditory nerve fiber to a tone burst at eight different levels. The 180-ms tone burst has a rise-fall time of 0.25 ms and a repetition rate of 100/min. The tone-burst frequency is approximately the fiber's characteristic frequency(CF). The post-stimulus time (PST) histograms are computed with a bin width of 1.4 ms from 480 stimulus presentations. They are smoothed by convolution with a five-point, unity-gain, triangular window (three-point smoothing). Arrows point to the "elbow" in the time-course of the discharge rate. The fiber had a threshold of -8 dB SPL and a spontaneous discharge rate of 66 spikes/sec. [from Delgutte, 1980].

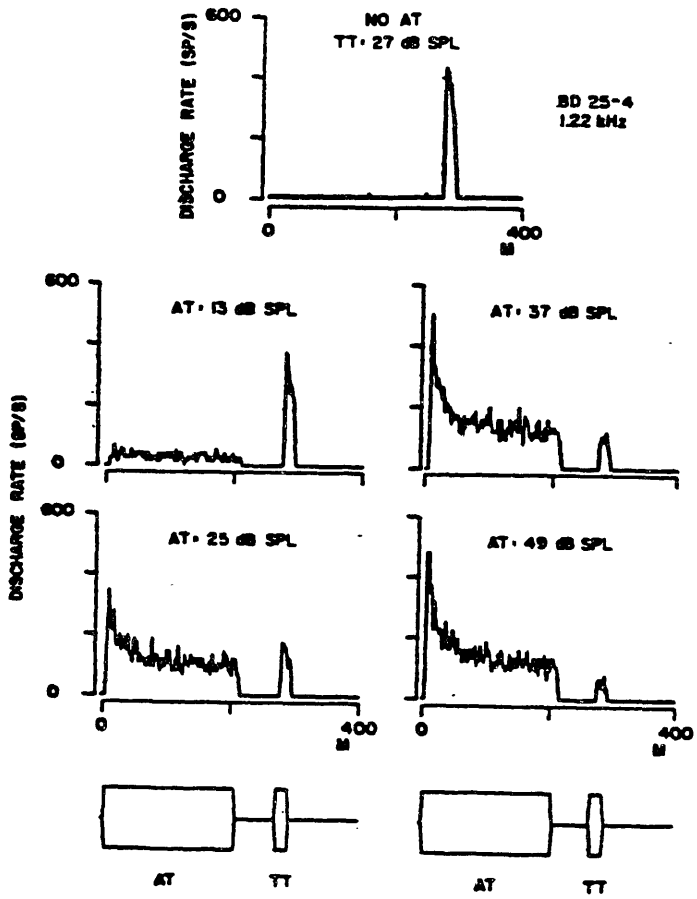

Figure 2.8: Response patterns of an auditory-nerve fiber to a 20-ms test tone (TT) preceded by a 200-ms adapting tone (AT). The envelope of the stimulus is shown in the bottom panel. Both tone bursts have a rise-fall time of 2.5 ms and a frequency approximately equal to the fiber CF. The time delay between the two tone bursts is 60 ms. The repetition rate of the stimulus is 100/min. The histograms are computed from 75 stimulus presentations with a bin width of 1 ms and are three-point smoothed. [from Delgutte, 1980].

of the preceding /m/.

A study of steady-state responses to single-formant vowel-like stimuli reveals several interesting phenomena, as illustrated in Figure 2.9. In the figure the responses, plotted as period histograms, for a series of increasing stimulus levels, are shown for two particular nerve fibers, one tuned to the formant frequency (800 Hz), and one whose CF (2.79 kHz) is well above the formant frequency. The stimulus has a fundamental period of 10 ms which is clearly evident in the envelope of the waveform. This envelope periodicity is also prominent in all of the responses, **except** the responses of the fiber tuned to the formant frequency at higher amplitudes. For example, the mean response rates at 77dB SPL for the 800 Hz CF fiber and the 2790 Hz CF fiber are about the same. However, the fundamental period of voicing is prominent in the period histogram of the high frequency fiber, but not in the histogram of the low frequency fiber. There are at least two possible explanations for this result: either the filter characteristics are sharpened as amplitude is increased, to the point where only the single harmonic at 800 Hz is passed, or the 800 Hz response fiber is operating continuously at saturation. The latter is the more likely explanation, because a corresponding sharpening of the high frequency fiber's filter characteristcs should have resulted in a **decrease** in response to the distal formant, a result which was not observed.

2.3 Higher Auditory System

A number of synaptic stations provide links through which signals received at the level of the cochlea are reinterpreted and transferred to the auditory cortex in the brain [Moushegian et. al., 1972; Jeffress, 1972]. Each such station is a mass of nerve cells which recode incoming signals in ways that are for the most part obscure, although the picture is rapidly changing at the present time.

A succinct description of the central auditory system can be found in Pickles [1982]. The first station after the cochlea is the cochlear nucleus, which divides into two distinct areas, the dorsal and the ventral nuclei [Pickles, 1982]. Most of the afferent input from the cochlea is directed to the ventral nucleus. Axons from the ventral nucleus terminate in the superior-olivary complex, a complicated relay and reflex center. Within this complex, the medial superior-olivary nucleus (accessory nucleus) has been implicated as the lowest level of the auditory pathway where binaural interactions occur, because of clear bilateral linkages at this level. The responses of cells in this nucleus to stimuli presented in sequence to the two ears suggest that these cells are sensitive to interaural timing differences. Another component, the lateral superior olivary nucleus, receives a predominantly excitatory input ipsilaterally, and inhibitory contralaterally. Thus this nucleus may be responsible for detecting interaural intensity differences. Some anatomical evidence indicates an orderly point-to-point projection from the ventral cochlear nucleus to the superior-olivary complex, thus reinforcing the importance of the concept of "place" preservation.

The ventral cochlear nucleus appears to be mainly a relay center because cells found there are similar to primary auditory nerve fibers, and secure, short latency synapses transfer the signals to higher levels [Pickles, 1982]. In contrast, the dorsal cochlear nucleus is much more complex.

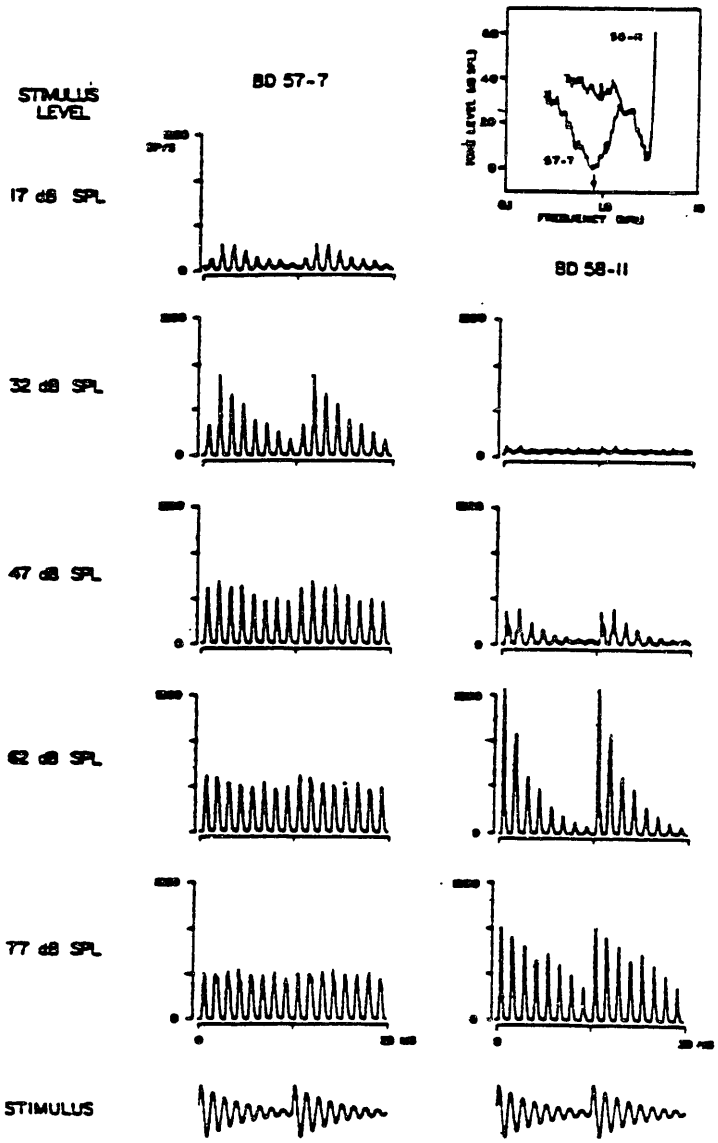

Figure 2.9: Response patterns of two auditory-nerve fibers to a single-formant synthetic stimulus at different levels. The stimulus has a 0.8 kHz formant frequency (F1), a fundamental frequency of 100 Hz, and a formant bandwidth of 70 Hz. The waveforms of two periods of the electrical signals to the earphone are shown in the bottom panels. The top right panel shows the tuning curves of the two units. The arrow points to the formant frequency of the stimulus. The CF's of units 57-7 and 58-11 are 0.78 kHz and 2.79 kHz respectively. Their spontaneous discharge rates are 50 and 32 spikes/sec. The response patterns are shown as period histograms, each of which is plotted twice. [from Delgutte, 1980].

Several different types of cells have been located here, and response patterns are complex.

One region of the posteroventral cochlear nucleus consists almost entirely of "octopus cells". Single auditory nerve fibers give rise to two types of synapse on these cells [Kane, 1973] – large primary endings and branching smaller secondary endings. When cells in this region are stimulated by tone bursts at CF, the response produces a sharp peak at onset followed by either very reduced or no activity. The suggestion has been made that the branching nerve fibers could be supplying both an excitatory and a delayed inhibitory response to account for the observed response pattern. When these cells are stimulated by a click train, the response will follow every click until the click frequency is increased to a critical point, beyond which response drops precipitously [Godfrey et al., 1975]. These cells have thus been implicated in the coding of stimulus periodicity.

Other cells have been located in the dorsal cochlear nucleus which show an excitatory response to a narrow frequency range and an inhibitory response to frequencies near the edges of the excitatory center [Evans and Nelson, 1973]. It has been proposed that the inhibitory input may be arising from other neurons within the dorsal nucleus [Voigt and Young, 1980]. Evidence has also been found for response patterns that are independent of overall signal level, and instead are related to the **contrast** in the stimulus pattern [Evans and Palmer, 1975; Evans, 1977].

Two further links in the auditory chain are the inferior colliculus and the medial geniculate body, which is the last synaptic station before the auditory cortex. The inferior colliculus combines the spatially coded input from the superior olive with the results of the complex sensory analyses of the dorsal cochlear nucleus. The medial geniculate body is usually classified into three subdivisions, ventral medial, and dorsal. The ventral division is probably mainly an auditory relay. Very little is know about the medial and dorsal divisions. At this stage, it is very difficult to sort out which aspects of a measured response are due to processing by the cell itself and which are due to previous processing at earlier stages in the pathway. Each cochlea is almost equally represented in both medial geniculate bodies and in both auditory cortices, because of the complex bilateral linkages that exist at previous levels.

Because our knowledge at this time of the central auditory system is fragmental, it is not feasible to incorporate this knowledge into the design of a "model" for higher auditory processing. As a consequence, a researcher exploring possibilities for further processing of the outputs of peripheral level models need not be constrained in any way in the choice of his or her strategies. Of course it also follows that such "models" have little hope of being "correct" in a neurophysiological sense. Still, it is of interest to see which types of strategies lead to the most promising results in terms of enhancing those aspects of the signal that are known to be important for perception. Furthermore, the results of such "models" may lead to an improved focus for researchers who are trying to understand the central auditory system.

Chapter 3

Auditory Modelling

Several researchers have developed models for specific aspects of human auditory processing. Some of these models are limited in scope to characterizing the steady state response of fibers in the auditory nerve to tones, including the detailed wave shape that reflects the phase locking property [Siebert, 1973; Colburn, 1973], and sometimes even including the generation of the spike sequence [Weiss, 1966]. Other models are concerned only with the envelope of the response, but deal with effects over time, such as adaptation phenomena [Smith and Zwislocki. 1975; Goldhor, 1983, 1985]. Some models, such as the model for the hair cell transduction process developed by Allen [1983], attempt to relate the neurochemical mechanisms to electrical circuit analogues. Finally, there are models that explore possibilities for further processing beyond the peripheral level of the auditory system, and thus often include some form of synchrony measure that would improve frequency resolution or extract relevant temporal information. Since little is known about the central auditory system, researchers have used considerable freedom in design strategies. These models are usually developed in the context of a specific task such as speech formant extraction [Sachs and Young, 1980; Srulovicz and Goldstein, 1983] , pitch estimation [Terhardt et al, 1982; Goldstein, 1973; Wightman, 1973], and sound localization [Lyon, 1983; Colburn, 1973].

One can evaluate a given model using either theoretical or experimental criteria. A theoretical approach developed by Siebert [1973], was based on the assumption that the neural spike sequence described a nonhomogeneous Poisson process. Siebert used mathematical formulations to predict whether rate response alone was sufficient to explain the known psychophysical ability to detect tones in noise. Colburn [1973] applied the same approach to predict the amount of sophistication necessary in the central processor to match the human performance in sound localization. Srulovicz and Goldstein [1983] used this Poisson formulation to predict frequency JND's (Just Noticable Differences), and the resolving capabilities for tones in tone complexes.

In contrast to the theoretical approaches that involve derivations based on mathematical formulations, there are computational models which specify a system through which any input signal can be processed to generate an output waveform that can then be subjected to further analysis. This approach characterizes the work of Searle et al. [1979, 1980], Lyon [1983] and Goldhor [1983]. Often the system is complex enough that it is difficult to sort out the individual effects of the various phenomena that are incorporated in the model. A major advantage, however, of this approach over the theoretical approach is that an output waveform is obtained for any input signal. Thus for example, processing of speech through the system can be compared with other, more standard, methods for speech analysis.

An ultimate test for such models would be to develop a speech recognition strategy based on the model outputs. In many cases, it is sufficient to demonstrate that a certain system can perform

limited recognition tasks, without attempting to solve the complete problem. Searle et al. [1980], for example, used stop identification as a criterion for demonstrating the utility of their system. Superior recognition performance, even for a restricted task, would be a good indicator that the model is an improvement over existing methods for speech analysis, and would generate further enthusiasm for auditory modelling among engineers interested in computer speech recognition.

3.1 Proposed Models for Peripheral Auditory System

Several researchers have attempted to design models for the response patterns observed in the peripheral auditory system. Most have concentrated on the response during steady state conditions to pure tones, and some deal only with an expression for the probabilistic rate response, whereas others generate an explicit spike train. There are also models that deal with phenomena that occur over time in non-steady state conditions. Such models attempt to characterize the effects of short-term adaptation on the **envelope** of the response characteristic, focusing on the observed linear incremental response characteristic. Yet another approach, one which characterizes the work of Jont Allen [1984], is to develop a circuit analogy for the neurochemical processes that take place in the hair cell transduction process.

One of the earliest models is the one proposed by Weiss [1966] consisting of a cascade of three fundamental elements: a linear time-invariant tuned filter, a transducer to model the action of sensory cells, and a spike generator to model neuron action potentials. Randomness was introduced by adding Gaussian noise to the filtered and transduced acoustic signal. Spikes were generated whenever the signal level exceeded a varying threshold, which was reset to a maximum level R_M immediately after each spike, and decayed exponentially to a minimum level R_m. This basic structure has been assumed by a number of researchers who improved upon Weiss's model only by modifications in one or more of these three elements.

Siebert [1973] characterized the response mathematically by assuming the spike train represented a nonhomogeneous Poisson process. He developed an expression for the Poisson rate function, $r_i(t)$, as a function of the stimulus waveform, $p(t)$, assuming the stimulus consisted of a tone in the presence of weak background noise. Since the form of the rate function resembles the one used in this thesis, it will be described here rather explicitly. The formula he used for $r_i(t)$ was as follows:

$$r_i(t) = \frac{f[p_i(t)]}{1 + K < p_i^2(t) >}$$

where $p_i(t)$ is the response to the stimulus $p(t)$ of a linear filter with the appropriate tuning curve response, and $f(x)$ is a positive non-linear function approaching Ax^2 for x large and positive, and approaching zero for x large and negative. The formula is only suitable for steady state response characteristics; it is thus assumed that the average power in the filter output, $< p_i^2(t) >$ is a constant for a given steady-state input stimulus. Furthermore, there is an implicit assumption that the response is completely synchronous to the input signal. Thus the formula is most appropriate

for low frequency stimuli. A major difference between this model and the one by Weiss is that the probabilistic spike pattern is determined by a mathematical formula rather than by an explicit threshold crossing of the filtered signal $y_i(t)$.

The most sophisticated model of this sort in the literature is one generated by Johnson [1974], consisting of a linear filter followed by a complex nonlinear device and a separate exponential rectifier. As in the Siebert model, an expression for $r(t)$, the probabilistic rate response, was obtained, and an explicit spike train was never generated. The nonlinear device consisted of three parallel branches, one representing the asynchronous component, one representing the synchronous component, and one accounting for responses to a second harmonic introduced by cochlear nonlinearities at high signal levels. Each branch contained an automatic gain control (AGC) whose form resembled in form although not in detail the formula used by Siebert. In contrast with Siebert's model, the AGC and rectification are separated into a two-stage process, and the rectifier has an exponential rather than a square-law form. Johnson was able to demonstrate a fairly good correspondence between output waveforms generated by the model and actual wave shapes derived from auditory nerve fiber histograms.

Jont Allen [1984] has expanded upon an original model by Davis [1958] to describe the transduction process that takes place in the hair cell in electrochemical terms. The relationship between receptor potential in the hair cell and hair deflection [Hudspeth and Corey, 1977] describes a variable resistance function which is a major component in Allen's system. Allen associates the current through this resistance with the neural response, after a lowpass filter to account for the reduction in phase-locking properties with increasing frequency. A cell capacitance and cell leakage resistance are also included in the circuit model. The model appears to match quite well neural data obtained by Johnson [1974] on responses of nerve fibers to pure tones.

Another important aspect of neural response patterns is the adaptation that occurs when a signal has been present for a long time. Smith and Zwislocki [1975] proposed a model for the **envelope** of the response over time which accounted for the observed linear incremental response characteristic. This model consisted of a linear filter, describing the underlying time course of adaptation, preceded and followed by memoryless saturating nonlinearities.

3.2 Speech Processing Systems based on Peripheral Models

In this section we describe existing computer speech processing systems which make use of knowledge about the peripheral auditory system in the design of the filter characteristics and/or some nonlinear amplitude compression schemes. Some pioneering work in this area was done by Chistovich et al. [1974] and Dolmazon et al. [1977]. More recently, Searle et al. [1979, 1980] developed a system that is motivated by the auditory system, and demonstrated its utility in a stop recognition experiment. The system consists of a bank of butterworth filters with 1/3 octave spacing and 1/3 octave bandwidths. These filters approximate the critical bandwidth fairly well for frequencies above about 500 Hz, but below this frequency the butterworth filters continue to improve in resolution whereas the auditory filters appear to maintain a constant bandwidth. The

filtering process is followed in the model by an envelope detector and a log compressor. Thus detailed timing information is lost.

The filter outputs were used as the input to a stop-recognition algorithm, in order to demonstrate the feasibility of the model for speech processing. The feature analyzer yielded approximately an 80% correct classification of the stops, /p/, /t/, /k/, /b/, /d/, and /g/, in a pre-stress context. The authors emphasize the important fact that the critical-band filter bank yields good temporal resolution at the high frequency end, suitable for detection of the rapid onsets that characterize stop consonants, and good frequency resolution at the low end, necessary for extracting formant information.

Two systems which go beyond the model of the critical band frequency responses to include some features designed to model short-term adaptation are the system developed by Lyon [1982] and the one designed by Goldhor[1983]. Lyon's system models the basilar membrane motion through a transmission line analogy; his filters consist of a cascade of notch filters with a number of parallel taps each passing through a resonator to yield the individual filter outputs. The filter responses can be made to be very steep on the high frequency end, reflecting the most recent data on tuning curves. The filter outputs are then passed through a half wave rectifier followed by a coupled AGC compression network. The AGC system includes feedback of the output through leaky integrators with differing time constants, to reflect both slow and rapid response processes, and also includes lateral inhibition effects, by allowing filter outputs to be affected by the outputs of neighboring filters. The system yields spectrogram-like display outputs, which Lyon refers to as "cochleagrams". An intermediate output of the model was used as the input to a model for binaural localization, which will be described in the next section on synchrony models.

Goldhor's system includes a linear critical band filter bank followed by a saturating nonlinearity and an adaptation scheme that is based on a hair cell transduction model. He was able to demonstrate that the model responded to incremental responses in a way that was consistent with the data obtained by Smith and Zwislocki [1975] on short term adaptation phenomena. His filter bank design strategy was to characterize the magnitude and phase response characteristics of each auditory filter as coefficients in the spectral domain, and then to inverse transform the result to yield the coefficients of an equivalent FIR filter. Each filter output was processed through an envelope detector, thus losing the fine-time response, and the slowly varying output was then passed through a memoryless nonlinearity to clip the high amplitude responses in a soft way. A raised hyperbolic tangent function was used for this step, although he suggested that many other functions could be used for nearly identical results. The final step was to process the compressed outputs through an adaptation circuit, consisting of three parallel paths, a capacitor in one path, a resistor in the second, and a diode, voltage-source, and second resistor arranged in series in the third parallel path. The current roughly corresponds to neurotransmitter fluid, and the capacitor models storage of the neurotransmitter in the hair cell. The voltage source, representing the receptor potential of the cell, is the input, and the current through the series resistor is the output, which should correspond to mean firing rate. He emphasized that the circuit does **not** act as an AGC, and, in particular, that response increments are independent of the amount of time that has

elapsed between the onset of the pedestal and that of the burst, as in the Smith and Zwislocki data [1975]..

3.3 Models for Central Processing

Models for processing that might take place at a more central level of the auditory system have been proposed for a variety of specific tasks, including sound localization [Lindemann, 1983; Colburn, 1973; Lyon, 1983], spectral peak enhancement for speech processing [Sachs and Young, 1980; Delgutte, 1984], and pitch detection [Licklider, 1959; Wightman, 1973; Goldstein, 1973; Terhardt, 1982]. Since pitch is the subject of the next chapter, models for pitch processing will be deferred until then. All of the models discussed in this section have in common the fact that they propose some form of further processing of the spike sequences generated at the peripheral level that involves an examination of the temporal fine structure.

A number of researchers have recognized the need for temporal processing to detect the time delay between the arrival of a signal at one ear and the arrival at the other, in order to localize the sound. The first serious proposal of a model was described by Jeffress et al. [1956] as follows:

> "This mechanism [for binaural localization] receives impulses from corresponding filter sections of the two ears and delays them progressively by small increments, either by means of fine nerve tissue with a slow conduction rate or by a series of synapses. The delay nets are in opposition, so that undelayed impulses from one side meet delayed impulses from the other. A time delay in the stimulus to one ear can therefore be matched by an equal delay in the neural channel from the other. A series of detectors, in the form of synapses requiring coincident impulses from both ears in order to respond, completes the mechanism. As is usual in such models, the device achieves precision statistically by the use of large numbers of elements."

This proposed model has been examined in greater detail by Colburn [1973], who used a theoretical treatment, and by Lyon [1983], whose approach was to implement a "computational model" of the proposed system. Colburn examined a variety of different models for lateralization, and compared performance with known human performance. He was able to demonstrate that an optimal processor, which made full use of the timing information available in the spike sequences, would obtain a performance far superior to the observed performance. His models were mathematical in nature, and included a detailed characterization of the steady state rate characteristic as in the Siebert model. He found that a model similar to the one outlined by Jeffress was able to explain the human localization capabilities quite adequately, and was more attractive than the optimal model because it could be more easily implemented in a biological nerve network. He pointed out that such a model would only require simple neural networks containing delays, coincidence counters (AND gates), and weighting constants, and that such elements are not unreasonable in the context of what is known about neural network capabilities.

Lyon [1983] proposed a model for lateralization based on a **cross-correlation** of the outputs from fibers with the same frequency selectivity from opposite ears. His model is thus in many

respects similar to the Colburn model, except that his method for demonstrating the utility of the model is to analyze complex data through the system, and to examine the resolving capabilities experimentally instead of theoretically. Peaks in the short-time cross-correlation function were interpreted as lateral directions, and at each time sample the delay parameter corresponding to the highest of the correlation coefficients is interpreted as the apparent direction of the signal. He extended the model to attempt to separate a signal into distinct sound streams, and applied the processing to a signal which was the sum of a speech signal, recorded binaurally under non-reverberant conditions, with the sound of a ping-pong ball being struck by a paddle in a highly reverberant room. The system separated the signal into four sound streams representing sounds from two presumed source directions and the left and right echos.

An extension to the cross-correlation model to incorporate lateral inhibition has been proposed and investigated by Lindemann [1983] in Blauert's laboratory in Germany. In the model, high valued cross-correlation products result in an inhibition of the gain in the signals proceeding down the delay lines on both sides. Thus an effective sharpening of the correlation peak is realized. A correspondence between the model and some observed psychoacoustical phenomena was also demonstrated.

Relatively little previous research has been done in the area of modelling mechanisms that might exist in the auditory system for enhancing formant information in speech processing. Sachs and Young's data [1980] on the response of a large population of nerve fibers to speech-like stimuli suggest that mean rate response yields an inadequate representation for the speech spectrum. These authors advocate further processing to filter the rate response such that only rates synchronized to the center frequency of the filter are included.

In contrast with the approach of zeroing in on frequencies close to the center frequency of the filter is an approach which seeks to determine the dominant frequency present in the output of each critical band filter [Carlson et al, 1975]. The supposition is that if the energy at a resonant frequency is high enough, then this frequency will dominate in the patterns of rate response of a large number of filters, both at and above the frequency of the resonance (read formant). A histogram of "frequencies", measured, for example, as average zero-crossing rates, should reveal clusters at the various formant resonances.

Sachs and Young adopted the approach of processing speech first through a large population of auditory nerve fibers of the cat's ear, and then applying computer algorithms to further process the response histograms in order to enhance formant information. Thus their "model" included as an initial step a portion of an existing auditory system. They first determined that rate response alone was insufficient, because formant peaks tended to merge into broad plateaus, even at levels that are typical of conversational speech. They showed that because of saturation and two-tone suppression, the rate response for the vowel /ɛ/ as a function of frequency varied greatly with overall signal amplitude. Since the perceptual quality of the vowel /ɛ/ does not change drastically as amplitude is varied, it is likely that rate response alone is not sufficient for the task.

The authors then analyzed the patterns in firing rates, by means of period histograms and interval histograms, and devised a processing method which exhibited substantially less variability

with signal amplitude, and overall spectral characteristics more like those of the original speech sound. The method consisted of averaging together the interval rate responses of a collection of fibers whose center frequencies were close to a given harmonic of the pitch. Only the energy in the spectrum of the interval histogram at the frequency of the harmonic under consideration was included in the average. Thus they obtained values for an "ALIR" (Average Localized Interval Rate), at discrete points in frequency corresponding to the harmonics of the pitch. Except for strong responses to the second harmonic of F_1 for high amplitudes (due to nonlinearities in the peripheral auditory system) the responses were much more similar across a wide amplitude range than were the rate responses alone. Furthermore, peaks in the response showed up at the appropriate formant frequencies.

One restriction in the method is that spectral analysis of the peripheral level outputs was performed at distinct frequencies equal to the harmonics of the fundamental frequency of the stimulus. It is unlikely (although not altogether impossible) that the auditory system uses pitch synchronous analysis. If Sachs and Young had sampled the spectrum at intervals independent of the pitch there probably would have been much greater fluctuations in the spectral shape, since most of the filters would be centered at frequencies intermediate to the harmonic frequencies.

Another serious problem with the Sachs and Young method is that it picks up a strong response at twice the first formant frequency. For normal speech sound levels, this response was in fact stronger than the response to F_2. The energy at $2F_1$ is introduced as a consequence of the half wave rectification inherent in nerve fiber response, plus nonlinearities causing wave shape distortion at high levels. Fibers located on the basilar membrane at $2F_1$ tend to respond mostly to the strong signal at F_1. Although the period of the resulting distorted sine wave is equal to the period of the first formant, Fourier analysis will reveal significant amounts of energy at $2F_1$, as a consequence of the missing negative half of the sine wave and the distortion in the positive half.

A more recent model for auditory processing of the spectrum is given by Srulovicz and Goldstein [1983]. This model is not based on pitch harmonic-synchronous spectral analysis, but would still suffer from the problem of energy at the second harmonic of the first formant. Like the Sachs and Young model, their model begins with the interval histogram of auditory-nerve spikes at each place on the basilar membrane. This interval histogram is then processed through a filter which is matched to the characteristic period of the nerve fiber. They suggest the following idealized form for the impulse response of the matched filter:

$$w(t) = 1 + \cos \omega_0 t$$

where ω_0 is the center frequency to which the nerve fiber is tuned. Thus the energy in the spectrum of the response at or near the center frequency of the peripheral filter is combined with the energy near DC to produce the final response. This procedure is very similar to the Sachs and Young strategies except that a mean rate response as well as a synchronized rate response is included in the second filter output.

Srulovicz and Goldstein have not tested their model on actual auditory nerve fiber response

data, nor have they attempted to implement a simulation through which actual speech data could be processed to provide insight into its performance characteristics. Instead, they chose the approach of a mathematical analysis based on the properties of Poisson processes, using equations for statistical rate response that had been previously derived by other researchers [Johnson, 1974; Colburn, 1973]. They were able to demonstrate a correspondence between the predicted JND's in frequency and actual JND's available from psychoacoustical data, except that the theory predicted a continued improvement as the time of the stimulus was increased, as opposed to a leveling off after a finite time interval observed in the subjects' performance. This difference was due to an assumption in the model that integration could take place over as long a time interval as was available.

In contrast to the above proposals which examine synchrony to the center frequency of the nerve fiber is an approach which determines the dominant frequency in each filtered waveform, producing a two-dimensional plot of frequency versus frequency. Carlson, Fant and Granstrom [1975] developed a system consisting of a bank of 120 linear filters, each of whose output was processed to obtain the mean zero crossing rate over an interval of 100 ms, to obtain an estimate of the dominant frequency. A histogram of the dominant frequencies, obtained by combining the data from all of the filters, should show peaks at formant frequencies. For example, for the synthetic vowel /i/, a strong peak in the histogram occurs at F_1 and another at F_2', the frequency of the best-matched second formant selected by subjects in a two-formant stimulus matching experiment.

This approach has been modified and expanded upon recently, with a new definition for spectral prominence based on the peak in the frequency-weighted DFT, instead of mean zero-crossing rate [Blomberg et al., 1983]. The speech waveform is first processed using standard DFT analysis techniques, and then a number of stages of spectral processing are introduced to model such features as Bark scale frequency distortion, psychoacoustical masking, and loudness perception. Each "critical band" filter output is generated by multiplying the processed DFT spectral representation by an appropriate frequency window. The peak of this weighted spectrum then represents the dominant frequency in the output at the corresponding place on the Basilar Membrane. A plot of dominant frequency versus center frequency, referred to by the authors as "DOMIN", then constitutes a spectral representation not unlike the zero-crossing plots as described above. The authors were able to use the DOMIN plots to obtain a "template" which could then be plugged into a standardized template-matching recognition scheme. The interesting result in the context of this thesis is that, for vowels, recognition accuracy for DOMIN was superior to recognition accuracy from the initial, smoothed DFT representation.

Delgutte [1984] has taken the Sachs and Young approach of using actual auditory nerve data as the input to the synchrony model, and has explored a number of different synchrony methods that fall into the two separate classes of examing synchrony to the center period or detecting the dominant periodicities. The stimuli consist of nine different two-formant synthetic vowels, where the formant frequencies are usually **not** multiples of the fundamental frequency. He examines four different synchrony measures, each of which is applied to the period histogram obtained from a group of fibers centered at a particular frequency. The methods differ only in the filter characteristics, which are shown here in Figure 3.1. Two of the methods use a bandpass filter with the

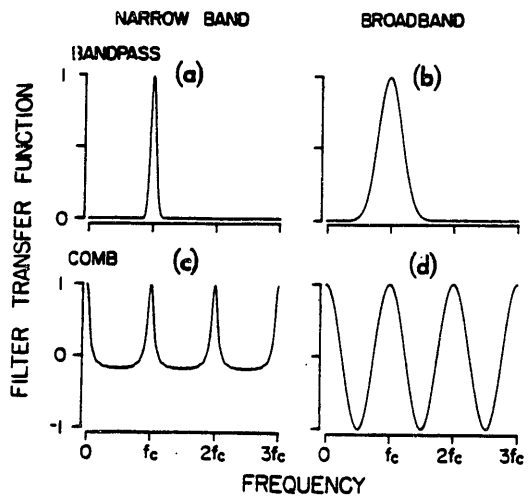

Figure 3.1: Transfer function of four different filters with center frequency f_c that were used to analyze period histograms for the computation of average localized synchronized measures (ALSM). [From Delgutte, 1984].
a) 1/6 octave Gaussian bandpass filters
b) 2/3 octave Gaussian bandpass filters
c) 1/6 octave comb filters
d) Cosinusoidal comb filters

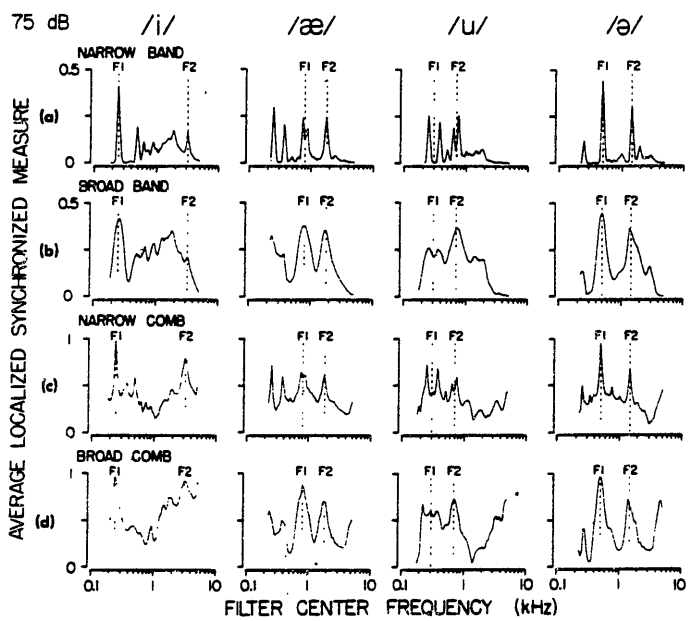

Figure 3.2: Four average localized synchronized measures plotted against center frequency of the analyzing filter for the /i/, /æ/, /u/, and /ə/ stimuli presented at 75 dB SPL. For each measure, the places of the formant frequencies along the f_c dimension are marked by dashed lines.
(a) ALSM obtained using the 1/6 octave Gaussian bandpass filters of Figure 3.1(a),
(b) ALSM obtained using 2/3 octave Gaussian bandpass filters of Figure 3.1(b),
(c) ALSM obtained using 1/6 octave comb filters of Figure 3.1(c), and
(d) ALSM obtained using the cosinusoidal comb filters of Figure 3.1(d).
[from Delgutte, 1984]

passband centered on the f_c of the fiber group, and the other two use comb filters that selectively pass multiples of the fundamental period, which is the center period. The other dimension that is varied is the width of the passbands, so that the four conditions are narrow band, broad band, narrow comb, and broad comb.

The methods differ from the original Sachs and Young approach in several ways. First, no assumptions are made about a knowledge of the fundamental frequency. Secondly, the filters are constant-Q instead of fixed bandwidth, and are in general considerably broader than the 50 Hz effective bandwidth of the Sachs and Young filters [Delgutte uses 1/6 octave and 2/3 octave for narrow and broad respectively]. Finally, mean rate response is normalized out of the final output, by dividing the estimate of the energy in the filtered waveform by the square of the mean rate response.

The results for four of the vowels are reproduced here in Figure 3.2. For the most part, peaks show up at formant frequencies, and the broad band analysis produces smoother, but broader peaks than the narrow band. The comb methods are to be preferred for the /i/, to the extent that the two largest peaks are the first two formants, which is not the case for the band pass methods. There are several peaks below the first formant for the /æ/, because individual harmonics are resolved by the narrow auditory filters in this region. Such information could be useful for determining the fundamental period, but would rule out the possibility of a template matching strategy for recognizing the appropriate pattern of the vowel. In the case of the /u/, the first formant frequency is approximately midway between two harmonics of the pitch. None of the four methods is able to produce a peak at the formant frequency. In the /ə/, both formants are positioned at harmonic frequencies, and thus both show up as the two strongest peaks in all four methods.

Delgutte also examined the approach of locating the dominant component in each fiber group histogram. To estimate the dominant component within a particular frequency band, a series of 1/6 octave Gaussian bandpass filters were applied to the period histogram, and the center frequency of the filter which yielded the largest output was chosen as the dominant frequency. For each fiber group, a separate dominant component below 200 Hz and above 200 Hz were detected, corresponding to the pitch frequency and the formant frequency respectively.

The dominant component below 200 Hz was always the fundamental frequency of 125 Hz. All of the formant frequencies were dominant components for several of the fiber groups, except the second formant of /i/ and the first formant of /u/. The fibers whose center frequencies were between the first and second formants of /i/ tended to respond to their f_c's, and the fibers near the first formant of /u/ responded to one or the other of the two harmonics stradling the formant. Responses to harmonics below the first formant of /æ/ also occurred. Thus all of the problems apparent in the synchrony-to-f_c method also show up in the dominant-component method.

Chapter 4

Pitch Perception: Possible Mechanisms

What could be said generally about the subject of pitch perception in the 19th century was that the pitch of a low frequency tone was essentially the same as its frequency but that a set of high frequency tones that were multiples of a common fundamental, if presented simultaneously, could be percieved as a simple entity with a "pitch" equal to that of the common fundamental. Helmholtz [1853] originally postulated that it was necessary that a component at the fundamental be present in the stimulus in order for the pitch to be heard, and that therefore a response at the place on the basilar membrane corresponding to the pitch frequency was responsible for the pitch perception. Experiments by Seebeck [1843] which used siren signals for which no energy existed at the fundamental seemed to contradict this theory, but Helmholtz maintained that nonlinearities in the cochlea generated difference frequencies which then stimulated the proper place.

Helmholtz's position remained largely unchallenged until Schouten [1938, 1940a,b] in Holland published his landmark work on "residue" pitch. He generated a complex signal with a clear pitch percept, but no energy at the fundamental, essentially a more accurate version of the Seebeck demonstration. But then he added to the signal a tone at a frequency slightly different from the missing fundamental, and demonstrated the absence of a beating phenomenon. A beating phenomenon would be expected if energy at the fundamental had been generated at the input to the basilar membrane through nonlinearities. Another similar experiment was demonstrated by Licklider [1954], who added to the high frequency harmonic complex a noise centered on the fundamental frequency. The pitch could be clearly heard in spite of the fact that the noise would mask any relevant low frequency energy in the signal.

Another line of attack that proved to be most fascinating and instructive was the study on perception of inharmonic signals, begun by Schouten, and continued in greater detail by deBoer [1956] and later Patterson [1973] and others. These experiments not only demonstrated that the pitch is not a consequence of nonlinear distortion, but showed that the pitch is **not** perceived by somehow measuring the spacing between peaks in the spectrum at the places of the harmonic complex. These studies suggest in fact that pitch is perceived by examining periodicity in a waveform or waveforms derived from the responses at the frequencies of the harmonics.

DeBoer's experiment consisted of generating a series of seven sine waves spaced by 200 Hz but offset in frequency from true harmonics by an amount ranging from -100 to +100 Hz. The perhaps surprising result is that the pitch that is perceived is not 200 Hz but rather something that differs by an amount that correlates with the offset of the harmonics. As the harmonic frequencies are increased from $f_0 n - f_0/2$ to $f_0 n + f_0/2$, the perceived pitch also systematically increases from about 190 Hz to about 210 Hz. In fact, the pitch is ambiguous when the harmonics are offset precisely $f_0/2$ (100 Hz).

Licklider [1959] proposed an auditory model for pitch perception based essentially on autocorrelations at a series of periods spanning the pitch range. The signals that were correlated were presumed to be the joint responses from the two ears that had arrived simultaneously at a place in an x-y plane, where x is the place dimension, roughly corresponding to frequency, and y is the time dimension, corresponding to interaural delay, used for spatial orientation. Following the autorcorrelations, there is a transformation of patterns of periodicity in space to points that correspond to frequently recurring stimulus conditions. After training, periodic patterns elicit strong responses at the relevant loci, that are interpreted as pitch. Stimuli that are fuzzier but still periodic stimulate a localized region of points centered near the focus of the perceived pitch.

Other researchers continued to collect more data on pitch perception, with the hope of elucidating the neural pitch mechanism. Nordmark [1963] compared the results of pitch and lateralization phenomena produced by filtered and unfiltered pulse pairs of same or different polarity. He predicted that if time delays between neural pulses were processed for pitch perception, then changes in the perceived pitch when the polarity of the two stimulus pulses is reversed should correspond closely with time differences derived from localization studies. Pulse pairs of either normal or reversed polarity were processed through lowpass filters of varying cutoff frequency and listeners were asked to adjust the spacing between the pulse pairs until the perceived pitch matched the pitch of a fixed pure tone. He found that the difference between the pitch of the reversed polarity pair was significantly greater when the pulse pair was filtered with a 600 Hz cutoff lowpass filter, than, for example, when the pair was unfiltered. This result could be explained by the amount of additional lag necessary to align a compression peak with the next rarefaction peak of a low frequency wave as contrasted with a high frequency wave. More significantly, a plot of the time difference between the period of the tone and the distance between the pulses as a function of cutoff frequency of the filter showed a remarkable correspondence with Flanagan's [1962] data on lateralization shift for similar antiphasic clicks.

Ritsma and Engel [1964] proposed that pitch is perceived as the spacing between peaks in the waveform that are near the peak in the **envelope** of the waveform. They compared a signal consisting of three equally spaced sine waves in cosine phase (AM modulated) with a signal having the central tone shifted by 90 degrees (quasi FM). In the latter case the perceived pitch is considerably weaker and more ambiguous. Using signals centered on a 2000 Hz carrier, they collected histograms of pitch matches by two subjects for a variety of spacing frequencies and relative amplitudes, and compared the peaks in the histograms with the spacing between peaks in the detailed wave shapes of the resulting waveforms. In all cases, whenever two peaks in the detail straddled a peak in the envelope the histogram showed two clusters at two periods stradling the period of the envelope, left and right.

Thus an association was made between peaks in the fine structure and pitch. However, the tacit assumption is being made that the auditory system has access to the original waveform. This is clearly not the case. If the frequency band of the signal is sufficiently high and sufficiently narrow, then the appropriate critical band filter should pass most of the signal, although with some phase shifts and amplitude distortions. If, however, the signal spectrum is in a band where individual

harmonics are resolved, then each filter output will reveal only the harmonic frequency, not the fundamental.

Studies on regions of dominance for pitch perception have shown that if information is present at the third to fifth harmonics, then such information will override conflicting pitch data at higher frequencies [Ritsma, 1967; Bilsen and Ritsma, 1967]. In this dominant region, the auditory filters are generally sufficiently narrow that the filter output at each harmonic frequency should resemble a half-wave rectified sine wave at that frequency. The envelope fluctuations in the stimulus with the fundamental period would be essentially lost in the output. The fundamental frequency could only be determined by integrating information across the place dimension.

A temporal approach could still be used in the dominant region for detecting the harmonic frequency, and then the outputs of a collection of fibers could be combined to obtain a decision about a common fundamental period. Moore [1982, pp140-144], proposed that intervals [not necessarily adjacent] between nerve firings could be detected and integrated over an appropriate time span for each filter. A histogram would then be constructed at some central location, including all of the observed time intervals from several channels in both ears. The interval equal to the fundamental period would be reinforced by all of the filters, and hence should show up as a peak in the histogram.

Most temporal models would require an accurate mechanism for detecting the delays between spikes, particularly models such as the one above where histograms of delay intervals are to be collected. Whitfield [1970] maintains that it is unlikely that the auditory mechanism for delaying nerve fiber responses is sufficiently accurate to obtain the kind of pitch resolution that psychophysical experiments yield. He suggests that inaccuracies in synaptic delays pose a major problem for temporal-based schemes.

An alternative approach to the problem is to obtain a spectral image from the outputs [perhaps mean rate response] of all of the relevant channels, plotted across the place dimension [Terhardt et al, 1982; Goldstein, 1973; Wightman, 1973]. Pitch is then determined as that frequency which presents a best match to the available harmonics, that are extracted as peaks in the spectrum, if these are all assumed to be integer multiples of this pitch. This approach seems to match the data on residue pitch quite well, and, in particular, naturally builds in the additional information generated by combination tones introduced through nonlinearities. The combination tone at $f_1 - (f_2 - f_1)$ generated by a two-tone complex appears to be an "essential nonlinearity" in the cochlea; i.e., is present even at low signal levels, and approximately proportional in amplitude to the amplitude of the stimulus [Plomp,1965; Goldstein, 1967]. Goldstein has suggested that in a harmonic complex several lower harmonics would be generated through nonlinearities, and that these could also be used by a central pitch processor to aid in detection of the fundamental. Particularly since better frequency resolution is available in the lower frequencies, such lower harmonics could be essential for a model that depends upon resolving peaks for determining pitch. If these generated harmonics are included in the series, then the pitch match obtained by a best least-common-denominator approach would correspond well with observed data.

This concept eventually led to several theoretical proposals of actual mechanisms that might be used to deduce a pitch as the best-fitting fundamental to a sequence of presumed harmonics.

Goldstein et al. [1978] proposed a method based on maximum likelihood procedures, to estimate first the harmonic numbers and then the fundamental frequency, assuming that the peaks corresponded to successive harmonics. The proposed analysis would take place in an optimal central processor which would absorb information from both ears in the form of a stochastic estimate of the locations (no amplitude information) of the spectral peaks in the place domain. Temporal analysis per se is not strictly ruled out, since timing information could be used to deduce the dominant frequency present in a given critical band filter. However, the temporal cues would only be used to determine the frequency of the **harmonic**, not of the fundamental. This approach can only work in regions where the individual harmonics are resolvable; thus the authors suppose that nonlinearities introducing additional harmonics below those present in the original signal may play a critical role.

A model proposed by Wightman [1973] involves the notion of a "PAP": a "Peripheral (neural) Activity Pattern", and thus views the problem as a pattern matching procedure. This PAP is interpreted as the pattern of average firing rate along the basilar membrane produced by a periodic signal, which will include undulations with the fundamental frequency. At the higher end of the spectrum the undulations will be less sharp, because of the limited frequency resolution, and, in fact, will be altogether missing for sufficiently high frequency and/or sufficiently low fundamental frequency. The pitch frequency is then extracted as the best estimate of the oscillation frequency of the waveform that is generated along the place dimension.

Both of these models would work very poorly if the periodic input signal consisted only of information above around the tenth harmonic of the fundamental, because of the limited resolution of the filters. An experiment by Moore and Rosen [1979] was developed expressly to demonstrate that pitch could be heard when the signal was restricted to regions above the place where harmonics would be resolvable. They generated simple melodies by varying the repetition rate of high-pass filtered pulse trains, to which had been added low frequency noise sufficient to mask out any combination tones. The melodies were heard quite accurately in spite of the lack of useful low frequency information.

Some neurophysiological support for temporal processing for pitch comes from a study by Smith et al. [1978] concerning frequency following responses [FFR's], measured from human subjects. The FFR is a low voltage neuroelectric wave that

> "is generally considered to be the aggregate envelope of the action potentials of a large group of phase- locking auditory neurons concentrated within major brainstem auditory nuclei".

Subjects were presented binaurally with two different periodic stimuli with the same fundamental frequency of 365 Hz. The signals consisted of a 365 Hz pure tone and a complex of four tones at 730, 1095, 1460, and 1825 Hz. Both were capable of evoking a strong 365 Hz periodicity in the FFR. Furthermore, low frequency masking noise centered at 365 Hz was able to significantly reduce the response to the 365 Hz pure tone, while having little effect on the response to the complex.

This physiological evidence shows not only that an envelope at the fundamental can be generated from high frequency components, but also that a similar envelope can be measured through an

identical procedure, from a low frequency sine wave. If a signal similar to the FFR can be made available to the central auditory processor, then the same mechanism (processing of the FFR-like signal) could be used to extract the pitch of both signals. What is perhaps even more striking is that the complex stimulated a region of the basilar membrane where individual harmonics should have been resolvable; even so, the envelope, exhibiting mainly periodicities at the pitch period, could be regenerated in the form of the FFR. Temporal processing of a single waveform would not necessarily require accurate delays, because a tapped delay line could be used to obtain an accurate measure for relative pitch, but not absolute pitch.

In summary, there is at the present time an open debate about the methods that may be used by the auditory system to detect periodicities that are perceived as pitch. There seem to be major problems with both temporal and place-based schemes. A system that relies upon the resolving of individual harmonics in the place dimension would work poorly for stimuli restricted to the high frequency region. Likewise, a system that depends upon detecting the fundamental period by temporal processing of the spike sequences would be inadequate in the low frequency region, where individual harmonics are resolved. Timing inaccuracies also are a problem for temporal methods. It may be the case that the auditory system makes use of a variety of different strategies that deliver their outputs to a single high-level moderator. However, the experiment by Smith et al. is suggestive of an approach which involves an initial combining of the outputs across the place dimension, with further temporal processing of the single resulting waveform. Such an approach will be developed as the pitch component in the thesis.

Chapter 5

Review of Pitch Detection Algorithms

5.1 Introduction

The task of classifying speech into voiced and unvoiced regions, and of determining the fundamental period of the glottal source during voiced regions, has proven to be much more difficult than was originally expected. A number of different methods have been reported in the literature, for applications ranging from vocoders to speaker identification to research tools for linguists and acousticians to aids for the deaf. Most of these methods have not sought to borrow from knowledge about how humans process pitch, but rather have viewed the problem strictly from an engineering standpoint. Some of the most recent work, however, has at least attempted to justify some of the methods used from the standpoint of human perception.

In this chapter some of the proposed algorithms will be described briefly, and some of the problems inherent in the various strategies will be brought up. A loose categorization into three classes is suggested:

1. **Waveform Methods:** Those that deal directly with the waveform, or a filtered version of it,

2. **Autocorrelation Methods:** Those that look for peaks in a function such as the autocorrelation, derived by comparing the original waveform with itself delayed through a series of delays covering the pitch range, and

3. **Spectral and Cepstral Methods:** Those that begin by computing a high resolution spectrum. Pitch is determined either from peaks in the log magnitude spectrum (spectral) or from peaks in the inverse transform of the log magnitude spectrum (cepstral).

5.2 Waveform Methods

The best example of a waveform pitch detector is the Gold-Rabiner pitch detector [1962], which combines the outputs of six parallel pitch period estimators into a single final decision based on majority rule. The six detectors are all derived from a cartoonized version of the lowpass filtered original waveform, which is obtained by setting to zero all samples which are neither peaks nor valleys. One of the six processors deals only with peaks; another only with valleys (negative peaks). Two others process waveforms obtained by subtracting the previous peak (or valley) from the current one. Finally, the last two deal with spikes generated by subtracting valleys from peaks, or peaks from valleys. All negative outputs are set to zero.

Each of these six spike trains is then evaluated for quality of its peaks. A variable threshold is set to the amplitude of the first peak, and then, after a blanking period, decays exponentially until a new peak is encountered that exceeds the threshold. The difference in time between the two peaks defines a pitch estimate. The three most recent estimates from each of the six detectors are fed to a final arbitrator which decides based on a majority rule. If the detectors are inconsistent, the decision is made that the segment of speech is unvoiced.

5.3 Autocorrelation Methods

Under the category autocorrelation we are also including other strategies such as comb filtering that, like the autocorrelation method, yield a zero-phase function with time [as opposed to frequency] as the ordinate. Strictly speaking, the cepstral method should perhaps also be placed in this category, except that a cepstrum, unlike an autocorrelation, cannot be derived from the original waveform without first performing Fourier analysis. The advantage that these methods have over methods that detect periodicities directly from the original waveform is that the absolute location of a single peak determines the pitch period. A peak in the autocorrelation at time t corresponds directly to a pitch period of t, because the origin is defined explicitly. Thus the algorithm for determining the value of the pitch is straightforward, and usually involves simply the detection of a maximum.

Autocorrelation methods do have disadvantages, however. The main problem is that a fixed time window is usually preselected for processing, regardless of the pitch frequency. It would be preferable to look over a very short time segment if the pitch frequency is high, but it is essential to look over a sufficiently long time for very low pitches. Usually the compromise is made at the high end, and thus a rapidly changing female pitch yields a broad ill-defined peak.

An additional problem is that the vocal tract resonances are sometimes changing rapidly over the interval of the window. In this case, the waveform is not well correlated with itself because of the changing spectral shape, and thus the peak at the fundamental period is diffuse. Furthermore, additional peaks in the autocorrelation occur at the periods of the formants and multiples thereof, which may confuse with the actual pitch period. For these reasons, many autocorrelation methods include a form of spectral flattening in order to remove the effects of the vocal tract as much as possible.

An example of a pitch detector that uses autocorrelation on a spectrally flattened waveform is a scheme proposed by Sondhi [1968]. For this scheme, the speech waveform is first center-clipped with a threshold whose value is equal to a percentage of the waveform peak value over the window. Dubnowski et al. [1976] implemented in real-time hardware a modified version which used infinite peak clipping in addition to the center clipping, thus allowing the correlation multiplies to be replaced with logic. The difficult step with these approaches is to set the center-clipping threshold correctly. Particularly when the overall gain is changing over the frame, a single constant may not be sufficient. Thus legitimate peaks at the end of a vowel may fall below threshold. If the threshold is set too low, too many peaks may pass on the high-amplitude side of the window.

Another method for spectrally flattening is to use Linear Prediction to obtain an all-pole model, and then process the original waveform through the LPC inverse filter. This approach characterizes the SIFT algorithm developed by Markel [1972]. The lowpass filtered (900 Hz cutoff) speech was passed through a fourth order LPC inverse filter, and then the resulting error signal was autocorrelated. A peak in the autocorrelation function is then the pitch estimate. One problem that was encountered is that, in the case of high-pitched voices, the inverse filter would sometimes actually filter out individual harmonics, thus removing pitch as well as spectral information.

Other pitch detection strategies that are similar to autocorrelation methods are those methods that filter the waveform through a series of comb filters, and detect a null in the output energy at the appropriate comb period. An example is the method proposed by Moorer [1974]. Instead of an autocorrelation, he uses the following function, often referred to as an AMDF for "Average Magnitude Difference Function", as the zero-phase function from which to estimate the pitch:

$$y_M[n] = \sum_{i=n-N/2}^{n+N/2} |x[i] - x[i-M]|$$

$y_M[n]$, plotted as a function of M, which is varied over the pitch range, should exhibit nulls at multiples of the pitch period. The first prominent null is then the pitch estimate. Some heuristics are usually needed, as is generally the case for autocorrelations as well, to distinguish between the fundamental period and multiples of the fundamental period. Additional nulls due to formant information remain a problem for this method, unless spectral flattening procedures are applied.

5.4 Spectral and Cepstral Methods

We are choosing to define spectral methods as those methods which obtain the pitch by direct processing of a spectrum of the speech; i.e., by detecting a series of peaks in the spectrum. Such methods typically derive the pitch from either the greatest common denominator of the set of spectral peaks, or from measurements of the spacing between the peaks in frequency. The cepstral method includes spectral analysis, but obtains the pitch estimate from a function (the cepstrum) obtained by inverse transforming the log magnitude spectrum, to yield a zero-phase time-domain waveform, similar to an autocorrelation.

When originally conceived, the cepstral method was expected to work well because it operated in the context of a theoretical framework in which the convolution of the source with the vocal tract resonance filters was converted to a sum. It was predicted that source information should be associated with the high time cepstrum and spectral shaping information restricted to the low time cepstrum. To a first approximation, these predictions were true; however, for high pitched voices the first peak due to the source is typically mixed in with the spectral-shaping information. Furthermore, the cepstral method, like the autocorrelation methods, suffers from the problem of a fixed window size. From a signal processing standpoint, the cepstrum is very costly, particularly because the spectrum must be oversampled in order to avoid aliasing of the negative time cepstral

peaks into the positive time half. It also was one of the worst in performance of the pitch detectors tested by Rabiner et al [1976].

Methods that detect pitch in the spectral domain typically begin with an algorithm for detecting prominent peaks in the spectrum; pitch is then determined through some heuristic procedures that look for consistent patterns in the peak frequencies. Seneff [1978] developed an iterative strategy for examining spacing between peaks, such that high amplitude peaks were weighted more heavily than weak peaks. The strategy begins with the two largest peaks in the region from 200 to 1200 Hz, and the spacing between them constitutes a single pitch estimate. The next step is to add the third largest peak to the set under consideration, and to add two new estimates, defined as the distance between **adjacent** peaks in the set of three, to the growing table of potential pitch values. A new peak is added with each stage of the iteration, and a new set of pitch candidates are added to the list. The final pitch decision is made, as in the Gold-Rabiner time-domain algorithm, by a majority rule.

Errors in single-side-band transmission can lead to a frequency shift of the speech spectrum, producing an effect which turns the speech into an inharmonic sequence, and has a profound effect upon the perceived pitch. A "feature" of the Seneff pitch detector was that it did not depend upon the peaks being harmonically related, and thus it could undo the error at the point of resynthesis. However, because the human is very sensitive to such a frequency offset [see Chapter 4 for a discussion], it is clear that spacing between harmonics is **not** the method used by the auditory system.

There has been a recent interest in designing spectrally based methods for pitch extraction that would make the same errors that humans make with inharmonic sequences. Noteworthy is the work by Duifhuis and Willems [1982], which makes use of a "harmonic sieve" procedure, a simplificiation of the "optimum" estimator proposed by Goldstein [1973]. In their algorithm, the first step is to detect a set of peaks for consideration in the spectral region up to 2500 Hz. The spectrum was obtained by straightforward Fourier analysis of a 40 ms window of speech. Thus no attempt was made to include critical-band like spectral analysis. A sieve was then created, containing "holes" in narrow regions around each harmonic of a given fundamental under consideration. Any peaks which passed through the holes in the sieve were considered as candidate harmonics. The scoring criterion was based on the highest harmonic number that passed through the sieve, M_l, and the number of peaks below the frequency of M_l that were passed or rejected by the sieve. The sieve was applied incrementally over the full range of f_0 in speech, and the best-scoring match was the estimated pitch.

An enhanced version of the above algorithm was developed by Scheffers [1983], in which more attention was given to the characteristics of the auditory system. Scheffer's method included a stage of additional level-dependent smoothing of the power spectrum, in order to give an appearance more like the outputs of critical band filters. As a consequence, the resolving power for low frequency pitches was insufficient in the higher regions of the spectrum, resulting in a poorer performance for vowels in noise than the observed human performance.

Scheffer was however able to demonstrate that the harmonic sieve strategy matched human

performance for inharmonic sequences, provided that two additional harmonics were introduced artificially into the candidate set below the lowest detected peak frequency. The signals consisted of three tones spaced by 200 Hz, with the center tone ranging in frequency from 1200 to 2400 Hz. Motivation for the addition of peaks below the lowest peak present in the spectrum came from studies demonstrating an "essential nonlinearity" that seems to introduce such peaks at an early stage in auditory processing [Goldstein, 1967].

5.5 Summary

The pitch detectors described in this chapter are only a representative subset of the available pitch detection schemes. However, the major problems associated with the different strategies are clear. Any method that begins with a fixed window on the input waveform has to deal with a compromise between male and female voices. A 40 ms window is much longer than the desirable integration time for a 3 ms pitch period. However, a minimum width to detect a 16 ms pitch period is a 32 ms window, and even this is inadequate if the window happens to be centered on a glottal pulse.

Methods that detect peaks directly from the original waveform have considerable difficulty if the waveform has been processed through phase shifts that tend to smear out the peaks. In contrast, methods based on the log spectrum or autocorrelation strategies are insensitive to such phase shifts. In particular, peak detection alone would be completely inadequate for signals consisting of random noise added to a delayed version of itself. The pitch of such signals is perceivable, however, particularly if the pitch is swept in time [Bilsen, 1966]. It is probably necessary to compare the detailed waveshape with the delayed version, in some way, in order to detect such "repetition pitch".

A problem, however, for methods that rely on a repetition of the detailed wave shape, as opposed to simply detecting a peak, is that the resonances of the vocal tract are continually changing over time. Thus such methods tend to rely on spectral flattening schemes to reduce the variability from period to period. Nonlinear schemes applied directly to the waveform, such as center-clipping, are probably preferable to methods such as inverse filtering, which accept a fixed vocal tract model for the entire frame. On the other hand, center-clipping with a fixed threshold over the frame is also subject to errors at onsets and offsets of voiced regions.

Thus it is clear that there are still many unsolved problems in the area of pitch detection from human speech. A method which attempts to follow the example of the human auditory system is not necessarily to be preferred, in terms of performance, over a method that does not. To date, there are no published methods, to my knowledge, based on temporal processing, that claim to be auditorily motivated. The general feeling at the present time is that frequency domain approaches are more likely to be correct in a psychophysical sense than time domain approaches. One goal of this thesis, as will be shown later, is to demonstrate that the time domain need not be ruled out as a possibility for human auditory processing of speech.

Chapter 6

Current Spectral Representation Methods for Speech Recognition

For the most part, current speech recognition strategies are based on a spectral representation for the speech signal that borrows heavily from models of speech production. Such models are very useful for characterizing the signal and for leading to an understanding of the basic acoustic attributes that are important for phonetic identity [Fant, 1970; Flanagan, 1972; Rabiner and Schafer, 1978]. Such models are also clearly appropriate in analysis/synthesis systems, where the goal is to produce as accurate a reconstruction as possible of the measured signal. Often, however, the form of spectral representation that is used for synthesis is assumed to also be appropriate for recognition. Yet the two tasks are really quite distinct, and there is no reason to believe that what works for synthesis is suitable for recognition. The human brain obviously recognizes the signal only after it has been processed through the human auditory system. It is quite clear from studies of the peripheral auditory system that what is available at the level of the 8th cranial nerve is not a close copy of the log spectrum. Yet a standard approach to the evaluation of the effectiveness of a particular method for generating a spectral representation, even when it is intended to be used for recognition purposes, is to apply some error metric to define the deviation of the representation from the log spectrum.

In this chapter, we will first review very briefly the fundamentals of the theory of the speech production mechanism, including a discussion of voice-quality aspects that are generally unrelated to phonetic content, but represent additional information available in the signal. We will then show how the theory of speech production has been used to justify certain traditional methods of speech analysis, which have been applied to the two distinct tasks of resynthesis and recognition. We will follow this with a review of methods used to extract the relevant phonetic information from the spectral representation. One common approach is to match the pattern of the spectrum of the unknown speech sound to a collection of "templates" representing a catalog of canonic speech sounds. Opposing this approach is one which extracts relevant features from the available unknown spectrum, and then relates these to acquired data on the distributions of these features in the catalog of sounds. We will discuss the advantages and disadvantages of these two approaches.

6.1 Review of Speech Production Model

The speech signal is typically modelled as the convolution of an excitation source function with a filter describing the transfer function of the vocal tract in conjunction with radiation characteristics. The source function is usually periodic during voiced speech sounds, such as vowels, nasals, and

the voiced consonants /l/, /r/, /w/, and /y/, whereas a noise source characterizes sounds such as the unvoiced fricatives /s/, /sh/, and /f/, and stop bursts, such as /p/ or /t/. The periodic voiced source is produced through vibrations of the vocal folds at the glottis, and the fundamental frequency of voicing, or pitch, is a time-varying attribute that, for English, contributes mainly at the prosodic level. During the production of unvoiced sounds, a noise source is produced at some point of tight constriction in the vocal tract.

The shape of the vocal tract changes continuously with the rapid movement of the tongue, lips, and jaws that accompany the production of speech. The vocal tract can be modelled as an acoustic tube, and, as it changes shape the resonance frequencies move in a somewhat predictable way. Because the noise source for unvoiced sounds tends to be at a constriction in the mouth, only the natural frequencies of the front cavity are excited. As a consequence, the resonance frequencies tend to be high. Hence there is a rough division of voiced from unvoiced sounds based on the ratio of high to low frequency energy in the spectrum.

The resonance frequencies of the vocal tract, or formants, are manifested as prominent peaks in the spectrum, and their frequency locations convey much of the phonetic information in the signal. For certain sounds, such as nasals and nasalized vowels, the nasal tract as well as the vocal tract is involved in the production mechanism. In these cases, the nasal tract acts as a separate branch, causing additional resonances and antiresonances. The excitation function may at times also show a low-frequency resonance, that can appear as a distinct peak in the vowel spectrum.

The excitation and vocal tract shape change continuously over time, but the speech signal has as its underlying representation a sequence of basic linguistic units, called phonemes. Each phoneme is characterized in the abstract by a set of specific acoustic features, but interactions with adjacent phonemes tend in many cases to modify considerably the acoustic characteristics. Vowel sounds and most sonorant consonants are characterized in large part by the positions of the lowest three or four formants. Thus, for example, there is a wide separation between the first two formants in the vowel /i/ as in "see", whereas these two formants are very close together for /ɔ/ as in "caught". Formant movements are very important, both in terms of direction and relative rate of movement, for characterizing diphthongs such as /aɪ/ in "type", as well as adjacent consonants, particularly the liquids and glides [/l/, /w/, /y/, and /r/]. Rapid formant movements at the onset of the vowel are also important for the identification of the place of articulation of the preceding consonant.

In addition to the phonetic identity of a speech segment, there are often several other attributes of the signal that contribute to variability in the spectral representation. For example, the listener can usually identify the sex of the talker. How this is done is not completely understood, but two factors that are clearly important are the range of the fundamental frequency of voicing and the extent of breathiness of the voice. Female speech tends to have a high fundamental frequency, and the source spectrum tends to be characterized by a prominent peak at the fundamental frequency, which contributes to the percept "breathy". Whispered speech is also quite intelligible, in spite of the gross differences between a noise excitation and a periodic source function. In English, no phonetic distinction is made between nasalized and non/nasalized vowels. Nonetheless, the presence of the feature nasalized can alter significantly the spectral shape, particularly in the first

formant region. Typically, the first formant of a nasalized vowel is characterized by a pole-zero-pole complex. In such cases, a "peak-splitting" of the first formant occurs, which complicates peak-picking strategies for formant tracking.

These complex factors are retained at some level of representation, because we are able to perceive the appropriate differences. Presumably, however, the processing in the brain that is restricted to the task of phonetic identification is able to overlook, to a great extent, the variability in the signal caused by these additional factors. Spectral representations that have been developed to be used in computer speech recognition often retain too much of the inherent variability due to speaker, environment, etc., such that most "successful" speech recognition systems are strongly speaker-dependent, and are in addition restricted in terms of recording conditions.

6.2 Standard Methods for Speech Spectral Analysis

In this section we will discuss briefly some of the currently available methods for processing the speech waveform, which usually are based on an implicit or explicit assumption that the goal is to reproduce an accurate representation of the envelope of the log spectrum of the speech signal. We will give several examples of the various processing methods, applied to certain speech segments that are selected to illustrate some of the difficulties that can arise. We restrict our discussion to sonorant sounds, where the most important goal is to represent the positions and the movements over time of the formant frequencies. This restriction is imposed mainly because the thesis system has been restricted to the low frequency region, in large part because it is not clear that an approach based on synchrony is appropriate for the higher frequency region.

The traditional method of analysis is to compute a log spectrum using the standard short-time Fourier transform, for which the only constraints are the size and shape of the window that is applied to the speech prior to the transformation. The length of the selected time window depends upon the intended application. For purposes of producing a spectrographic display which the human can then examine visually, it is usually preferable to use a short time window, thus preserving crisp onset characteristics, and minimizing the blurring of the spectral characteristics of short duration sounds such as stop bursts. The duration of this window is typically on the order of 6 or 7 ms, and the resulting "wide-band" analysis typically retains information about the fundamental frequency of the excitation in the temporal, but not in the spectral, domain. In contrast to wide-band analysis is narrow-band analysis, with the window length set to around 25 ms. In this case, the window typically encompasses several glottal excitation pulses, and therefore the harmonic structure is preserved in the spectrum. However, the spectral representation is much more stable from frame to frame than wide-band analysis, which gives this representation a distinct advantage when the intent is further computer analysis of the spectrum.

Some of these points are illustrated in Figures 6.1 and 6.2. Part (a) of Figure 6.1 shows the entire waveform on a compressed time scale for the word "majority", spoken by a female speaker. Part (b) shows a small section of the /ɔ/, on which are superimposed two hamming windows centered at the same point in time. The short window is a typical size for wide-band analysis, and the longer

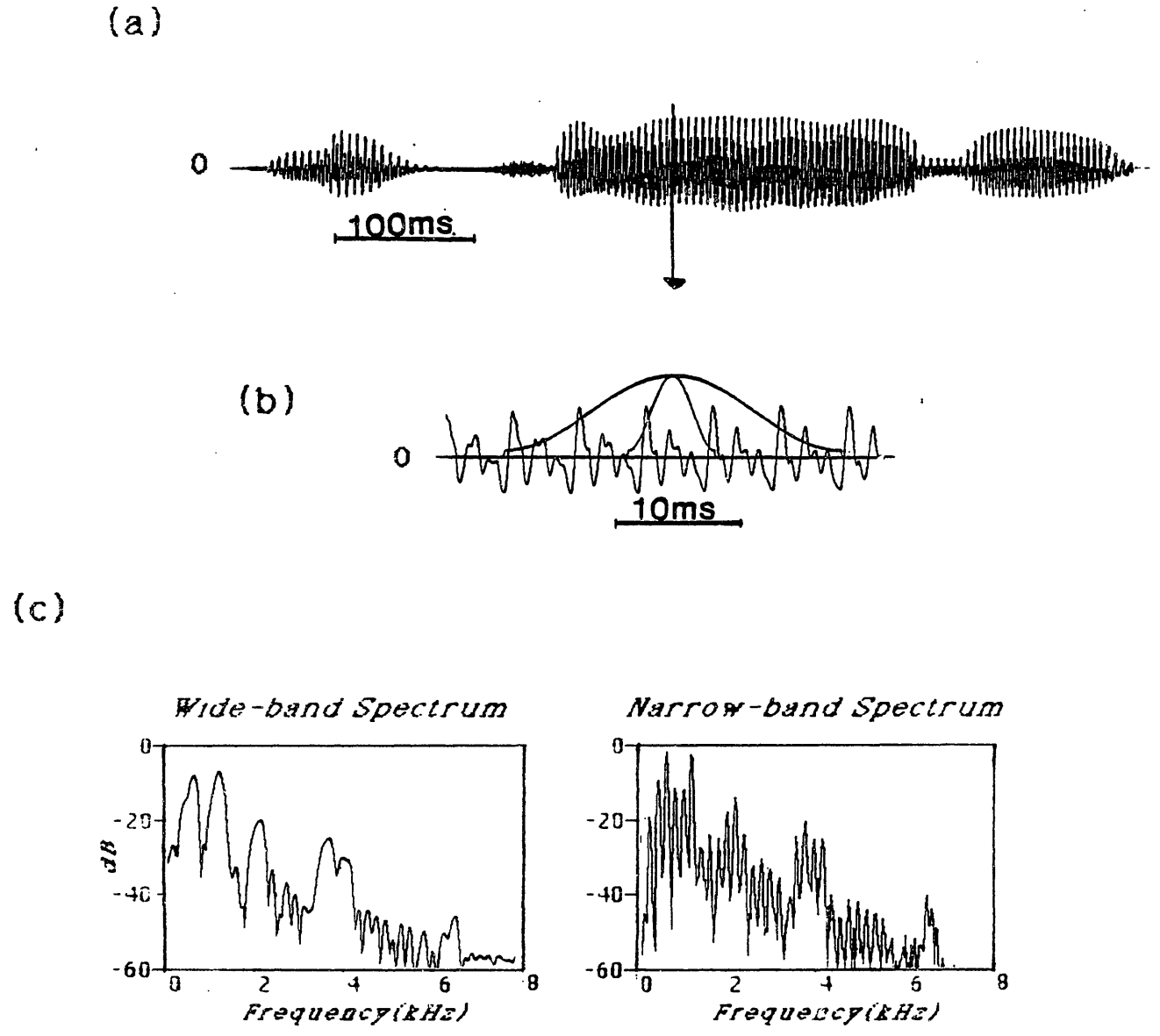

Figure 6.1: Example illustrating wide-band and narrow-band spectral analysis.
a) Original waveform for word "majority" spoken by a female speaker.
b) Time-expanded segment of waveform at the time-point indicated by the arrow, during the /ɔ/, with hamming windows appropriate for wide-band [short] and narrow-band analysis superimposed.
c) Resulting wide-band and narrow-band spectra for the segment of speech shown in part (b).

window, which encompasses several periods of the excitation, is typical for narrow-band analysis. Part (c) shows the resulting wide-band and narrow-band spectra, for the speech segment in part (b). The narrow-band spectrum shows clear harmonic structure, but nonetheless the formants are visible as peaks in the envelope of the spectrum. The wide-band spectrum appears in some sense to be more attractive, because each major peak corresponds to a formant. It would appear to be easier to extract the formant frequencies from this particular example of wide-band analysis than from the corresponding narrow-band spectrum.

Figure 6.2 illustrates a serious problem with wide-band analysis, which is that it can be extremely unstable over time. Part (a) of the Figure shows the wide-band spectrogram for the phrase "This is", spoken by a male speaker. The frequencies of the four first formants are quite clear in both vowels. Part (b) shows a small segment of the waveform at the time corresponding to the vertical bar in part (a). Part (c) shows wide-band spectral cross-sections centered at the two vertical bars indicated on the waveform in part (b). The spectrum centered on the peak in the waveform, corresponding to the closed phase of the glottal cycle, shows the formants quite clearly. However, the other spectrum, centered on the open phase of the cycle, shows no clear valleys below the frequency of the fourth formant. The human is able to ignore the inferior frames, and follow the lines of the formant frequencies over time even when the situation is much worse than is shown here. However, it is very difficult to conceive of a plan for training the computer to accomplish what the human visual system does so well. In theory, it should be possible to perform pitch-synchronous analysis, always selecting a placement of the window that is aligned with the closed phase of the glottal cycle. In practice, however, such an approach is extremely difficult to realize reliably, in part because of the strong dependence on an accurate estimate of the fundamental period of voicing. Instead, most researchers prefer to begin with a longer time window, and then apply a smoothing filter to the log spectrum to partially remove the harmonic structure.

The method of smoothing the log spectrum to obtain an estimate of the vocal tract filter shape was motivated theoretically by the need to deconvolve the vocal tract filter with the excitation function. The convolution of the excitation with the vocal tract response characteristic is converted to a product in the spectral domain. Taking the log of the spectrum effects a transformation of the product into a sum. An inverse transform of the log spectrum then yields a zero-phase time domain waveform, the "cepstrum", which, ideally, contains the information characterizing the vocal tract response mostly in the low time portion, and the information characterizing the excitation in the high time portion. By then applying a time window to the cepstrum, restricted to a narrow region [2 to 3 ms], it should be possible to remove most of the information related to the excitation, while retaining the relevant spectral shaping information. Such "low time liftering" can be accomplished in practice by applying a smoothing filter to the log spectrum, with a cutoff "quefrency" of around 2 ms. For a further discussion of the theoretical framework behind such "homomorphic" analysis, see Oppenheim and Schafer [1975], Chapter 10.

Another popular method for speech analysis is based on an all-pole model for the vocal tract filter, where poles in the model then correspond ideally to resonance frequencies of the vocal tract. This method, "Linear Prediction" [Makhoul, 1975; Markel, 1976], results in smooth log spectra

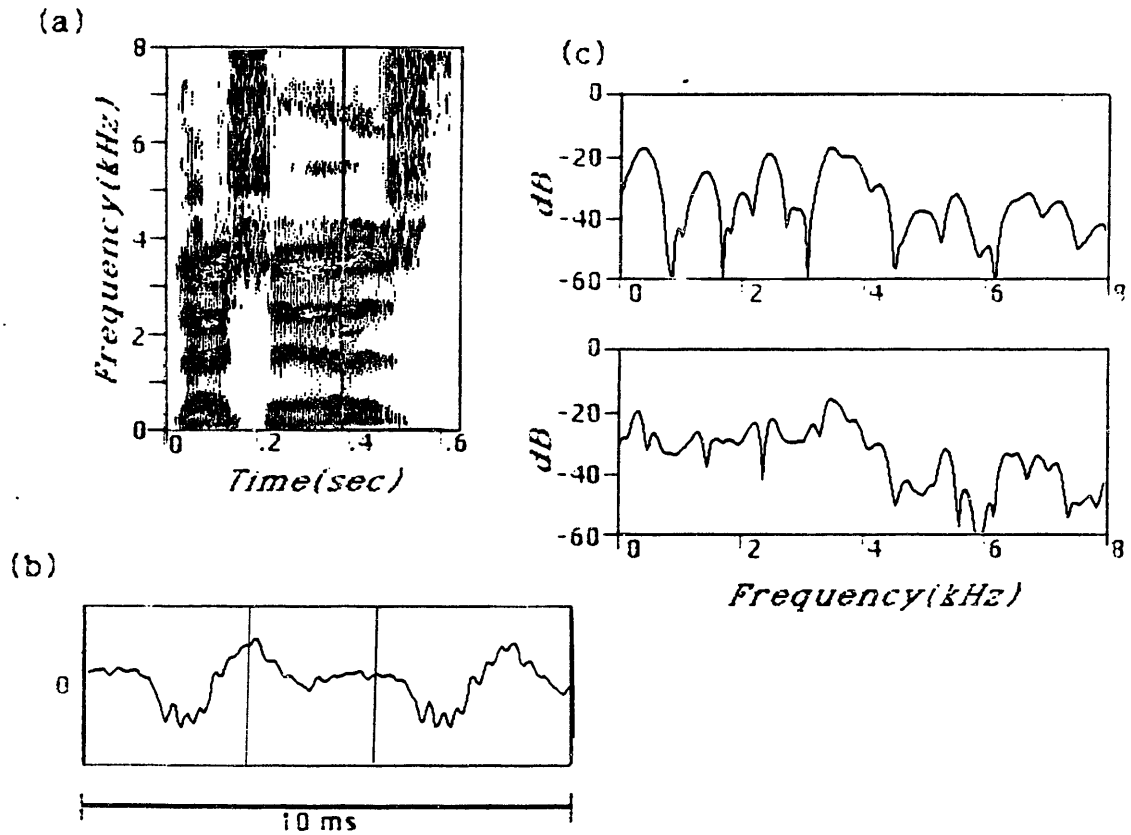

Figure 6.2: Example showing lack of stability over time in wide-band spectral cross-sections.
a) Wide-band spectrogram for phrase "This is", spoken by a male speaker.
b) Time-expanded segment of waveform at time-point marked by vertical bar in spectrogram in part (a).
c) Wide-band spectra computed at two points in time indicated by vertical bars in part (b). In spite of close proximity of two analysis windows, results are quite different. Window for upper cross-section is centered on vertical bar on the left.

that generally outline the shape of the Fourier log spectrum fairly accurately, with more emphasis on the peaks than on the valleys. The main advantage of Linear Prediction is that a model with P poles is guaranteed to have no more than P/2 peaks in the spectrum. Thus it becomes feasible to consider tracking the formant frequencies, asserting a one-to-one correspondence between the first N [where N is usually three or four] underlying formant frequencies and N specified peaks in the LP spectrum.

Figure 6.3 shows examples of wide-band, narrow-band, cepstral, and LP analysis of the vowel /ɔ/ in "ordered", spoken by a male speaker, to illustrate some of the points discussed above. The wide-band analysis was performed with a 7 ms hamming window, and the other three methods were all applied to a 25 ms hamming windowed segment of the speech. The number of poles, P, for the LP analysis was 19. The smoothing "lifter" for the log spectrum had a response characteristic that was flat to 2 ms, followed by a taper to the stop band which began at 4 ms.

The first five formants are represented by the first five peaks in the LP spectrum. The smoothed log spectrum has an additional low peak between F_3 and F_4, but the first three formants are aligned with the first three peaks. The cepstral analysis preserves the detailed shape of the envelope spectrum, including the deep valley at about 5 kHz. Linear Prediction, in contrast, because it is a model, is unable to reproduce spectral details intact. This limitation is often viewed however as a feature, because it is difficult to devise a method for extracting the formant frequencies when extra "spurious" low amplitude peaks are prevalent.

6.3 Application to Speech Recognition

Once a representation of the log spectrum has been obtained, the next step is to use the representation in computer speech recognition. There are several alternative methods for capturing the relevant features necessary for phonetic identification, each of which has certain liabilities and assets. It is customary to distinguish between methods that approach the problem through "template matching" and methods that are more concerned with "feature extraction". With template matching, the time sequence of log spectra describing an unknown word or phrase are compared against a set of stored template sequences representing the set of words or phrases in the predefined vocabulary. The comparison is usually made through some distance metric, such as the Itakura metric [Itakura, 1975] or a euclidean distance, with some sort of dynamic time warping procedure used to find the best temporal alignment of the matching sequence. Feature-extraction approaches are based on a knowledge of the important acoustic attributes for each phonetic distinction. A specified set of features would be called upon at any given instance in time, depending on the proposed hypotheses available at that time point.

Template-matching strategies are strongly speaker-dependent; attempts to achieve speaker independence through formations of cluster groupings have met with limited success. This limitation is in part related to the fact that the absolute amplitudes of the formants vary greatly from speaker to speaker. Listeners are much more sensitive to changes in formant frequencies than in their amplitudes. The amplitude of the first formant peak is implicated in the nasal/nonnasal percept,

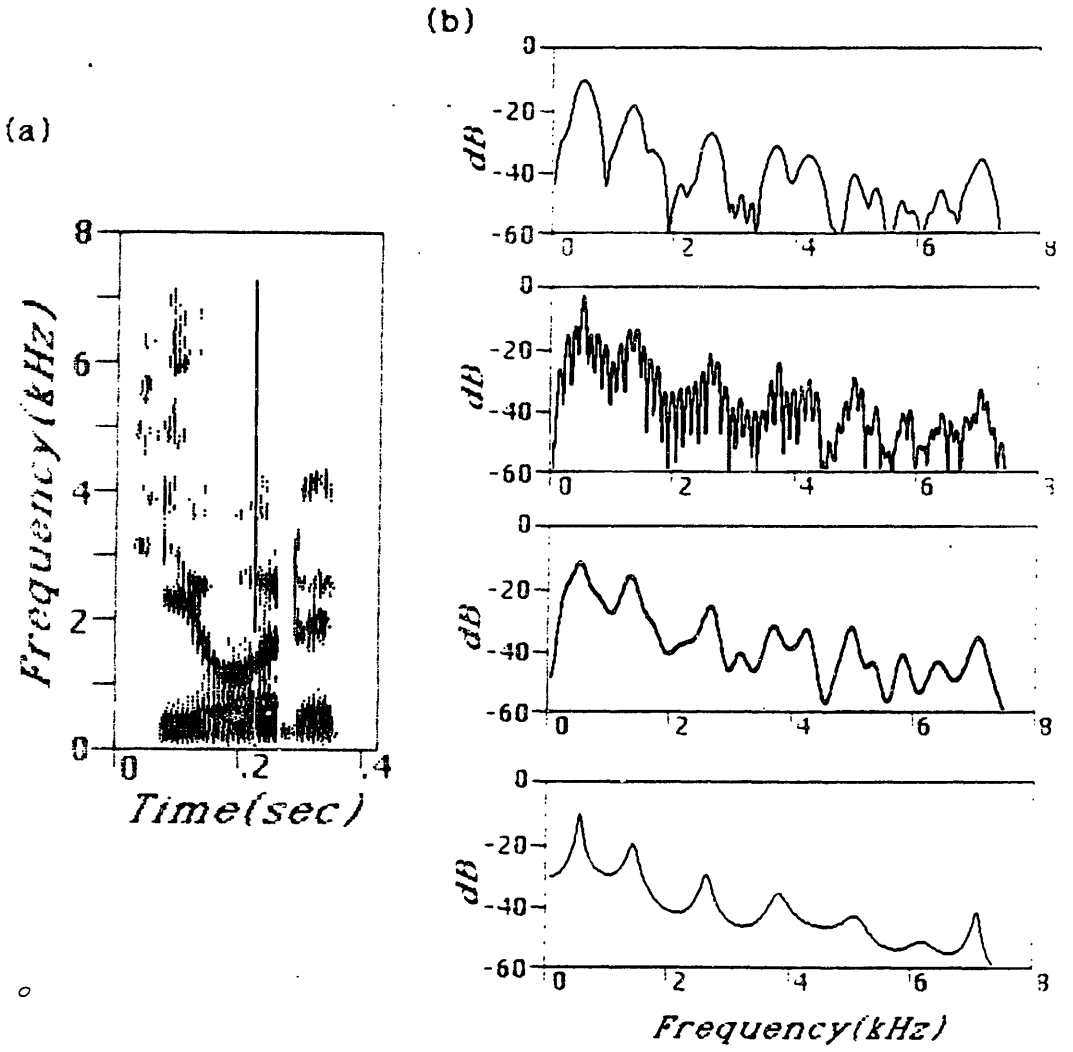

Figure 6.3: Example illustrating several processing methods for spectral analysis.
a) wide-band spectrogram for phrase "He ordered", spoken by a male speaker.
b) Spectral cross-sections taken at vertical bar in part (a), during the /ɔ/.
i) Wide-band analysis, ii) Narrow-band analysis, iii) Homomorphic analysis, iv) Linear Prediction analysis

but, except for this important discrimination, formant amplitudes are relatively unimportant for phonetic identity.

A feature-extraction approach to speech recognition might be able to overcome to a certain degree the apparent speaker-dependencies. Most of the significant features for sonorant regions can be derived from a knowledge of the frequencies of the first three or four formants. An error-free formant tracker would therefore be a wonderful tool for subsequent recognition. Not only are the formant positions in the middle of steady-state regions important, but also directions of formant movements near phoneme boundaries are often essential for the identification of adjacent consonants.

The problems in formant tracking reduce generally to two categories: the presence of spurious peaks in the spectrum and the merger of two formants into a single peak. Sometimes a peak representing a formant can be very weak, making it difficult to distinguish between true formants that happen to be weak and peaks that are known to be extraneous, given a knowledge of the appropriate formants for the underlying speech sound. Nasalization poses a particularly difficult problem, because it is often manifested by a double-peaked first formant, with the underlying formant frequency being somewhere between the two peaks. An example of an algorithm for extracting the frequencies of the first three formants from LP spectra, as well as a discussion of some of the above-mentioned problems, can be found in McCandless [1974].

Figures 6.4 through 6.7 show several examples that were selected to illustrate many of the points mentioned above. In each case, a wide-band spectrogram is given, as well as three log spectra computed using narrow-band, cepstral, and LP analysis, respectively. The log spectra are all computed at the time point indicated by the vertical bar on the spectrogram.

Figure 6.4 shows two examples where one or more of the formants are disproportionately weak. Part (a) of the Figure shows the phone /I/ in the environment of a preceding and following /z/. The second and third formants are invisible on the spectrogram, and are manifested as extremely low peaks in the LP spectrum. The narrow-band spectrum shows that there are actually present two distinct harmonics, representing the second and third formants, that are well above the surrounding harmonics in amplitude. A measure of local prominence might have been able to produce a more spectrally balanced representation of the vowel in this case. Part (b) of the figure shows a similar situation for a section of the /a^{w}/ in the word "found". Here the pole at the second formant in the LP representation is manifested as an inflection point on the upper edge of the first formant peak. The peak could perhaps be recovered by taking the second derivative with respect to ω. The cepstral analysis retains the peak at F_2, but its amplitude is weak enough that it could be confused with the many spurious peaks that are present in this analysis method.

Examples of problems with a nasalized first formant are given in Figure 6.5. Part (a) shows the wide-band spectrogram of the word "pound", spoken by a female speaker, and cross sections taken near the end of the /a^{w}/. Here the nasalization in the first formant region is compounded by the fact that the harmonics are well separated in frequency due to the high pitch of the voice. The cepstral analysis retains some harmonic structure, and thus extracts several peaks in the first and second formant region. LP analysis detects two peaks representing F_1, followed closely by a more

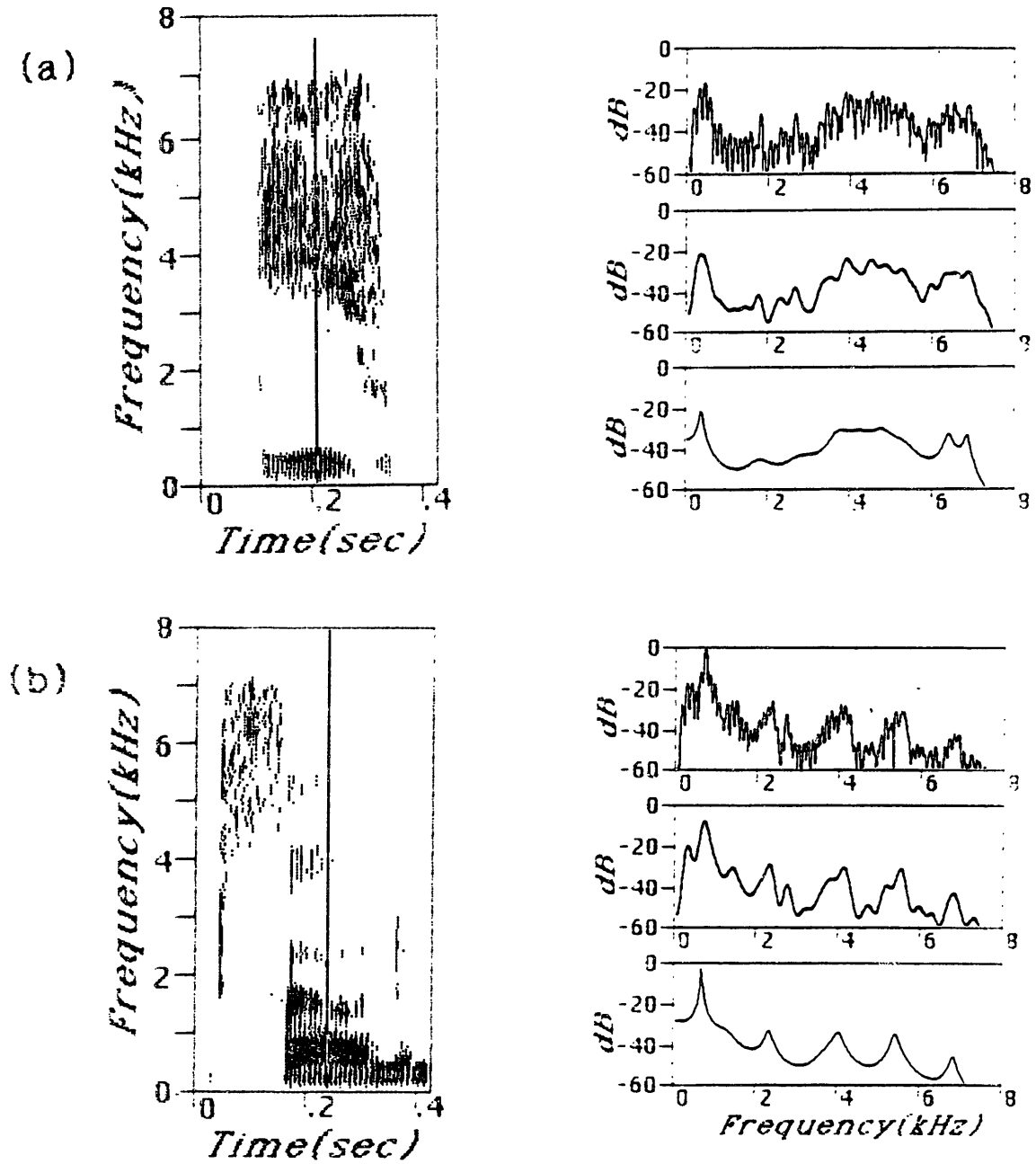

Figure 6.4: Examples illustrating disproportionately weak formants.
a) [left] Wide-band spectrogram during phone sequence /z/ /I/ /z/, spoken by a male speaker. [right] Three spectral cross-sections taken at time point indicated by vertical bar in spectrogram, during the /I/. [top] Narrow-band analysis, [middle] Homomorphic analysis, [bottom] Linear Prediction analysis.
b) Same as in part (a), for the word "found" spoken by a male speaker. Cross-sections are taken during the /a^w/.

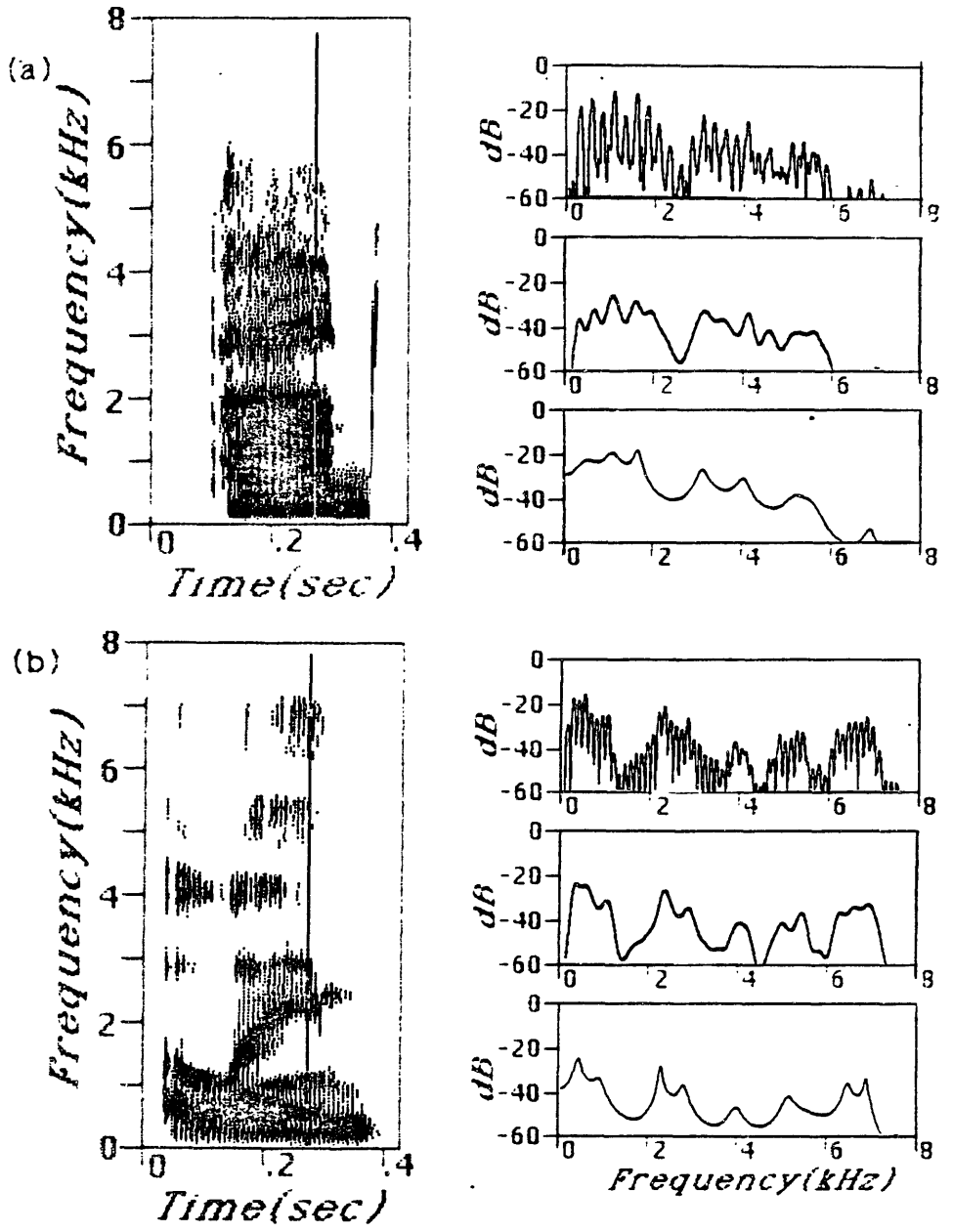

Figure 6.5: Examples illustrating effects of nasalization on vowel spectra.
a) [left] Wide-band spectrogram of the word "pound", spoken by a female speaker.
[right] Three spectral cross-sections taken at time point indicated by vertical bar in spectrogram, near the end of the vowel. [top] Narrow-band analysis, [middle] Homomorphic analysis, [bottom] Linear Prediction analysis.
b) Same as in part (a), for the word "lame" spoken by a male speaker. Cross-sections are taken during the /e/.

prominent peak at F_2. A similar example is given in Figure 6.5(b). The word is "lame", spoken by a male speaker, and both cepstral and LP analysis represent the first formant region by two peaks.

Another problem is formant mergers that occur when two formants are very close in frequency. Sounds which have a narrow spacing of F_1 and F_2 are /a/ as in "box" and /ɔ/ as in "caught". F_2 and F_3 are close together for /ɜ^/ as in "bird", /r/, and /y/. Two examples are given in Figure 6.6. Part (a) of the Figure shows the word "rare", with the cross-sections taken near the middle of the vowel. The /r/ color has caused the third formant to remain close to the second throughout. Indeed, the LP spectral analysis rarely produced two peaks over the entire duration of this vowel, for the energy concentration representing the second and third formants. However, at least for the frame shown in cross-section, there is an underlying pole in the model accounting for the third formant. Spectral enhancement techniques might be able to recover a separate peak [McCandless, 1974]. The cepstral analysis has retained separate peaks in the illustrated cross-section, as well as in general, for this vowel. Part (b) of the Figure shows a similar case for the inserted /y/ in the phrase "He ordered". There is evidence in both the cepstral and LP spectra for the presence of two formants, but it is not easy to reliably extract this evidence.

There is often a trade-off associated with the number of poles that are used to represent the LP model. Too many poles lead to spurious peaks that complicate the peak-picking strategy; whereas too few result in formants that are left out of the model. An example where LP analysis produced a surprisingly poor representation of the important low frequency region of the spectrum is given in Figure 6.7. The cross-sections are taken in the /aɪ/ of "tried", spoken by a female speaker. While the cepstral analysis has produced three peaks corresponding to the first three formants, the LP analysis has represented F_2 and F_3 as a broadly sloping upper edge on the first formant. This speaker happens to generate a significant amount of high frequency energy during the voiced regions, the details of which are consuming too many of the poles of the LP model.

The above cases have been selected because of their pathological nature. We do not mean to imply that these are typical of the performance of the LP model. In fact, formant tracking from LP spectra can probably yield a better than 90% success rate. Unfortunately, when an error is made in formant tracking the results are often devastating, because the estimates are off by a gross amount. To accurately abstract directions and rates of movement of formants depends upon the accurate tracking over a sequence of frames, thus compounding the effects of errors.

We thus arrive at an apparent dilemma. If we could somehow balance out the spectral shape so that the formant amplitudes did not vary so widely from speaker to speaker, then a template-matching approach would be much more attractive, because it would not be nearly so dependent on the speaker and the environment. If, on the other hand, we could extract the frequencies of the formants with an accuracy approaching 100%, then the variabilities in amplitude would no longer be a problem. Yet these amplitude variabilities contribute substantially to the lack of success in formant tracking. Furthermore, there are situations when two formants are so close together that it seems almost impossible to resolve them. These are some of the major unsolved problems that confront the researcher interested in computer speech recognition. It is our belief some answers

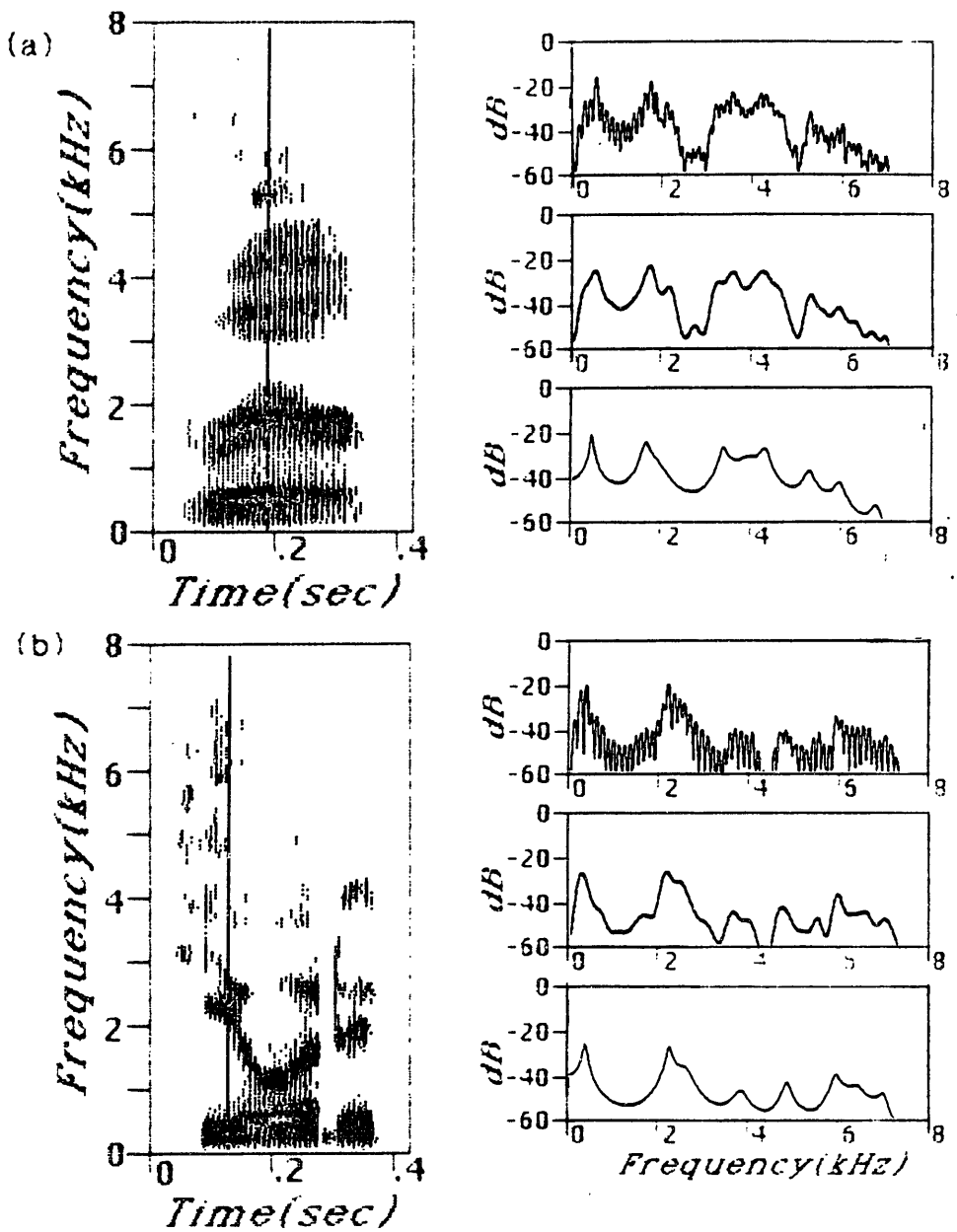

Figure 6.6: Examples illustrating problems with formant mergers.
a) [left] Wide-band spectrogram of the word "rare", spoken by a male speaker.
[right] Three spectral cross-sections taken at time point indicated by vertical bar in spectrogram, during the /ɛ/. [top] Narrow-band analysis, [middle] Homomorphic analysis, [bottom] Linear Prediction analysis.
b) Same as in part (a), for the phrase "He ordered," spoken by a male speaker. Cross-sections are taken during the /y/ inserted between the two vowels.

may be found in a method for spectral analysis that departs considerably from the linear domain. The design goal should not be defined in terms of obtaining a spectral representation that matches the log spectrum, but should instead be developed from strategies that make use of our available knowledge about how the auditory system processes speech.

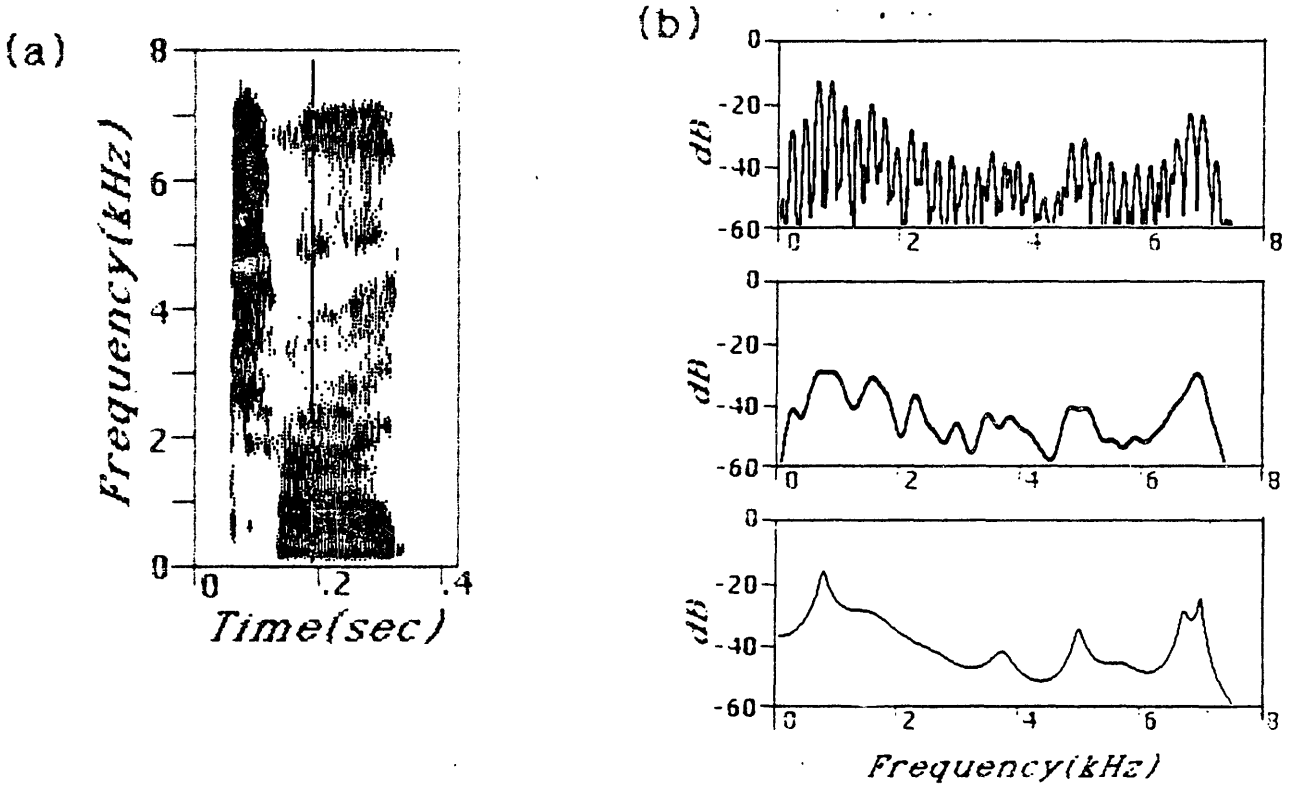

Figure 6.7: Examples illustrating breakdown of LP analysis due to inadequate number of poles in model.
a) Wide-band spectrogram of the word "tried", spoken by a female speaker.
b) Three spectral cross-sections taken at time point indicated by vertical bar in spectrogram, during the /a/ portion of /aʸ/. [top] Narrow-band analysis, [middle] Homomorphic analysis, [bottom] Linear Prediction analysis.

Chapter 7

General Description of Thesis System

Sachs and Young's work on auditory processing of speech-like stimuli seems to indicate that a simple rate response scheme for speech spectral analysis is insufficient for the task of accurately detecting spectral prominences associated with formant resonances. As has been discussed in previous chapters of this thesis, several researchers, including Sachs and Young, have suggested some further processing of the peripheral outputs in order to sharpen the peaks. Such filtering is beneficial because the peripheral outputs respond in synchrony to the input stimulus, thus preserving additional frequency information in the detailed waveshape.

A major portion of this chapter is concerned with the development of a set of arguments advanced to support the proposed synchrony measure that was used in the thesis to detect prominent periodicities in the outputs of the peripheral model. This "Generalized Synchrony Detector" [GSD] is used both to obtain an improved spectral image over the one obtained from an estimate of overall energy out of each channel [rate response scheme] and, under slightly different circumstances, to determine a fundamental period of voicing for the input speech signal.

In addition to the discussion of the synchrony measure, the entire system will also be described at a somewhat superficial level, in order to set the stage for the detailed discussion of the system given in Chapter 8.

7.1 Overview

A block diagram of the system is given in Figure 7.1. The initial stage processing is intended to correspond to peripheral auditory processing. There is available, as has been indicated in the review chapters, an extensive amount of measured data to guide the design of this part of the system. The outputs of the peripheral level processor are delivered to both a spectral estimator and a pitch estimator. In the design of these estimators, very little attention was given to known auditory processing. However, the design was deliberately limited to structures that allowed for a simple interpretation in terms of neurological analogs. Thus, for example, complex mathematical procedures such as maximum-likelihood analysis were avoided. Furthermore, an attempt was made to unify the approaches for spectral and pitch estimation; thus, there is considerable parallelism in the procedures for these two separate tasks. The underlying assumption is that a small number of nerve cell types could function in multiple tasks, with minor modifications as, for example, in time constants.

The spectral analysis was restricted to the region between 200 and 2700 Hz. Thus, the system is especially suitable for use in sonorant segments of speech, such as vowels, nasals, and liquids. It may also be useful for certain compact consonants such as the velars, which are characterized

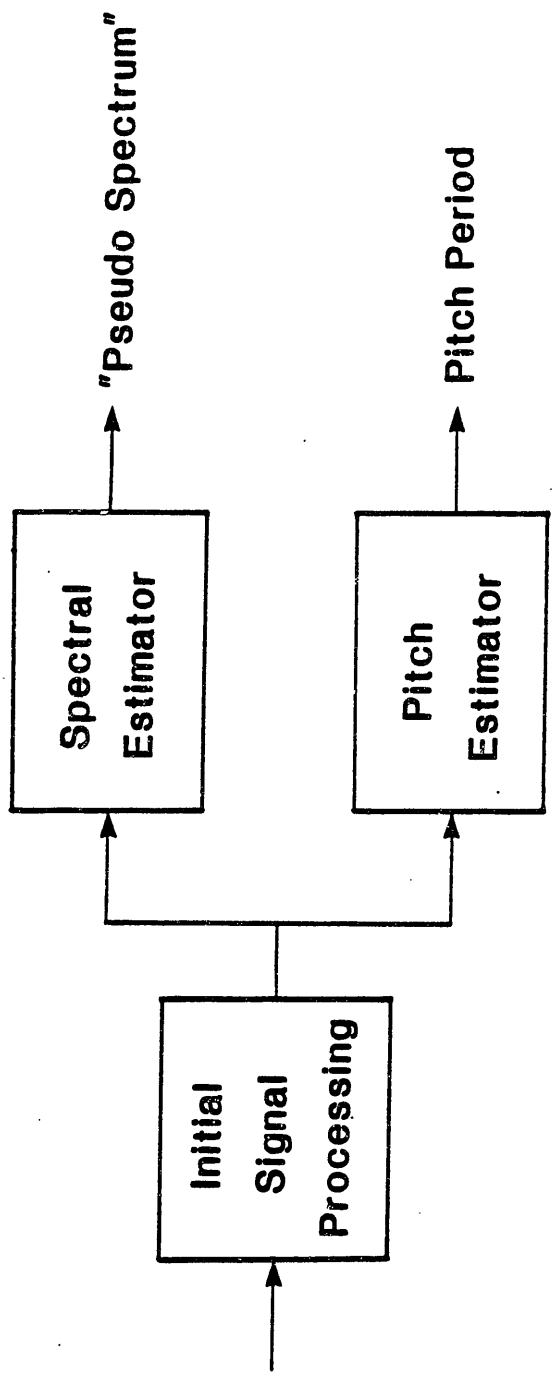

Figure 7.1: Block diagram of overall system implemented for thesis.

by the presence of a promir ; peak in the low frequency region. In the peripheral model, total synchrony in the response ; ern was assumed. Because of the 2700 Hz upper bound in frequency, this assumption is not unr >nable, if we can take the response of cats' ears as representative. The peripheral model includes a bank of linear bandpass filters with approximately critical bandwidths, a dynamic amplitude compression scheme, and a nonlinear half-wave rectifier.

The outputs of the peripheral model are never reduced to a spike sequence, and therefore can be viewed as representing a probability of firing of the nerve fiber. For the design of the half-wave rectifier, the shapes of peaks in period histograms were used as a reference. Although period histograms are obtained by combining the outputs of a single nerve fiber over a long time interval, a similar result could be obtained in principle by averaging together the simultaneous outputs of several adjacent fibers with the same frequency response characteristics. It has been shown that the responses of neighboring fibers are statistically independent [Johnson, 1970;Johnson and Kiang, 1976], and therefore combining the responses of two fibers over the same cycle of the signal would be very similar to combining the responses of a single fiber over two cycles in sequence.

In the design of the spectral estimator, the assumption is made that mean rate response is an insufficient measure for the speech spectrum in terms of preserving formant information. A mechanism for detecting the extent of phase-locking to the center frequency of the fiber in the peripheral-level output is developed, and it is demonstrated that this mechanism produces a spectral-like image with peaks in the spectrum corresponding to the formant frequencies. The same mechanism is also applied to detecting periodicities at the fundamental period of voicing, but the input in this case is a single waveform generated by combining the outputs of the peripheral model. To estimate the pitch period, the phase-locked response to each of the periods appropriate for pitch is examined for a best match.

7.2 Generation of Pitch Waveform

The "pitch waveform" is generated simply by adding the weighted outputs of all of the filters across the spectral or place dimension. In terms of a possible neurophysiological model, such a procedure follows naturally as an extension of the proposed local averaging to obtain an adequate statistical sampling at a given place. The summing process preserves the periodicities at the fundamental for pure tones and for harmonic sequences that are contained in the region below 2700 Hz. The dynamic range compression that is realized in the peripheral model results in an effective spectral flattening in the sum waveform. Thus periodicities at the formant frequencies tend to be reduced relative to periodicities at the fundamental.

The filters that are summed are spaced by half a Bark, i.e., half a critical bandwidth. As such there is significant overlap between adjacent filters. Differing phase response characteristics at the same frequency for adjacent filters could cause some complex cancellations of energy. However, the detailed shape of the spectral envelope is of no interest to the task of estimating the fundamental frequency. An attempt was made to delay the various filter outputs so as to remove the differing linear phase components of the filters, but other than that phase differences were ignored. The

half-wave rectification process introduces multiple higher harmonics, but these should all reinforce the periodicity at the fundamental period. It is impossible to control the phase relationships among such higher harmonics and the same frequency component present in the original signal and filtered through a higher CF filter. Surprisingly, in spite of all these distortions and phase differences, the pitch waveform often bears a strong resemblance to the original waveform.

The restriction to signals below 2700 Hz is an artificial consequence of the limitations in the computer simulation. In the auditory system, fibers tuned to frequencies above 2700 Hz continue to maintain synchrony to the envelope period even after they have lost synchrony to any of the individual harmonics responsible for the fundamental periodicity. Thus these higher frequency filter outputs could also be included in the sum waveform.

One of the arguments advanced against temporal processing for pitch is the apparent difficulty in maintaining accurate delays at the long delay intervals necessary for pitch estimation [Whitfield, 1970]. Such a problem disappears if there is a single waveform from which to extract the pitch period, rather than a set of waveforms whose estimates must be combined. The waveform, processed through a single tapped delay line, would be compared at multiple taps with the undelayed waveform to detect an alignment that registers a prominent periodicity. The tap that obtained a good match would correspond to a delay that was with certainty longer than the delay for any of the preceding taps, and shorter than the delay for the subsequent taps. Thus relative pitch is well defined, although absolute pitch would remain difficult to detect.

The notion of adding up the individual filter outputs to generate a pitch waveform is not completely new. Searle [1980] proposed a summing of the **envelope** responses of the high frequency filters to generate a single waveform for pitch analysis. Since his simulation system only generated the envelope of the response for each filter, the low frequency filter outputs had very little information relevant to pitch preserved in the time domain, so these were not included in the sum.

7.3 Synchrony Measures for Formant Enhancement

The spectral flattening process that comes about as a consequence of dynamic range compression in the peripheral model is a welcomed aid to the pitch detection task. However, this same process serves as a hindrance to spectral estimation, because the formant peaks become very diffuse and broad. A possible solution to this apparent dilemma is to assume that the amplitude information is far less important for spectral processing than the additional information available in the detailed wave shape of the period histogram or probabilistic output. Sachs and Young [1980] and Srulovicz and Goldstein [1983] essentially proposed that the peripheral stage outputs be passed through a second, more sharply tuned, filter, also at CF, in order to sharpen formant peaks. As has been discussed already, such a second filter tends to pick up a strong response at the place of the second harmonic of the first formant. In addition, it is likely that the neural mechanism responsible for the second filter would also include dynamic range compression, or, at least, saturation phenomena, which would lead to problems very similar to those encountered at the peripheral level.

There are potentially a number of different time domain mechanisms for comparing the wave-

form with a delayed version of itself, which offer certain advantages over a simple second filter strategy. Two obvious methods, which we will discuss here, are autocorrelation and Average Magnitude Difference Function (AMDF) [Moorer, 1974]. If we consider again Figure 2.9 from Delgutte [1980], it is evident that, for 77dB SPL,

1. The mean rate response for the fiber at 800 Hz CF, tuned to the formant frequency, is, if anything **lower** than the mean rate response for the fiber tuned to 2.79kHz, and

2. The response in both fibers shows a strong periodicity with the 800 Hz formant frequency; the 800 Hz fiber response histogram is almost perfectly periodic with this frequency.

A possible procedure for detecting the 800 Hz formant in the Delgutte example would be to autocorrelate the period histogram derived from each nerve fiber using a specific delay τ_{CF}, equal to the inverse of the center frequency of the filter:

$$R_\tau = \sum_{k=n-N/2}^{n+N/2} |x[k]\, x[k-M]|\, w[k+\frac{N}{2}-n]$$

where $x[n]$ is the input signal, $w[n]$ is a window function, M is the delay in samples equal to τ in ms, and N is the window size in samples. Such an autocorrelation could be normalized with respect to the RMS energy, R_0. Thus, amplitude information would be removed altogether. The fiber tuned to the formant frequency would obtain a response very close to the 1.0 theoretical maximum, whereas the high frequency fiber output is asynchronous with the high frequency CF, and therefore its response would be low.

An appealing aspect of the autocorrelation is that it does not pick up a false response at twice the first formant frequency. Consider, for simplicity, a pure tone at 800 Hz frequency and high signal level. Fibers tuned to 1600 Hz will pick up a significant response, because the auditory filters tend to have tails on the low frequency side. The half-wave rectification process then introduces a response at the 1600 Hz frequency, which would be passed through the second filter. However, the 1600 Hz component is phase synchronous with the 800 Hz fundamental. An autocorrelation of the waveform with the period of the 1600 Hz component will yield a very weak response because the near-zero outputs of the negative half of the fundamental cycle will be multiplied by the high level outputs of the positive half of the cycle.

Interestingly, the autocorrelation would produce an erroneous strong response at the place corresponding to **half** the frequency of the fundamental. Any waveform that is perfectly periodic with period τ is also perfectly periodic with period 2τ. However, because of the very steep slope on the high frequency side of the critical band filters, it is unlikely that a response above spontaneous would be obtained for filters below the frequency of the stimulus.

Such half-frequency periodicity would remain a problem, however, if amplitude information were ignored altogether. Therefore, an equation such as the following is suggested:

$$S_1(\tau) = \frac{R_\tau}{R_0 + K} \tag{7.1}$$

where $S_1(\tau)$ is the synchrony measure, R_τ is the autocorrelation of the peripheral level probabilistic output, suitably windowed, at period τ, the center period of the peripheral filter, and K is a constant which should be large compared to the energy in the spontaneous response.

R_τ can be interpreted in the frequency domain by making use of the Fourier transform pair relationship between autocorrelation and magnitude squared spectrum. An autocorrelation at a period τ, R_τ, is equivalent to summing the cosine-weighted magnitude-squared spectral coefficients of the windowed signal, where the weighting function is of the form $cos2\pi f\tau$ [see Figure 7.2a]. Frequencies in the input at half the correlation frequency get weighted by a negative factor, -1. The strongest positive weight, $+1$, is obtained for frequencies at multiples of the correlation frequency.

A preliminary version of the system used autocorrelation in order to detect periodicities. The problem with autocorrelation is that the width of the lobe of the cosine peaking at the correlation frequency is too broad. Since the input has already been bandpass filtered by the critical band filter, the correlation picks up most of the energy in most signals that are passed by the initial stage. Thus there is very little sharpening of the initial peak.

There are other alternatives to correlations that involve the principle of comparing the waveform with a delayed version of itself. One function that has been applied successfully to the detection of periodicities with the pitch period [Moorer, 1974], is the "Average Magnitude Difference Function" (AMDF), which can be defined as follows:

$$AMDF_M[n] = \sum_{k=n-N/2}^{n+N/2} |x[k] - x[k-M]|\, w[k + \frac{N}{2} - n]$$

where $x[n]$ is the input, $w[n]$ is the window function, M is the delay in samples, and N is the window size in samples. The AMDF can be interpreted as an estimate of the envelope of a waveform, $y[n]$, generated by subtracting the input waveform, $x[n]$, delayed by M, from the input waveform.

$$y[n] = x[n] - x[n-M]$$

Thus $y[n]$ is the result of processing $x[n]$ through a comb filter with a series of zeros equally spaced around the unit circle at multiples of the radian frequency, $\omega_0 = 2\pi/M$.

If the AMDF function were defined as a magnitude **squared** rather than a magnitude, then it could be related to the energy in $y[n]$, or, equivalently, in $Y(e^{j\omega})$, invoking Parseval's rule. Thus an exact analogy to the analysis for the autocorrelation would hold, except that the spectral weighting function is now $2 - 2cos2\pi f\tau_M$, where τ_M is the time delay corresponding to M samples. Thus broad **nulls**, instead of peaks, occur at multiples of the frequency corresponding to the delay period, as illustrated in Figure 7.2b.

Because the AMDF computes a magnitude rather than a magnitude squared, it cannot be analyzed using procedures such as those discussed above. However, it seems experimentally that the magnitude function produces sharper nulls than the magnitude squared function. Amplitude normalization can be imposed naturally by dividing the AMDF by a similar function, in which

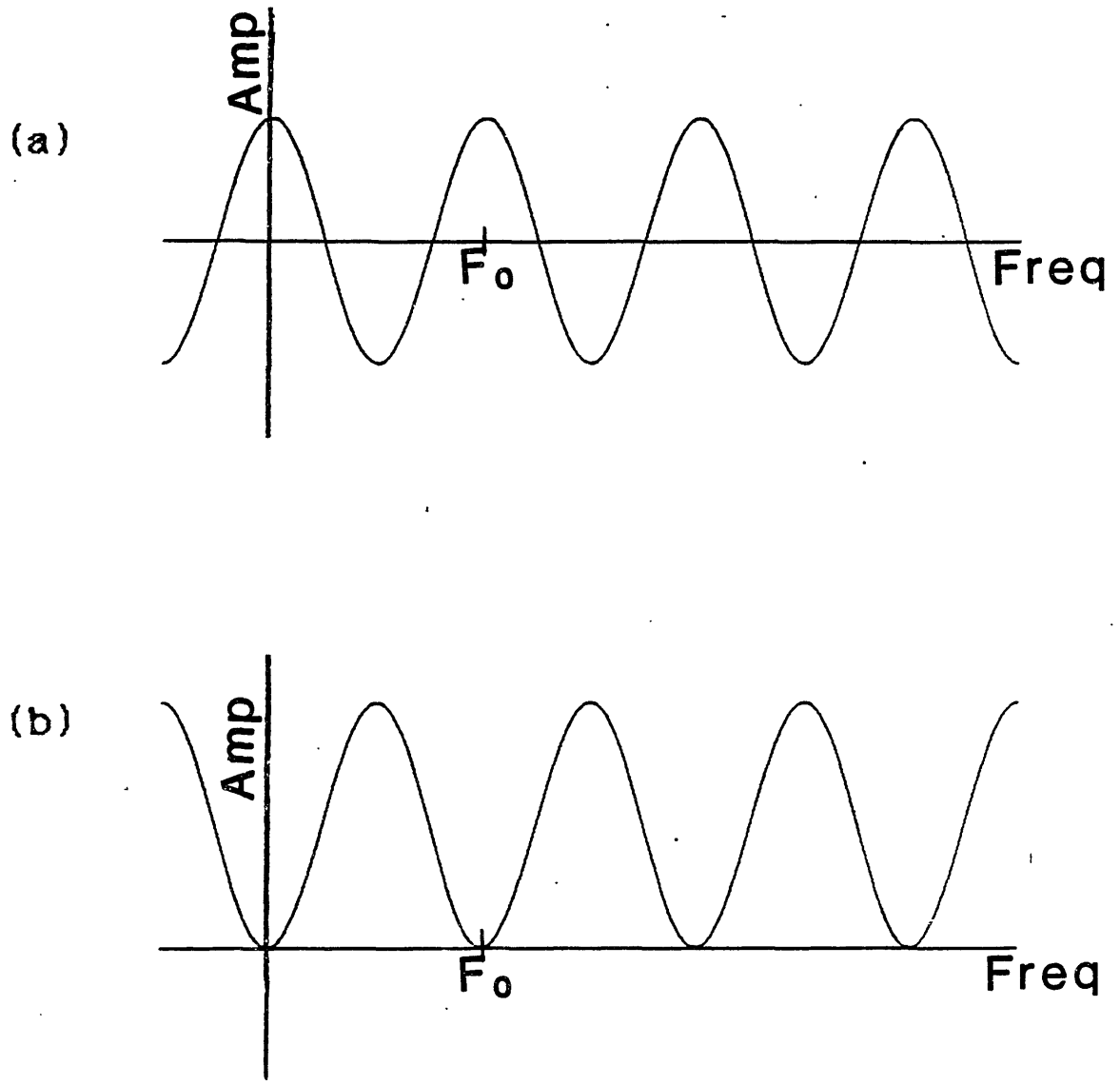

Figure 7.2:
a) Spectral weighting function corresponding to autocorrelation at period $\tau = 2\pi/F_0$.
b) Magnitude squared response characteristics for filter with N equally spaced zeros around unit circle at spacing F_0.

the subtraction is replaced by addition. Since the input waveform has been half-wave rectified and hence is always positive, the waveform sample and the delayed waveform sample would always add constructively.

The AMDF function produces nulls rather than peaks at the delays corresponding to strong periodicities in the waveform. Therefore, spectral representations using AMDF algorithms would appear to be upside down. A solution is to invert the AMDF, putting the sum term in the numerator and the difference term in the denominator. Of course there is the possibility of a zero in the denominator; this problem can be corrected by adding a small constant to the AMDF measurement.

Although it cannot be proven mathematically that the inverted AMDF function, as described above, produces sharper spectral peaks than the autocorrelation, it is possible to demonstrate experimentally that such is the case for simple inputs. To this end, an experiment was run on the computer using as input a swept sine wave changing in frequency from 300 to 1800 Hz. This swept sine wave was then half-wave rectified to produce a crude approximation to auditory peripheral outputs. The rectified output was processed through the autocorrelator at a fixed 1 ms delay [1000 Hz frequency] and an inverted AMDF at the same 1 ms delay.

Figure 7.3 shows the results. The autocorrelation output as a function of stimulus frequency closely resembles in shape a cosine function, as might be expected. The inverted AMDF, on the other hand, shows a very sharp peak near the 1000 Hz frequency, corresponding to the delay of the system.

A similar comparison is made in Figure 7.4 for a slightly more complicated signal. In this case, the input is a pulse train at 100 Hz frequency, filtered through a single resonator [complex pole pair] whose frequency is varied from 300 to 1800 Hz. As before, the signal is half-wave rectified prior to processing through the synchrony detectors. Again, the peak at 1000 Hz for the inverted AMDF is much sharper than the peak for the autocorrelator. Note also that for both signals, both the AMDF and the autocorrelator show a zero response at 500 Hz. This zero response is obtained in spite of the fact that both half-wave rectified signals contain significant energy at the 1000 Hz frequency of the detectors. That is to say, neither the AMDF nor the correlator pick up a response to the second harmonic frequency of the input.

The Generalized Synchrony Detector (GSD) that was developed for the purpose of detecting specific periodicities in the outputs of the peripheral stage processing, is a derivative of the AMDF function. It can be defined as follows:

$$S_2[r] = \text{soft-limit}\left[\frac{< x[n]+x[n-M] > -\delta}{< x[n]-x[n-M] >}\right] \qquad (7.2)$$

where $x[n]$ is the input signal, M is the delay in samples equal to the center period of the peripheral level filter, <> represents "envelope of", and δ is a silence threshold which depends upon the spontaneous response that would be obtained in the absence of a signal. Subtracting a δ from the numerator is a more precisely controlled mechanism for dealing with weak responses than adding a constant to the denominator, as was proposed for $S_1[r]$ [equation 7.1]. This feature turns out

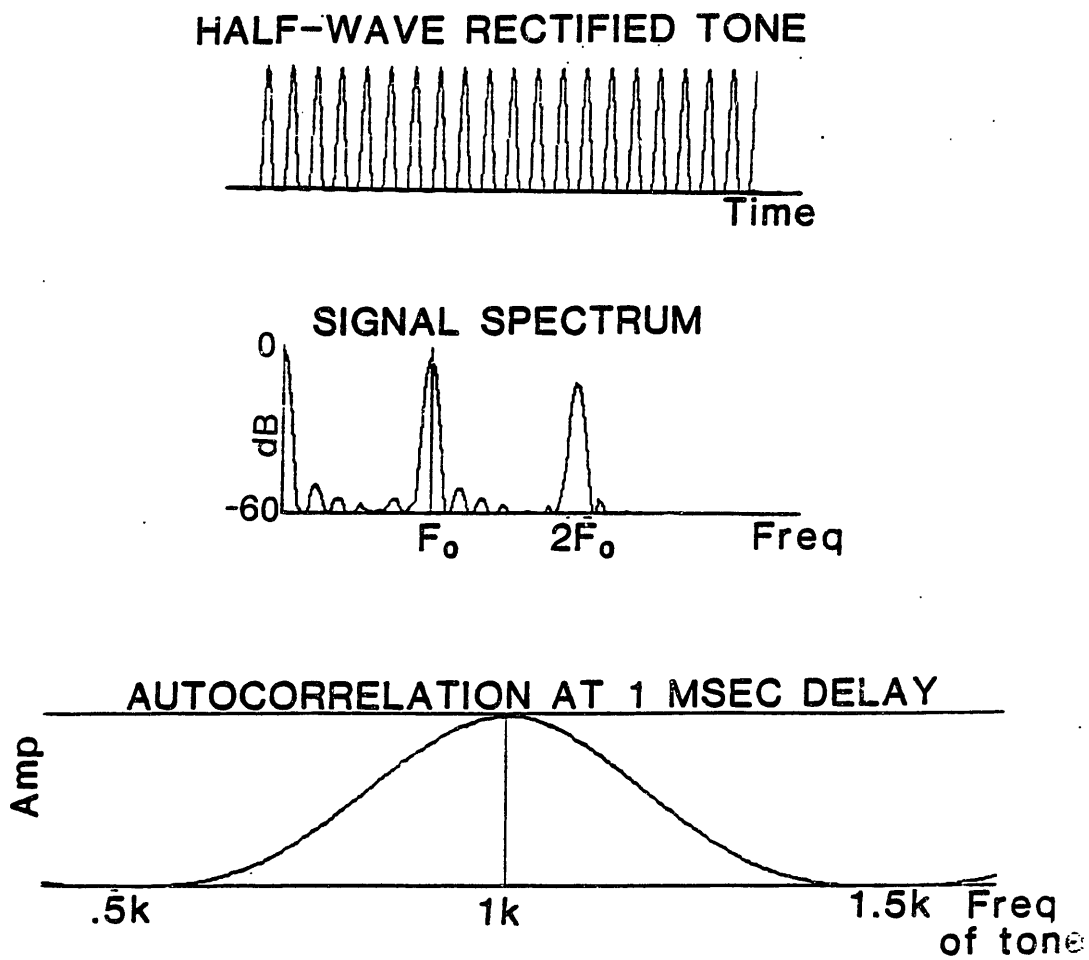

Figure 7.3: Response of autocorrelation function and inverse AMDF, both at fixed 1 ms delay, to an input signal consisting of a piecewise linear half-wave rectified sine tone, as a function of tone frequency.

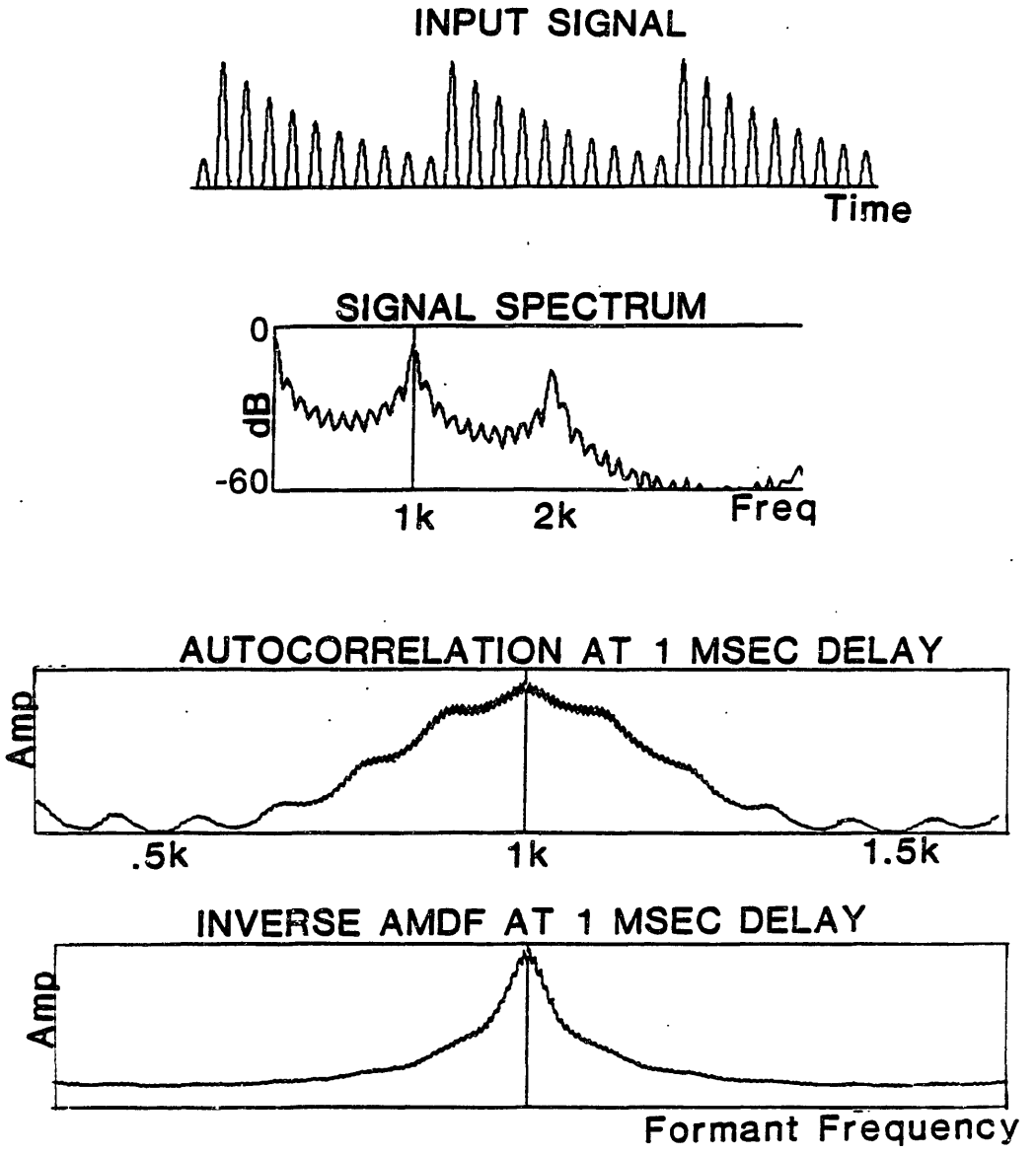

Figure 7.4: Response of autocorrelation function and inverse AMDF, both at fixed 1 ms delay, to an input signal consisting of a sequence of pulses spaced by a fundamental period of 100 Hz, passed through a single resonator, and half-wave rectified, as a function of resonance frequency.

to be more significant than might at first be expected; it will be discussed in more depth later in Chapter 11.

Whereas a zero in the denominator of $S_1[r]$ corresponded to weak responses, a zero in the denominator of $S_2[r]$ represents perfect synchrony. The ratio will then be huge regardless of the overall energy level, as long as the energy is significantly above the silence threshold. A partial solution to this problem is to move the zeros slightly in from the unit circle. An additional clipping of high amplitude responses can be obtained by passing the final output through a soft-limiter, such as an arc tangent function.

Any signals whose overall amplitude is less than δ will yield a negative response out of the synchrony measure. While such a negative response may be construed as inconsequential, it seemed preferrable to envision all responses in a neural domain as being positive. The final synchrony measure output can be passed through another half-wave transformation, identical to the one used for the peripheral model, which will accomplish both the removal of negative responses and the clipping of large responses to perfect synchrony.

In the thesis system, the synchrony measure is applied independently to the outputs of each of the channels of the initial stage processing, where each r is equal to the center r of the corresponding peripheral filter. The N channel outputs are plotted against center frequency as the abscissa to produce a "pseudo spectrum", which should show peaks at formant frequencies. A "pseudo spectrogram" is constructed by plotting each output as an intensity [darkness] level at a point in a two-dimensional grid of time versus frequency.

7.4 Pitch Estimation

The pitch estimator is very similar in structure to the spectral estimator. One major difference is that the initial stage outputs are first added together to produce the "pitch waveform", which is then fanned out to a series of synchrony detectors covering the pitch range. The outputs of the synchrony measures can be plotted as a "pseudo autocorrelation", with the delay τ as the abscissa. The first prominent peak in this function represents the pitch period. A voiced/unvoiced decision is made from a threshold on the amplitude of the pitch waveform in conjunction with a measure of the prominence of the peak in the pseudo autocorrelation at the pitch period delay.

In the case of pitch estimation, it was found to be difficult to decide upon a single δ to subtract from the numerator in order to suppress a response to weak signals. For pitch estimation the issues are slightly different; since the same signal is applied at the inputs to all of the synchrony measures, a weak input to one will be a weak input to all. A weak but strongly periodic input should nonetheless produce a strong positive response at the correct period. If δ is greater than the overall amplitude of the signal, then the peak at the fundamental period will be **negative**, a clearly undesirable situation. A solution is to set δ to zero, but in this case the responses will generally be elevated by an amount approximately equal to the overall level of the stimulus. Such a wandering of the baseline based on overall amplitude was felt to be undesirable, and hence a solution was to continually monitor the signal envelope, and subtract such an overall "rate" estimate from the

pitch waveform, prior to applying the GSD algorithm. This can be accomplished in practice by processing the pitch waveform through a linear filter, consisting of a single pole on the positive x-axis at a radius slightly less than 1.0, and a single zero at r = 1.0, to remove the DC component from the signal. This step would become more essential if higher CF filter outputs were also included, because the partial loss of synchrony of the high frequency fibers results in a large DC component in the peripheral level outputs.

With the DC component removed from the pitch waveform, the numerator of the GSD algorithm acts as a filter in addition to the denominator. When the input signal to the GSD is always positive, the two inputs to the sum integrator in the numerator always add constructively. However, once DC is removed, the two input samples will often be of opposite sign. The numerator then involves a filter of the form $(1+z^{-N})$ applied to the input signal. Such a filter has N zeros equally spaced around the unit circle, but offset from the zeros in the denominator by a radian frequency shift of π/N. Thus the numerator filter filters out the information **between** the proposed harmonics, accentuating the information at the proposed harmonics. Such a filter seems attractive for enhancing the contrast between signals which are periodic with the period of the GSD delay and signals which are not.

7.5 Summary

In this chapter we provided motivation for the proposed synchrony model for pitch and spectral estimation of speech. In particular, we described a "Generalized Synchrony Detector" (GSD), and illustrated that it appears to have certain advantages over other synchrony methods such as autocorrelation. We also described a method for generating a single waveform from which to extract periodicities at a fundamental frequency, and provided arguments for why it is advantageous to simplify the pitch extraction problem by reducing the information in all of the independent channels down to a one-dimensional signal.

In the next chapter we will fill in most of the details of the system design, so that the interested reader could reimplement the system on a different computer facility.

Chapter 8

Details of System Structure

8.1 Introduction

In Chapter 7, the system that was developed for the thesis was discussed at a descriptive level. In this chapter, the implemented computer system will be described in sufficient detail that the interested reader should be able to reproduce the essential characteristics of the system. A few examples of the outputs (both pitch and spectrum) for simple tone stimuli and speech tokens will be used to illustrate system performance characteristics. In subsequent chapters, more detailed analyses of specific tasks will be given, including some studies of the effects of varying certain parameters of the system.

The system obtains a spectral representation for the frequency region from 228 to 2667 Hz, which is referred to as a "pseudo spectrum", because it is not a spectrum in the usual definition of the word, and a pitch contour, derived from an analogous "pseudo autocorrelation". The output is intended for use in the analysis of **sonorant** segments of speech. The design of the system is motivated by the knowledge that auditory nerve fibers in the low frequency range respond synchronously with the stimulus waveform; i.e., are phase-locked to the stimulus.

At the core of both the spectral and pitch estimators is the "Generalized Synchrony Detector" (GSD), a mechanism for detecting specific periodicities in an input signal that eliminates the problem of picking up a strong response at the place associated with twice the frequency of an input stimulus. The GSD mechanism is also effective in normalizing out energy, and thus generates a response which is stable over wide ranges of input signal level. In addition, the measure requires only simple building blocks, that are not inconsistent with what is known about more central levels of auditory processing. The pitch estimation process and the spectral estimation process are similar; thus a unified approach to both problems is proposed, with only system constants being varied to suit the requirements of the one or the other task. The system was implemented on a Symbolics Lisp Machine in conjunction with a Floating Point Processor, manufactured by Floating Point Systems, which was responsible for most of the numerical computations.

8.2 Overview

The computer system consists of an initial stage of signal processing, whose outputs are delivered to both a spectral estimator and a pitch estimator. The initial stage processing is a terminal-analog model of responses at the peripheral level of the auditory system, i.e., the responses of actual nerve fibers as measured through period histograms. The model incorporates such effects as half-wave rectification and saturation, but no attempt was made to model these effects in any

physically meaningful way. The waveforms are never reduced to spike sequences; a value at time t thus represents a probabilistic response. A series of outputs $y_i[n]$ are produced, each of which is identified with a place on the basilar membrane corresponding to the center frequency of its bandpass filter.

Both the spectral estimator and the pitch estimator make use of the GSD algorithm to detect appropriate periodicities. The spectral estimator examines each of the outputs from the initial stage independently, and measures the synchrony in the response to the center frequency of the corresponding critical band filter. It attempts a sharpening of the peaks in the spectrum by detecting enhanced synchrony at the places tuned to formant frequencies. The pitch estimator first combines all of the outputs of the initial stage processing into a single waveform, which is then fanned out to a series of GSD's which detect periodicities at frequencies appropriate for human pitch. The summing process results in a reenforcement of the periodicities at the pitch frequency at the expense of periodicities at the formants.

8.2.1 Initial Stage Processing

The characteristics of nerve fiber response along the basilar membrane have been described in detail in Chapter 2. In this section we describe the portion of the system that is intended to model this response. The model can be broken up into three subtasks: 1) the linear filtering, due mainly to basilar membrane vibrations, 2) the envelope response characteristic due mainly to adaptation effects, and 3) the detailed wave shape of the phase-locked response.

A block diagram of the model is given in Figure 8.1. The diagram describes the model for a specific place on the basilar membrane; the input is thus the original stimulus, and the output corresponds to the probability of firing of a fiber at that place. The filter bank consists of 30 filters spaced by approximately half a critical band and spanning the frequency region from 228 to 2667 Hz. The analog speech is filtered at 6.5 kHz cutoff and sampled at 16 kHz. All of the tuned filters share a common linear phase lowpass filter, which removes high frequencies and provides a negative slope with frequency in the overall response characteristic. Each individual tuned filter consists of a double complex pole pair at center frequency. and a double complex zero pair above the center frequency. In addition, there are several zeros on the x-axis to help shape the slopes on the low and high frequency sides. As discussed previously, limited phase data on basilar membrane vibrations and nerve fiber responses suggest a 2π phase shift through resonance. It was this criterion that led to the initial choice of a **double** complex pole pair at the resonance frequency. The setting of the frequencies for the zeros was somewhat empirical; but the process to compute the radius of the poles was defined explicitly.

In order to specify the critical bandwidth criterion, the frequency scale in Hz was first converted to a Bark scale [Zwicker, 1961], through a nonlinear mapping function. In the Bark scale, a frequency difference of one Bark is a critical bandwidth. The conversion was obtained by means of

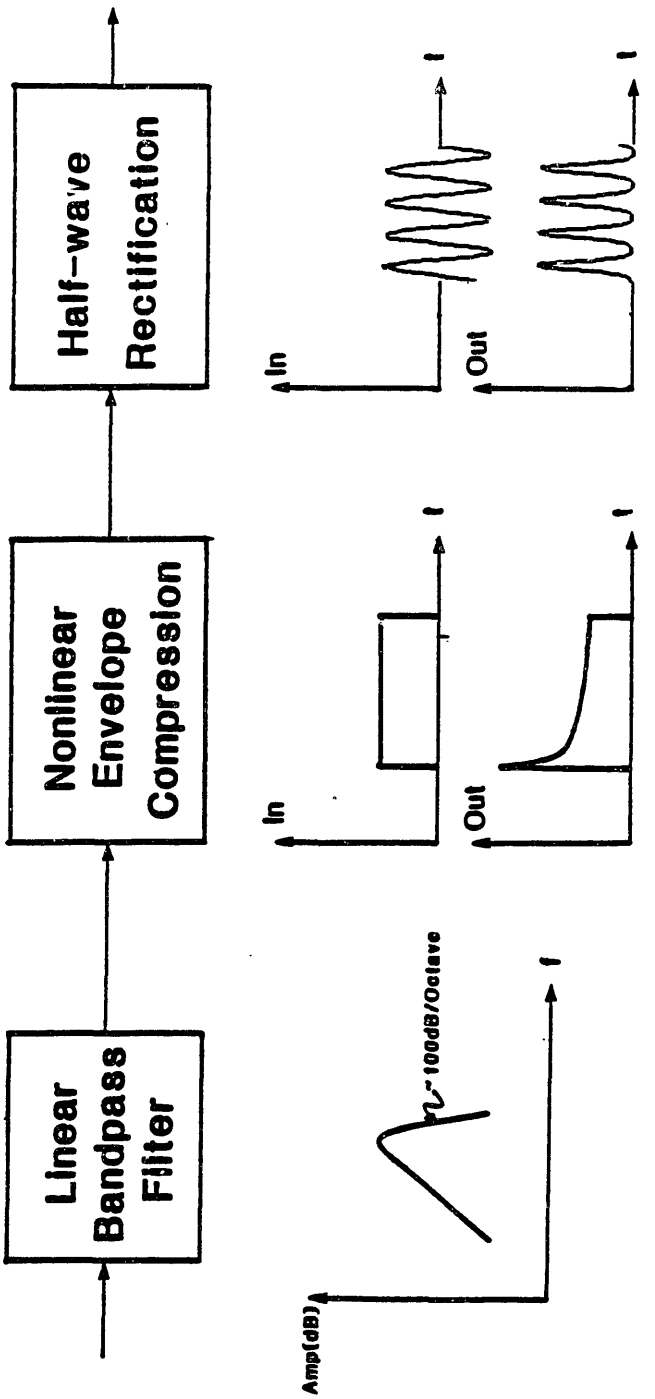

Figure 8.1: Block diagram of single channel of computer model for peripheral auditory system.

the following set of equations derived by Goldhor [1985].

$$B(f) = \begin{cases} .01f, & 0 \le f < 500 \\ .007f + 1.5, & 500 \le f < 1220 \\ 6\ln f - 32.6, & 1220 \le f \end{cases} \tag{8.1}$$

where f is the frequency in Hz, and B is the frequency in Barks. For a given filter, centered at frequency f_0, the critical bandwidth in Hz is obtained by first evaluating $B_0 = B(f_0)$, and then inverting the process to obtain $f(B_0 - 1/2)$ and $f(B_0 + 1/2)$. The difference between these two frequencies is then the critical bandwidth in Hz.

The overall gain of each filter, derived from the height of the tip of the tuning curves of auditory nerves, was specified as a function of characteristic frequency as follows:

$$G(f) = \begin{cases} -\frac{1}{48}(\frac{f}{100})^2 + \frac{1}{3}(\frac{f}{100}) - \frac{1}{3} & f < 800Hz \\ 1 & Otherwise \end{cases} \tag{8.2}$$

i.e., the gain increases parabolically from 200 to 800 Hz, with a value of 1/4 at 200 Hz, and a value of 1.0 at the peak of the parabola at 800 Hz. The gain is then constant at 1.0 above 800 Hz. The gain was reduced at low frequencies because tuning curves tend to have an elevated tip in the low frequency region [see Figure 2.3].

The major design task was to determine the radius of the double complex pole pair that would provide the critical bandwidth, given the existence of the zeros. This was accomplished through linear approximations as follows:

1. Compute the slope of the response characteristic of the filter near center frequency (f_c) when all poles and zeros are included except the double pole at $+f_c$, and the two poles at $-f_c$ are assumed to have a radius of 1.0.

2. From a linear approximation to the slope, and using Equation 8.1 above to determine critical bandwidth, compute two locations above and below f_c where the 3dB points should be. These two frequency locations are separated by a critical bandwidth but are not equidistant from f_c. [Assuming equidistance from f_c is incorrect if there is a strong negative slope superimposed on the resonance characteristic].

3. By linearizing the unit circle near f_c, use geometric considerations to determine the radius of the resonance poles that will yield critical bandwidth.

4. Adjust the gain according to 8.2 above.

Figure 8.2 shows an example of the magnitude (a) and phase (b) responses of the filters. Because of the flexibility in the placement of the zeros, it is easy to manipulate the filter shapes somewhat,

although total control of the shapes, as would be the case in a DFT implementation, is not available. In subsequent chapters, some effects on the final system output of changes in the characteristics of the filters will be shown. A complete derivation of the method used to compute the filter coefficients is given in the Appendix.

The output of each linear filter is next passed through a cascade of two identical Automatic Gain Controls [AGC's], in order to effect amplitude compression and adaptation phenomena. This stage corresponds to the "Nonlinear Envelope Compression" box in the Figure. Following the AGC's is a half-wave rectification scheme. Several different rectification strategies were tried; the final version uses a raised hyperbolic tangent, which tends to increase the peakiness of the positive cycles of the input waveform.

The first AGC has a short ($< 5ms$) time constant for the integration, and thus can be equated roughly with rapid adaptation [Smith and Zwislocki, 1975]. The time constant for the second AGC is around 40 ms, a typical value for short-term adaptation. Long-term adaptation effects are not included in the model.

The AGC's are described by the following equation:

$$y[n] = \frac{x[n]}{k + < |x| >_\tau}$$

where $y[n]$ is the output, k is a constant, $< |x| >$ refers to an integration of the magnitude of the input, $x[n]$, and τ is the time constant over which integration is performed. This form is similar to models that have been proposed by Siebert [1973] and Colburn [1973] for similar tasks. It was decided to integrate magnitude rather than energy because the former corresponds more closely to a mean rate response. Integration is performed through a "leaky integrator", an infinite impulse response filter of the form:

$$H(z) = \frac{1 - \alpha}{1 - \alpha z^{-1}}$$

The constant k controls the amount of amplitude compression that is realized. When k is large relative to input signal level, the response is approximately linear, but when k is insignificant, increments in y are proportional to increments in the log of x. The input signal was stored in 16-bit integer form, so that the largest input signals peak at a value of 32767. The value of k for the first AGC should be relatively large [ranging from 400 to 1000], whereas k for the second AGC can be fixed at 2.0.

The final step in the peripheral model is half-wave rectification. Histogram data consistently show a sharpening of the peak shape relative to the input sine wave shape; an exponential rectifier has been a commonly used model to describe the shape distortion (Johnson, 1975; Young and Sachs, 1979). Unfortunately, when dynamic aspects are considered, it is impossible to use an exponential rectifier alone because the response becomes unacceptably large at sudden onsets. A solution is to follow the exponential rectifier with a soft limiter, for which the arctan function is an appropriate

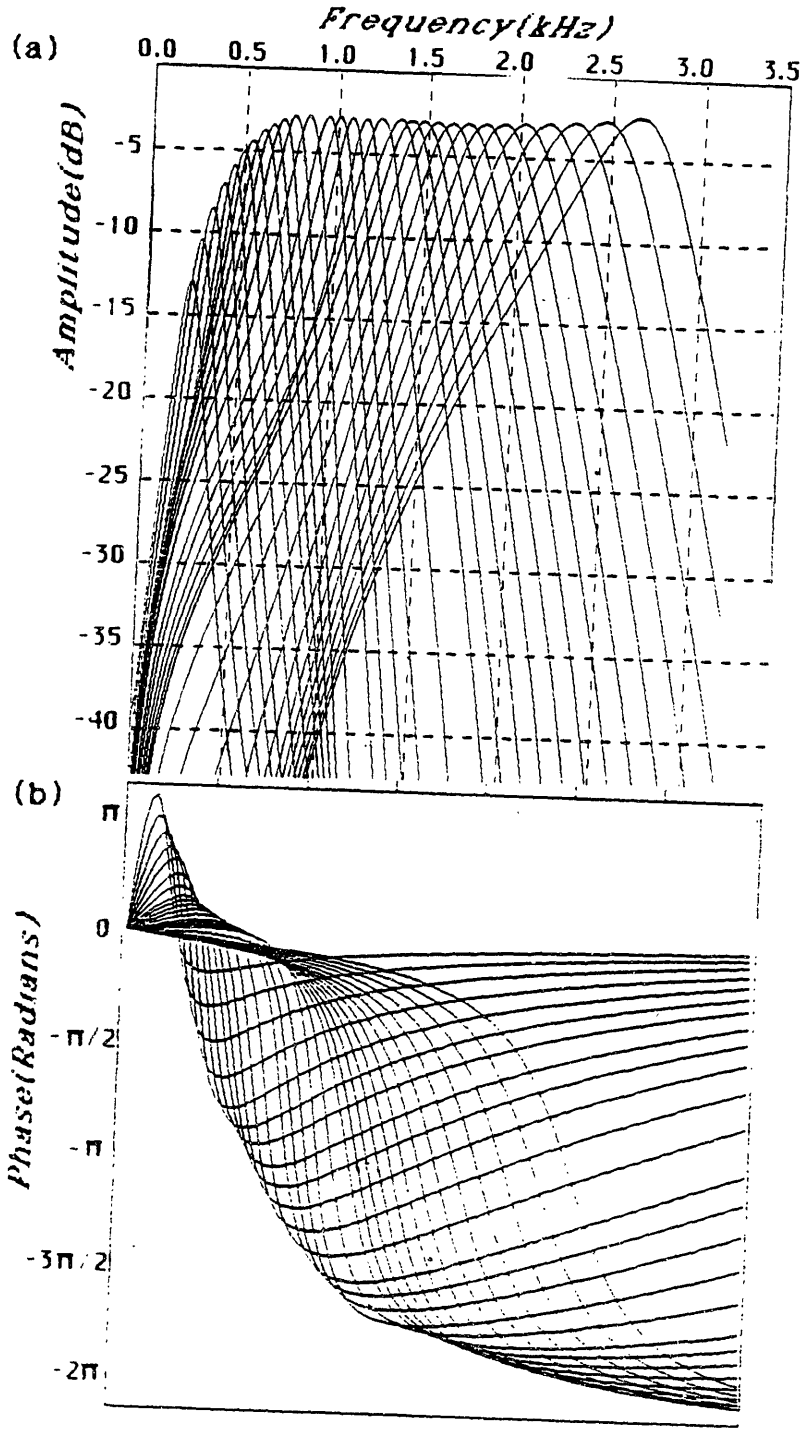

Figure 8.2: Magnitude (a) and phase (b) response characteristics of bank of 30 filters used to model auditory filters. Linear phase term has been removed.

choice. Thus a possible form for the half-wave rectifier is as follows:

$$y = G\left(atanA\left[e^{Bx} - 1\right] + atanA\right) \tag{8.3}$$

where x is the input, y is the output, A, and B are constants, and G is the gain. By adding the term $atanA$, one assures that the output is always positive. This term becomes, in effect, the spontaneous rate.

Data on the relationship between hair deflection in the hair cell and receptor potential (Hudspeth and Corey, 1977) can be modeled quite accurately by a hyperbolic tangent function, with suitable adjustments on the origin [Weiss and Leong, in preparation]. If one can argue for a more-or-less linear relationship between receptor potential and probability of firing, then the hyperbolic tangent might be viewed as a reasonable model for the half-wave rectifier. A proposed formula is the following, a raised and horizontally shifted hyperbolic tangent [Searle, personal communication]:

$$\frac{L_{Sat}\, e^{x-\Delta}}{e^{x-\Delta} + e^{-(x-\Delta)}} \tag{8.4}$$

with L_{Sat} as a gain term and Δ controlling the ratio of spontaneous to saturation levels, R. It turns out, as shown in Figure 8.3, that these two rectification strategies are very similar, if the factor B in Equation 8.3 is set to 2.0, and R is matched for the two strategies.

The simplest form of half-wave rectification is the piecewise linear form, for which negative input values are simply set to zero and positive values left unchanged. This form has been used implicitly, for example, by Rose et al. (1974), in order to estimate the level of each independent component of the response when the input was a two-tone complex. The piecewise linear rectifier is the easiest to control; for example, unlike the other two rectification schemes, it does not distort the overall gain. However, it exhibits neither a spontaneous nor a saturation level. Furthermore, the sharp corners in the waveshape that are introduced whenever the input crosses the origin are undesirable.

Some examples of outputs at several stages of the peripheral model for some simple input stimuli are shown in Figures 8.4 through 8.6. In all of the figures, the half-wave rectifier is the raised hyperbolic tangent form. This is in fact the form that was always used for the experiments described in the thesis. Figure 8.4 shows the response to a stimulus consisting of a pure tone at 500 Hz, with 6dB amplitude increments at fixed 20 ms intervals in time. The linear filter is centered at the 500 Hz frequency, and hence this represents a response to a tone at f_c. The envelope compression stage greatly reduces the dynamic range, but this compression is offset somewhat by the final half-wave rectifier. There is a more pronounced response to the first cycle after each amplitude increment than to later cycles, because the nonlinearities of the AGC's are not instantaneous. Irregularities in the peak levels after half-wave rectification are mostly due to insufficient sampling in the plot.

Figure 8.5 shows the detailed temporal response characteristics for a steady state tone at a center frequency of 1.5 kHz, for three different stimulus levels. It is clear that the envelope compression scheme preserves the sinusoidal wave shape. The half-wave rectification expands the dynamic range,

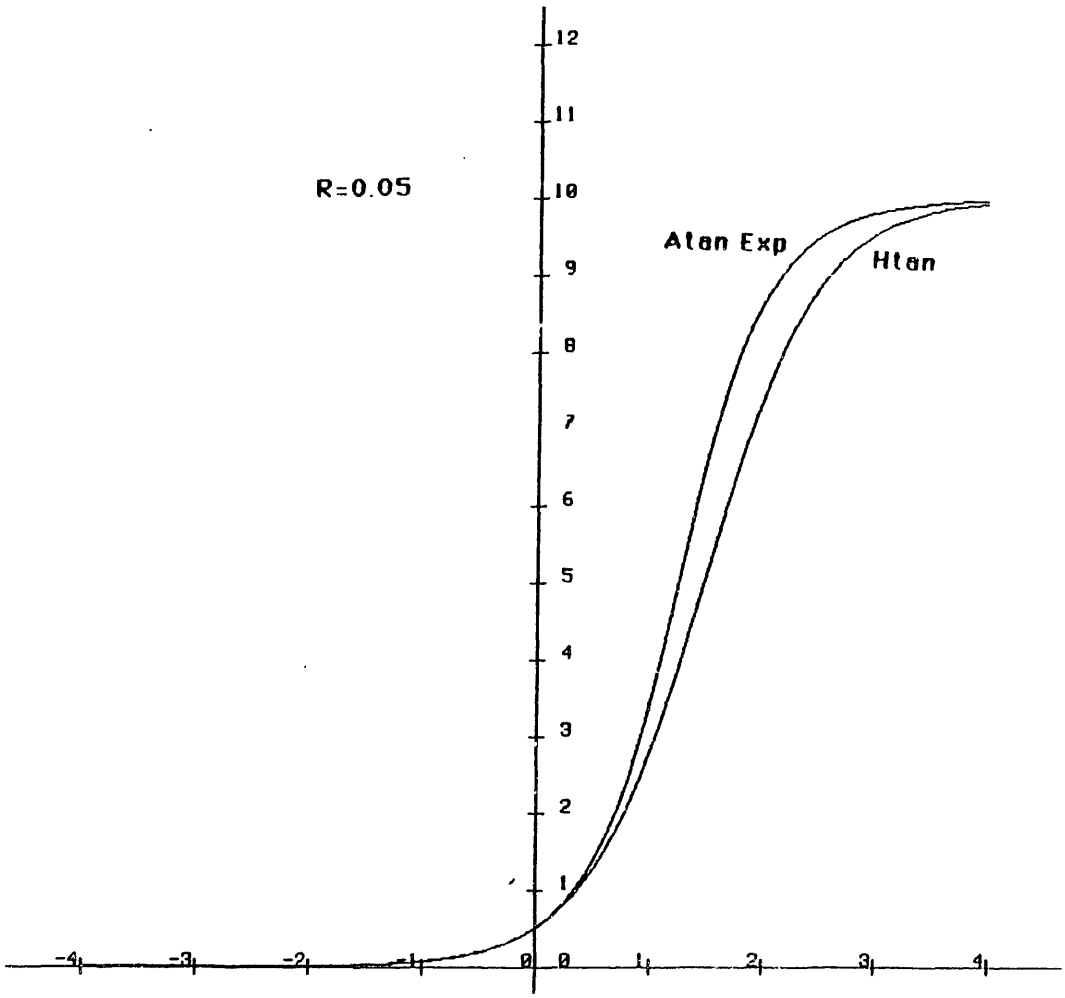

Figure 8.3: Comparison of two formulas, "Atan Exp" [Equation 8.3] and "Htan" [Equation 8.4] in text, for instantaneous nonlinearity characterizing half-wave rectification scheme. A value of 2 is used for B in Equation 8.3. A and δ are derived from R, the ratio of spontaneous to saturation levels.

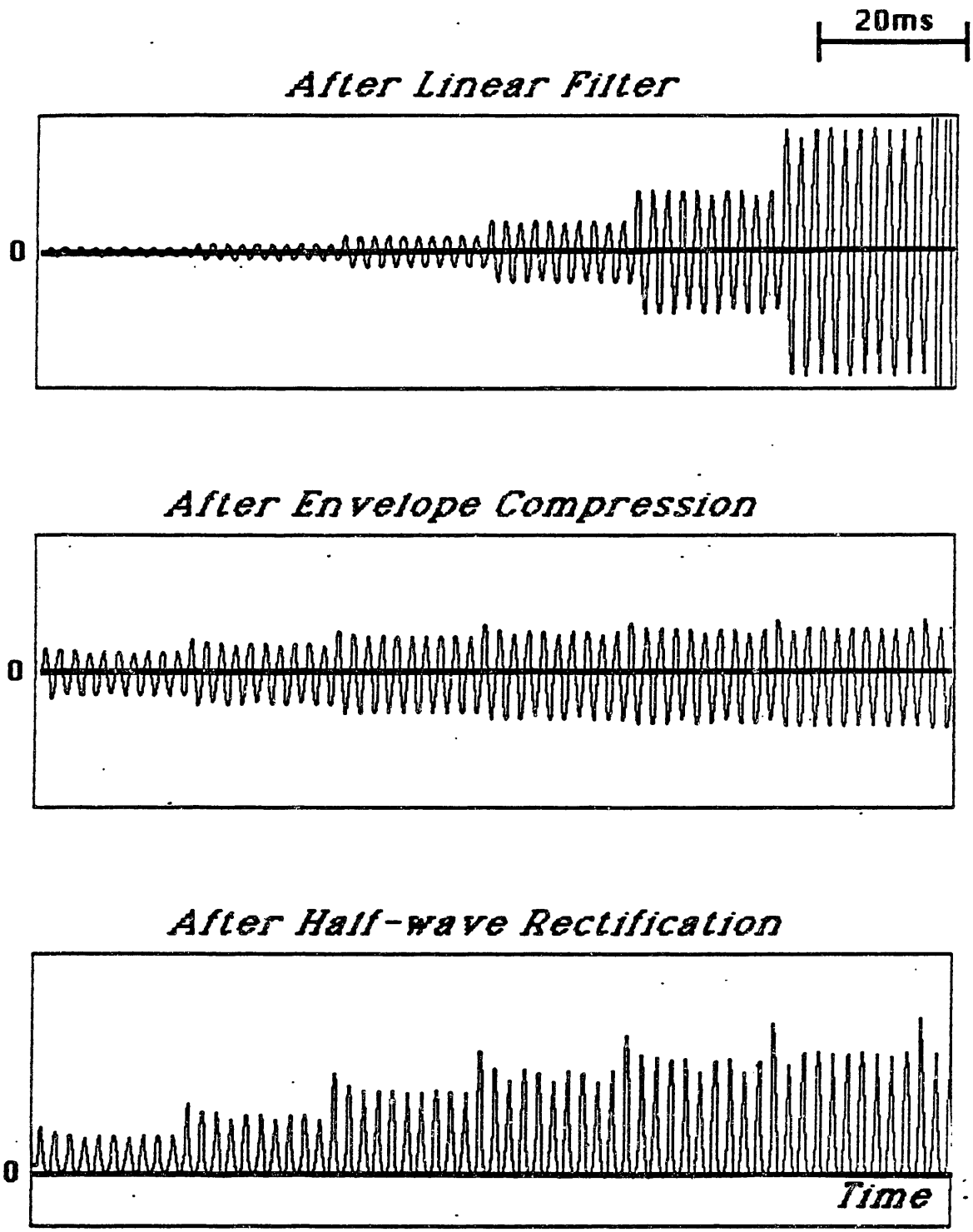

Figure 8.4: Response of channel in peripheral model at 500 Hz f_c to a tone at f_c, which increases in amplitude by 6dB every 20 ms.

and the final wave shapes for weak signals are more similar to sinusoids than for high amplitude stimuli.

Figure 8.6 shows responses of a single channel to stimuli at three different frequencies. The filter is centered at 800 Hz, and the stimuli are sine waves at 400 Hz, 800 Hz, and 1600 Hz. The response to the tone at twice f_c is greatly reduced because of the sharp tail of the filter on the high frequency side. The final response to the tone at half f_c is substantial. Because of the distortions introduced by the half wave rectifier, the final output contains a significant amount of energy at the f_c of the filter, which, as discussed previously, could introduce problems for certain schemes for synchrony detection.

8.2.2 Generalized Synchrony Detector

In this section we will describe the Generalized Synchrony Detector (GSD) which was developed for the dual purpose of enhancing formant peaks and detecting periodicities in the pitch waveform. The task for this detector is to compare two incoming waveforms and to produce a strong response if the two waveforms are of sufficient amplitude, similar in detailed shape, and time-aligned. In the use of the GSD in the system, the two inputs are always a signal and that same signal delayed by some specified time interval, although the GSD could also be applied, for example, to two distinct binaural inputs.

Motivation for the choice of this particular synchrony measure was given in Chapter 7. It can be described through the following equation:

$$S(f_0) = A_s \, atan \frac{1}{A_s} [\frac{< |y[n] + y[n+n_0]| > - \delta}{< |y[n] - \beta^{n_0} y[n-n_0]| >}] \tag{8.5}$$

where $S(f_0)$ is the synchrony output for frequency f_0, n_0 is equal to sr/f_0, where sr is the sampling rate, <> represents envelope detection, and A_s, β and δ are constants.

A block diagram of the GSD algorithm is given in Figure 8.7. A sum and a difference waveform are constructed from the two input waveforms, and the full-wave rectified outputs of both of these are passed through identical lowpass filters, to obtain an envelope response. The constant β is set to a value slightly less than 1.0, and serves to position the zeros in the filter characteristic of the denominator slightly inside the unit circle, thus reducing the sharpness of the nulls at multiples of n_0. This modification from a simple difference was found to be advantageous in the low frequency [F_1] region in particular, where sharp nulls combined with narrow peripheral filters could lead to overly precise tuning.

The form of the lowpass filters is a cascade of two identical leaky integrators; i.e., a double pole on the real axis near z = 1.0. Thus the temporal characteristic of the filter is as follows:

$$h[n] = \begin{cases} (n+1)\alpha^n & n \geq 0 \\ 0 & n < 0 \end{cases}$$

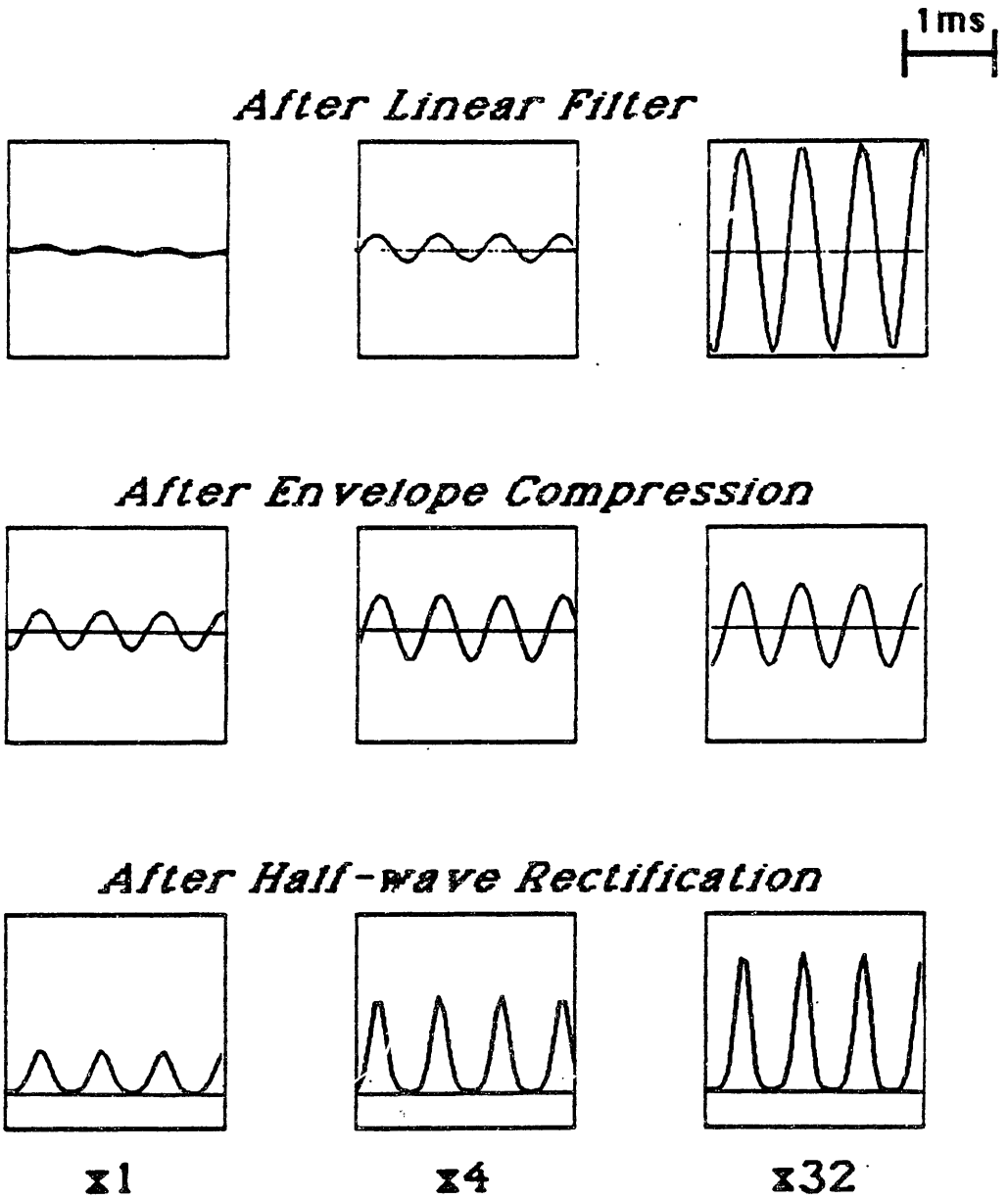

Figure 8.5: Steady state response of peripheral model to sine wave at f_c, for channel tuned to 1500 Hz, and for three different signal levels.

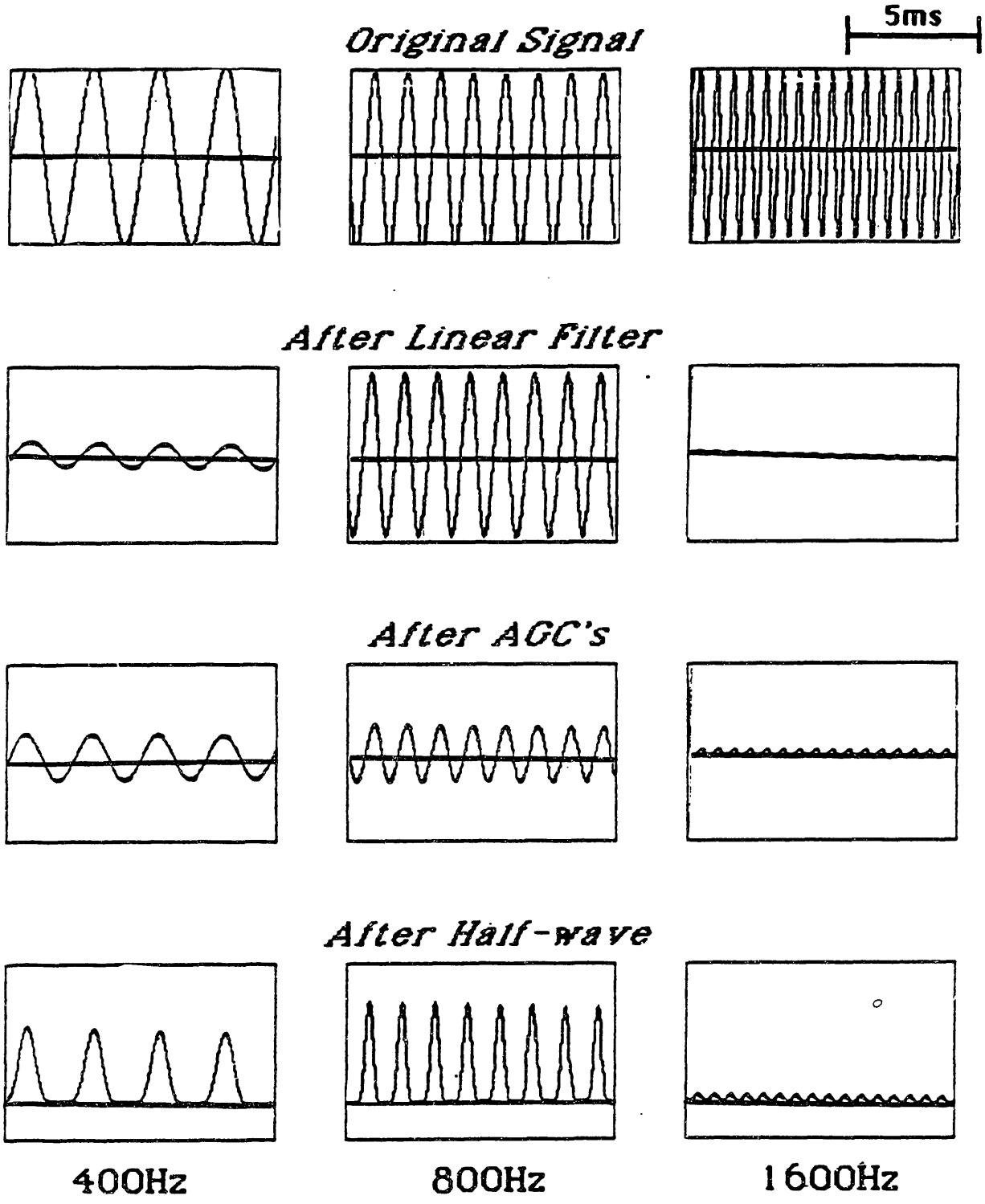

Figure 8.6: Example showing response of peripheral model as a function of tone frequency. All input signals are steady state tones of equal amplitude, and the channel is tuned to 800 Hz. The tones are at frequencies of half f_c [left], f_c [middle], and twice f_c [right].

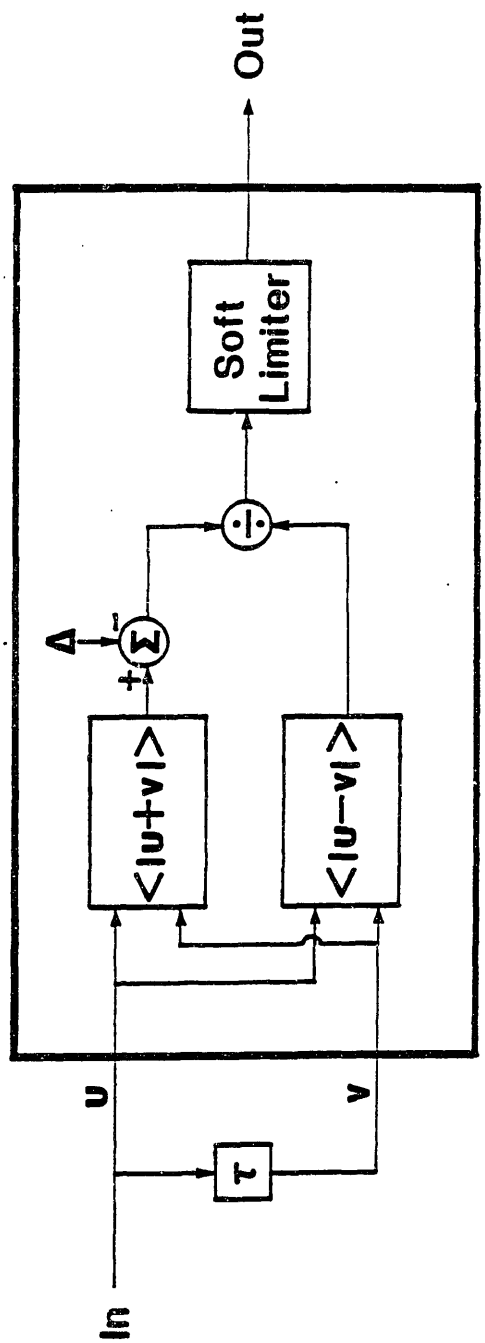

Figure 8.7: Generalized Synchrony Detector

The variable that is manipulated to control the integration window size is τ_{pk}, the time delay of the peak in $h[n]$. α can be derived from τ_{pk} as follows, where sr is the sampling rate:

Let $n_{float} = (sr)(\tau_{pk})$

Then, $\alpha = exp[-1/(n_{float} + 1)]$

The reason for choosing a double leaky integrator rather than a single one is to prevent the immediate past from carrying too much weight in the effective summation window for y[n]. The effective window shape for a double leaky integrator, although still infinite in duration, is more like a typical Hamming window than for a single leaky integrator, as shown in Figure 8.8. A filter with $\tau_{pk} = \tau_0$ is best matched to a Hamming window of duration $4\tau_0$, as shown in the figure.

A small threshold, δ, set at a level slightly above the spontaneous rate, is subtracted from the sum envelope, in order to suppress a response to small amplitude signals. The sum envelope is then **divided** by the difference envelope, so that, in the event that the two signals are very similar, a response approaching infinity is obtained. The soft limiter, to clamp potentially infinite outputs, is an instantaneous nonlinearity defined as follows:

$$y = A_s \, atan(x/A_s)$$

Thus, the output, y, is equal to the input, x, when the input is near zero, but saturates at $A_s\pi/2$ when the input is large. An increase in the value of A_s causes an increase in the linear range for the input.

8.2.3 Spectral Estimation

A block diagram of the procedure used to generate pseudo spectra is given in Figure 8.9. The output of each of the 30 channels of the initial stage processing is compared with itself delayed by a "center tau", equal to the inverse of the resonance frequency of the linear filter associated with that channel. The comparison is made by feeding the waveform and the delayed waveform through a GSD, and the output of the GSD is recorded every time a new spectral image is desired (typically 10 ms in the implemented system). A time sequence of the 30 pseudo spectral samples can be used to generate a pseudo spectrogram, with the output of the GSD as the intensity dimension. Linear interpolation in both frequency and time were used to generate a higher resolution spectrogram display, for visual integration purposes.

The 30 channels are spaced by approximately, but not precisely, half a Bark. The center frequencies of the filters are of necessity quantized in accord with the nearest integer delay available at the discrete sampling rate. The actual values of the center frequencies are given in Table 8.1. It was in fact necessary to upsample some of the high frequency filtered signals to 32,000 Hz in order to obtain adequate frequency resolution.

The system contains several parameters that can be manipulated. In the initial stage processing are included the two integration time constants for the AGC's, τ_1 and τ_2, the term, k_1, controlling the amount of compression preceding the rectification, L_{Sat}, the saturation level of the rectifier, and R_Δ, the ratio of spontaneous to saturation level in the rectifier. In the GSD computation, there

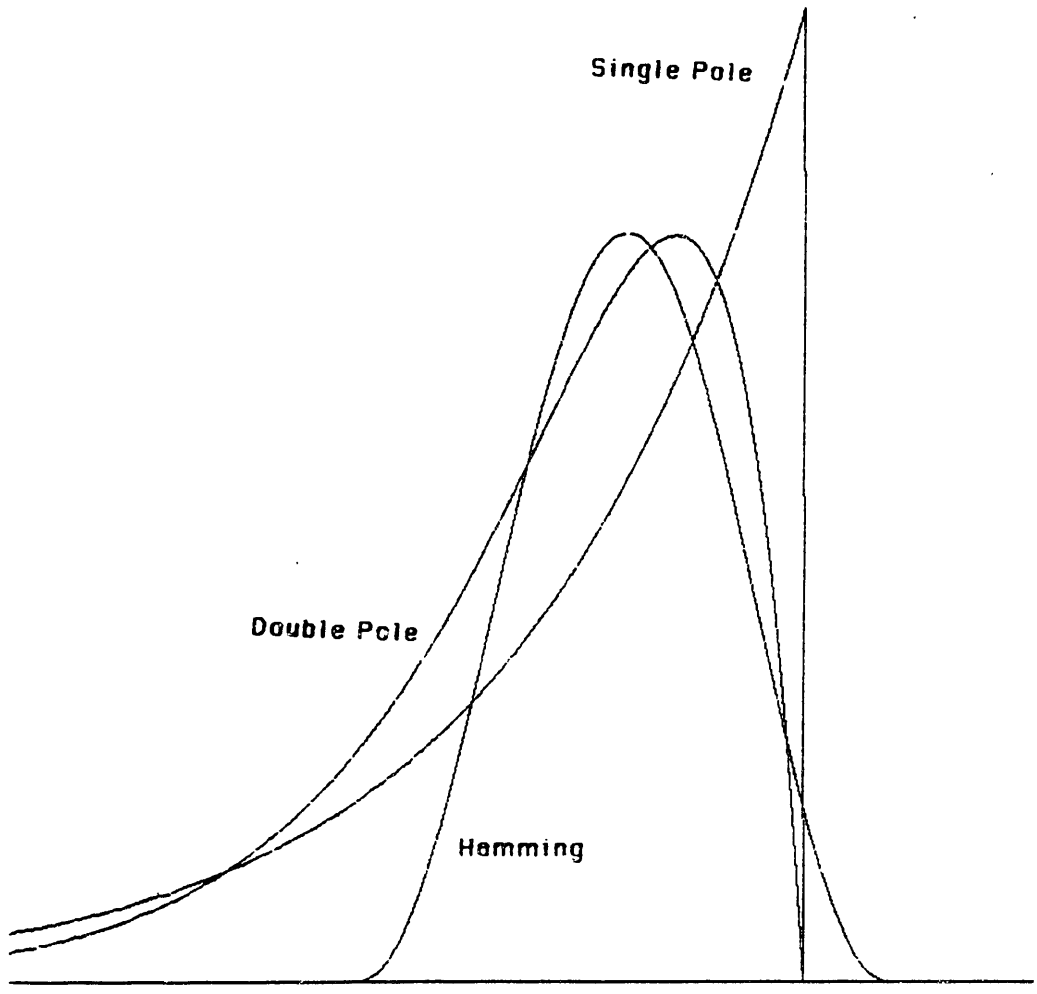

Figure 8.8: Comparison of two recursive filter window shapes with standard Hamming window. Radius of single pole equals square root of radius of double pole. Hamming window length is four times distance from origin to peak in double pole filter (r_{pk}).

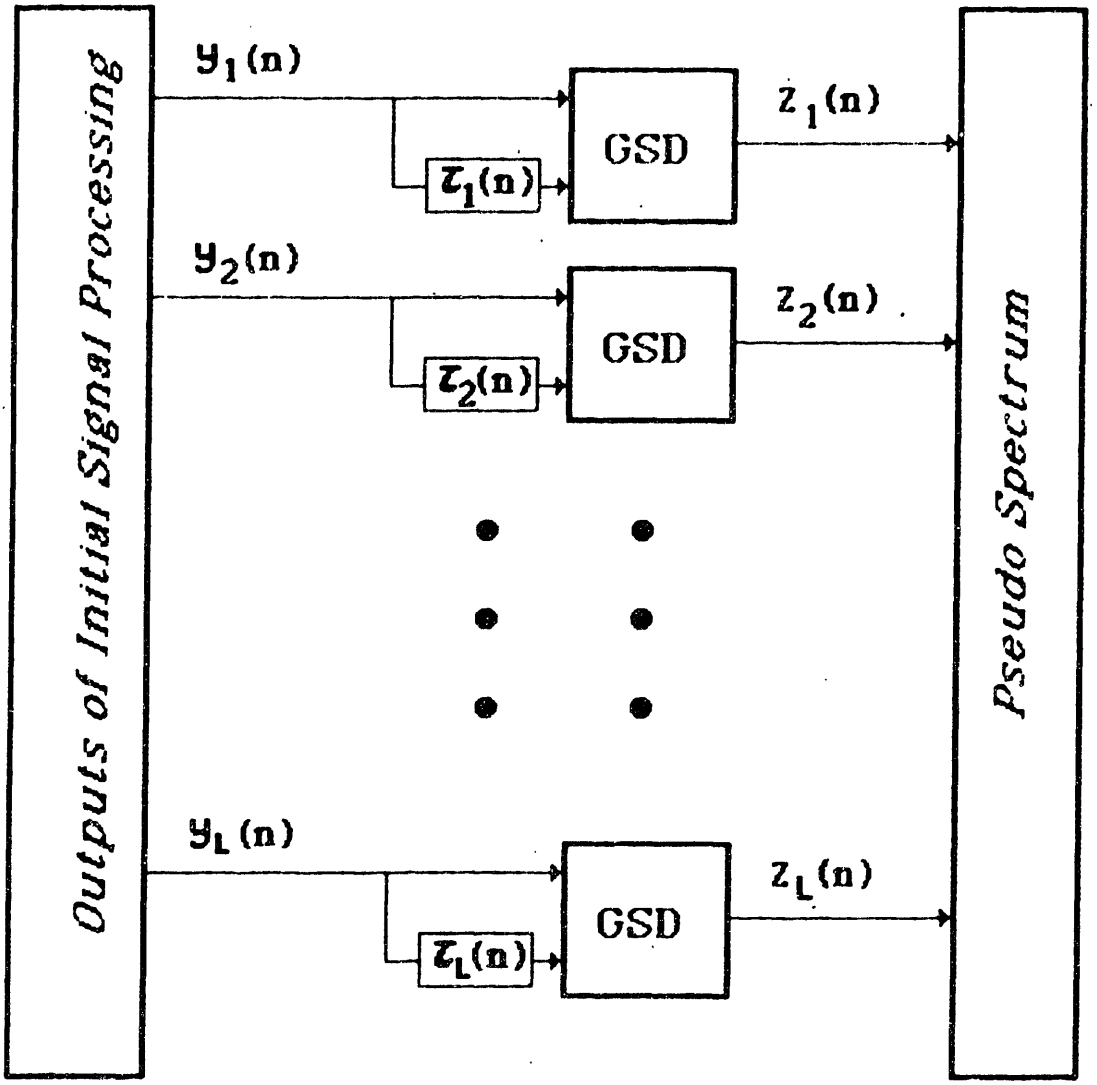

Figure 8.9: Block diagram of pseudo spectral analysis procedure.

228.6 Hz	1142.9
275.9	1230.8
326.5	1333.3
381.0	1391.3
432.4	1454.5
484.8	1523.8
533.3	1600.0
592.6	1684.2
640.0	1777.8
695.7	1882.4
761.9	2000.0
842.1	2133.3
941.2	2285.7
1000.0	2461.5
1066.7	2666.7

Table 8.1: Center Frequencies of 30 filters used in pseudo spectral anaysis. Center frequencies are chosen to be integer divisors of the basic discrete sampling rate, which is usually 16 kHz. When necessary, the higher frequency filter outputs are upsampled to 32kHz, in order to allow a smaller spacing of the filters.

are the integration time constant, τ_{pk}, the ratio r_δ of δ to the spontaneous rate, and the factor $A_\bullet$ controlling the soft limiter. The actual values that were used for these parameters are as follows:

$$\begin{aligned}
\tau_1 &= 40ms \\
\tau_2 &= 3ms \\
k_1 &= 800 \\
L_{Sat} &= 6 \\
R_\Delta &= .05 \\
\tau_{pk} &= 4ms \\
r_\delta &= 1.01 \\
A_\bullet &= 4
\end{aligned}$$

An impression of how the system responds to signals with sharp resonances can be obtained by examining the response to a sine wave, and comparing the outputs of the initial stage processing to the outputs obtained after processing through the series of GSD's. A display of the outputs in spectrogram form reveals the ability of the GSD algorithm to enhance spectral peaks, as illustrated in Figure 8.10. The input was a 1000 Hz tone, whose amplitude increased step-wise by a factor of two in quantal stages over a two second time interval. The upper spectrogram was produced

from the envelope of each channel output, $y_i[n]$, of the initial stage processing. As the amplitude increases, the response pattern expands upward above 1000 Hz, because critical band filters have a slow fall-off on the low frequency end, an effect that is compounded by the saturation of filters near the tone frequency. Filters below 1000 Hz do not pick up the signal because of the steep high frequency skirt of each filter. The lower spectrogram displays the outputs, $z_i[n]$ of the GSD's. There is a narrow peak at 1000 Hz that is maintained constant as overall amplitude is increased. This constancy is achieved because there is an effective energy normalization procedure inherent in the divide of the GSD. These effects are also demonstrated in cross-section at three different amplitudes, as indicated in the Figure.

In Figure 8.10, there is no response to the 1000 Hz tone at 2000 Hz, and thus responses to second harmonics are suppressed by the GSD algorithm. In fact, when the half-wave rectified sine wave is compared with itself delayed by **half** its period, the sum and difference waveforms are nearly identical. Hence the denominator in Equation 8.5 is near its theoretically maximal level, and consequently the overall response is minimized. This result is critically related to the fact that the distortion component is in phase with the original sine wave. There begins to be a response at the half-frequency [500 Hz] when the amplitude of the tone is very large. This response is still considerably less than the response at 1000 Hz, but it points out the need for a sharp cutoff on the high frequency side, to avoid passing data at double the center frequency, which would be perfectly synchronous to the center period.

Figure 8.11 shows the synchrony analysis applied to a token of natural speech. In the Figure, a comparison is made between the GSD output, the output of the initial stage processing, and standard Fourier analysis, for the word "ice", spoken by a male speaker. Part (a) of the Figure compares a wide-band spectrogram with a spectrogram obtained using mean rate response from each filter output as the intensity, and a spectrogram obtained from the outputs of the GSD's in the synchrony model. The vertical lines in the wide-band spectrogram are a consequence of fluctuations in the amplitude over time due to the pitch. Such fluctuations are missing in the middle spectrogram because of integration over a broader time window. However, the peaks at the formant frequencies have become somewhat obscured due in part to saturation phenomena. The synchrony spectrogram shows the outputs $z_i[n]$ of the GSD as the intensity dimension. The pitch striations are still absent, and, in addition, the formant peaks have been sharpened considerably.

Part (b) of the Figure shows a set of spectral cross sections taken during the /*a*/ and during the /I/ of the off-glide. The top spectra are obtained by Fourier analysis with a 25 ms long Hamming window applied to the speech. Thus, the fundamental period of voicing is evident as a sequence of harmonics in the spectrum, but there are clear peaks in the envelope at the formant frequencies. The middle slice from the initial stage processing output shows broad peaks at the formant frequencies, with most of the harmonic structure gone because the filters are broader in frequency than in the case of the narrow-band spectrum. At the bottom are spectral slices taken from the synchrony spectrogram. The peaks at the formant frequencies have clearly been sharpened by the synchrony processing. There is in addition, especially in the /*a*/, a low frequency peak below F_1, probably corresponding to a glottal resonance, which shows up in all three spectral representations.

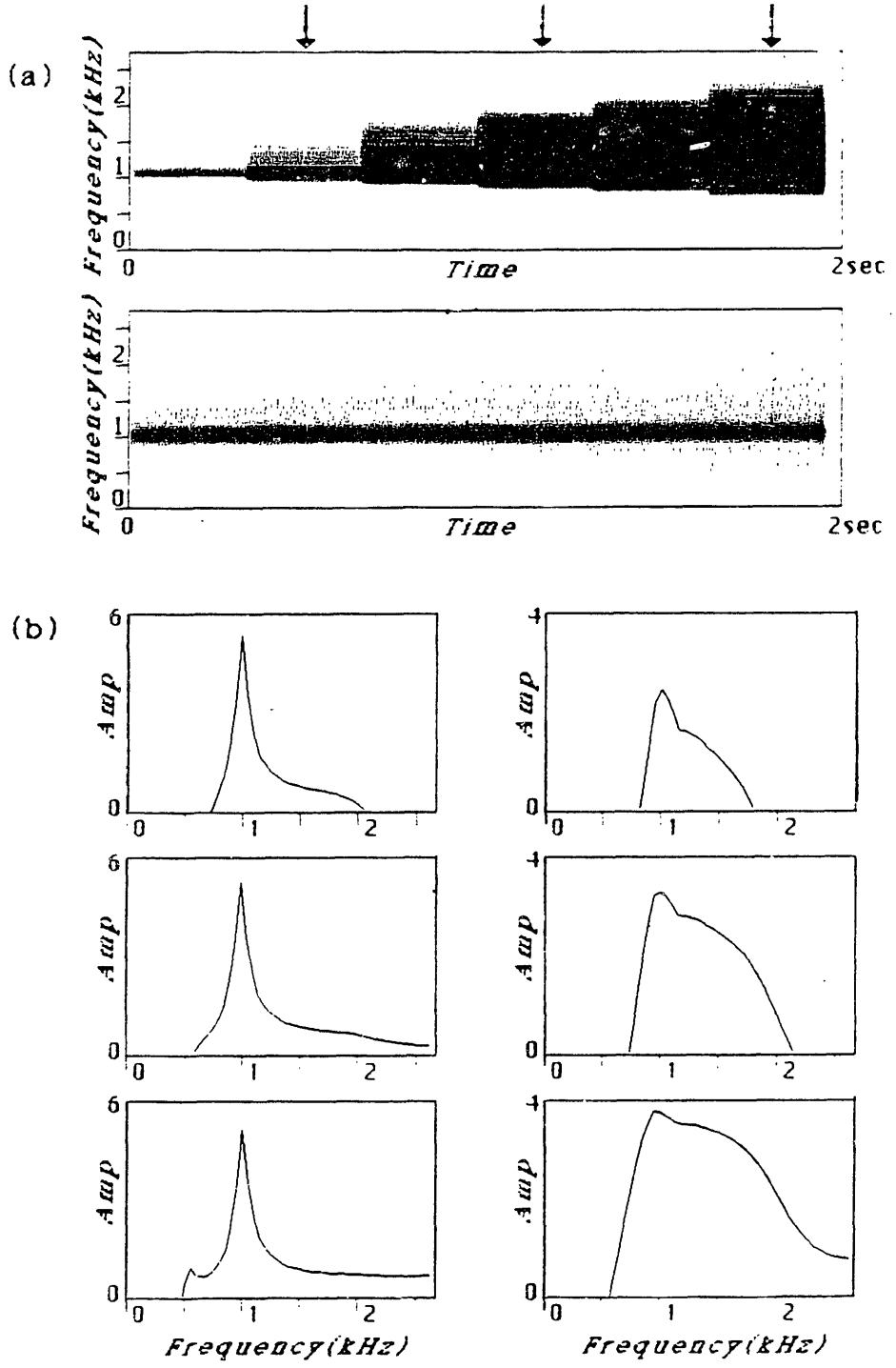

Figure 8.10: Response characteristics of system when input is a 1000 Hz tone.
a) Spectrogram computed from integrated outputs of initial stage processing [top] compared with pseudo spectrogram, computed from outputs of GSD's.
b) Time slices taken at three points indicated by arrows in part (a), comparing outputs of GSD's [left] with outputs of initial stage [right].

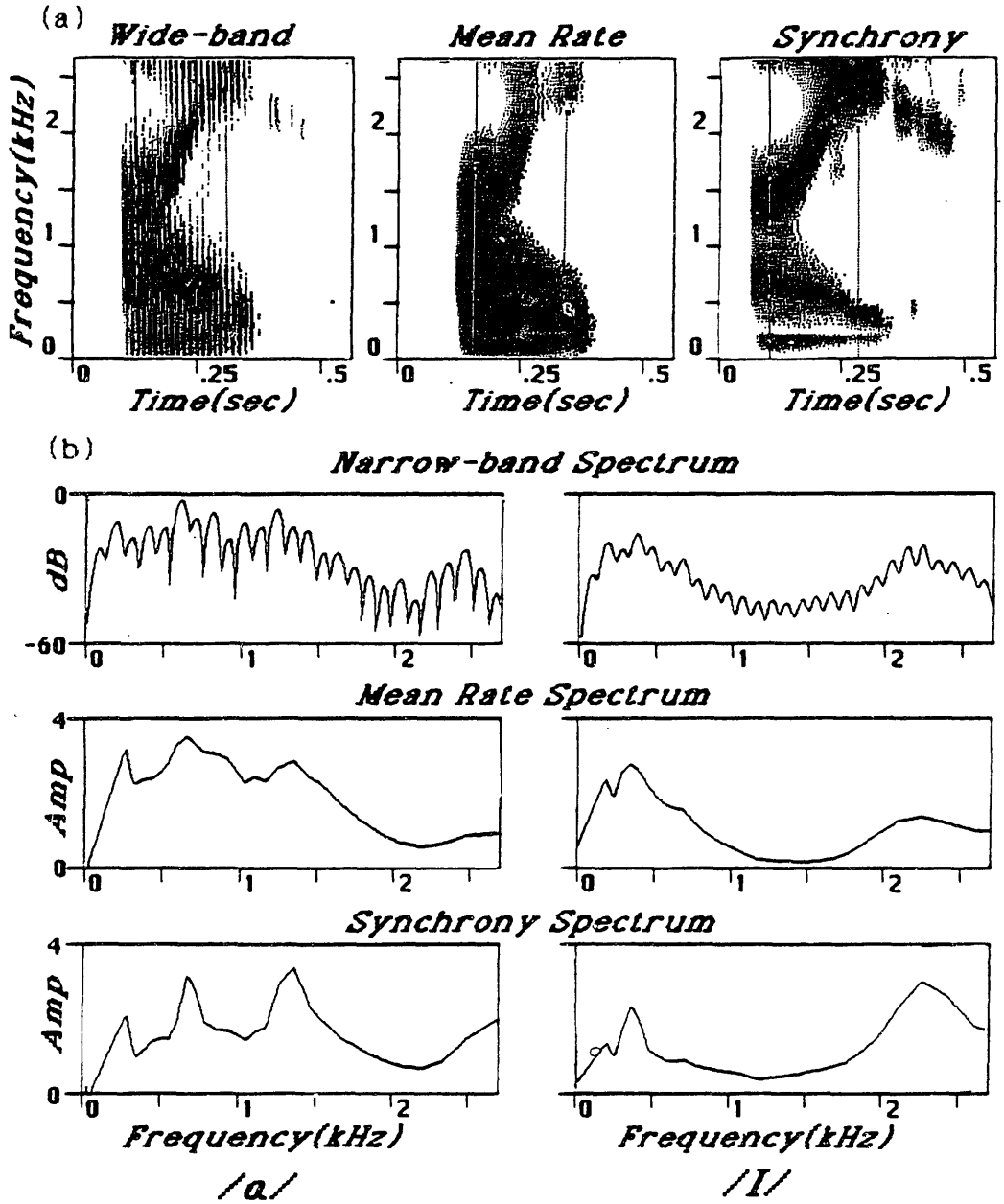

Figure 8.11: Example of synchrony analysis of natural speech, the word "ice", spoken by a male speaker. a) Wide-band spectrogram compared with spectrogram obtained using mean rate response of peripheral channel outputs for intensity levels, and spectrogram obtained from the outputs of the spectral GSD's. b) Spectral cross-sections in /a/ and /I/, at two time points indicated by vertical bars in part (a).

8.2.4 Pitch Estimation

A block diagram of the pitch estimation process is given in Figure 8.12. The process is quite similar to the spectral estimator, except that the [weighted] outputs of **all** of the channels are initially summed, and then passed through a linear filter to remove DC, as discussed in Chapter 7. The resulting "pitch waveform" is then fanned out to each of the GSD's for synchrony processing. The filters overlap by half a Bark; thus there is concern about the phase relationships of the individual filter outputs. It is for the pitch waveform that it is important to use a "correct" model for the phase of the basilar membrane filters. The limited data available on phase suggests a large linear phase component on which is superimposed a 2π phase shift through the resonance frequency. In adding the initial stage outputs, additional delays were introduced to offset the delays due to the linear phase terms; thus an attempt was made to align pitch pulses in the various filters as much as possible.

The delays for the GSD v-inputs (c.f., Figure 8.7) are considerably longer than the delays in the spectral estimator, spanning the periods appropriate for pitch in speech, from about 2 ms to about 16 ms. Each GSD also has its own time constant, γ, for the peak in the integration window of the sum and difference waveforms, which is monotonically related to the time delay, T of the v-input, according to the following empirical formula:

$$\gamma = 3ms + .4T$$

Thus it is possible to detect rapid changes in high fundamental frequencies without sacrificing the ability to detect low fundamental frequencies.

A "pseudo autocorrelation" can be constructed by plotting the outputs of the series of GSD's as a function of the time delay of the v-input. The pitch period is determined as the first prominent peak in this pseudo autocorrelation function. Because peaks will show up at integer multiples of the fundamental period, a few heuristics are needed to ensure that a fundamental is accepted when the response at a higher harmonic is slightly larger. A voiced-unvoiced decision can be made from the overall level of the pitch waveform and from a measure of the prominence of the peak in the pseudo autocorrelation. The details of the pitch extraction algorithm will be deferred until Chapter 12.

As in the case of the pseudo spectrum, some insight into how the pitch estimator processes waveforms can be gained by examining the outputs of the system at various stages when the input is a sine wave. Figure 8.13 shows the pitch waveform and pseudo autocorrelation for such a case; on the left the frequency is 100 Hz, and on the right 500 Hz. In both cases, the pitch waveform is periodic with the same period as the input, but with a distorted shape. The pseudo autocorrelation for the low frequency tone shows only a single peak at the period of the tone (10 ms), whereas the high frequency tone generates a series of equally spaced peaks at multiples of the fundamental period. From this example, it is clear why the heuristics are needed to pick the first prominent peak. At the bottom of the Figure is a trace of the pitch estimates that are produced for a chirp, consisting of a sweep in frequency from 60 Hz to 550 Hz over a two second interval. The algorithm

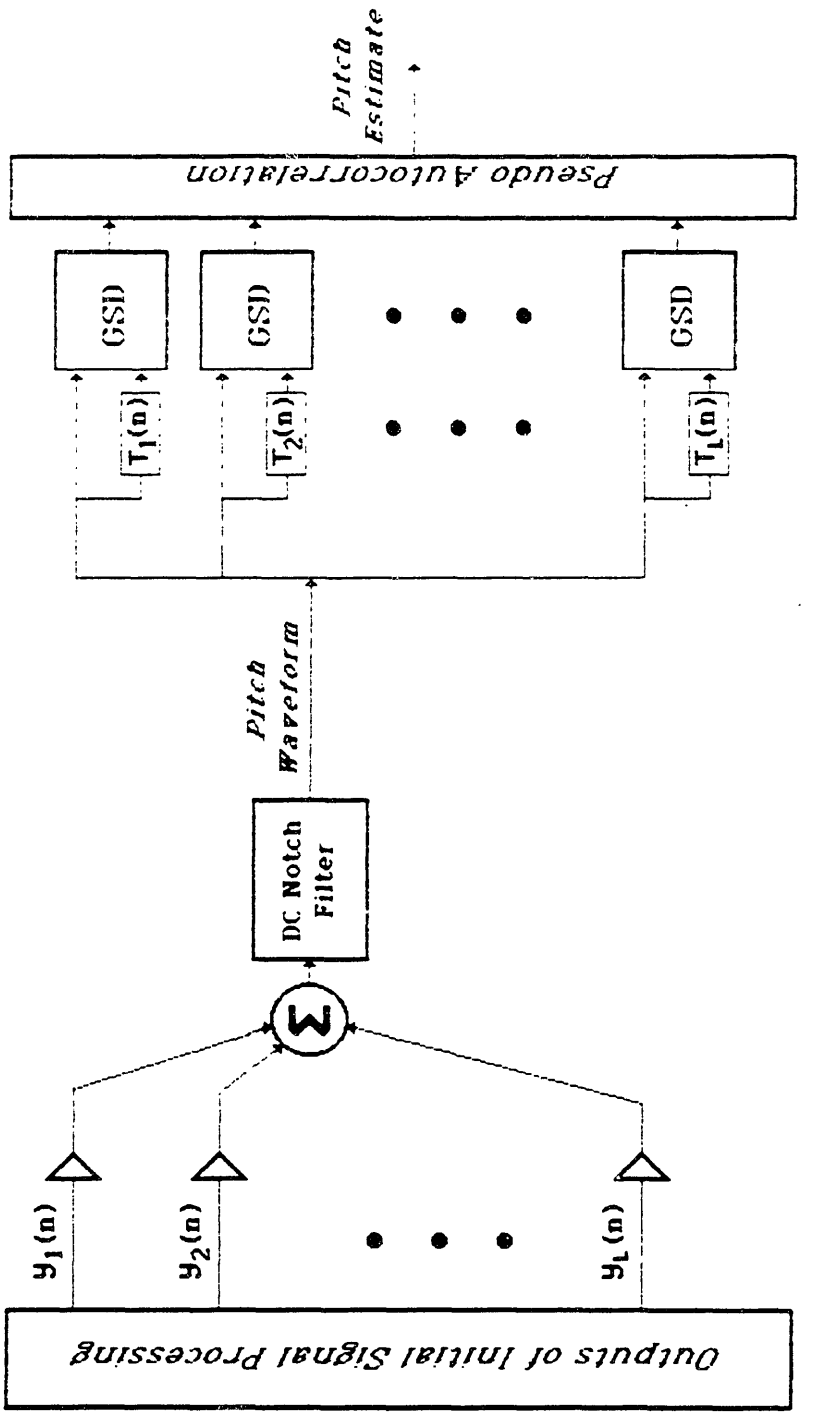

Figure 8.12: Block diagram of pitch estimation procedure.

is able to detect the correct pitch for all frequencies; staircases observed at the high frequency end are due to the fact that only discrete sampling intervals are available in time.

Figure 8.14 shows the pitch estimator applied to a sample of natural speech, the word "wish" spoken by a male speaker. The pitch waveform [part (b) of the Figure] bears a family resemblance to the original waveform shown in part (a). Because of the adaptation and saturation effects that were implemented in the first stage of the model, the pitch waveform tends to begin more abruptly at an onset of a voiced region than the original waveform. An increased abruptness at onset can lead to an enhancement of the peak in the pseudo autocorrelation at the fundamental period. As soon as there are two periods of the waveform, the denominator of the GSD tuned to the correct pitch period rapidly approaches zero, while the numerator becomes large due to the strong onset amplitude response. The pseudo autocorrelation applied at vowel onset is shown in part (c) of the Figure. There is a prominent peak at the fundamental period of voicing. At the bottom is the original waveform on a compressed time scale, and above that is the aligned pitch track that is produced by the estimator. The pitch waveform is derived through a highly nonlinear procedure; however, it can be conjectured that a form of spectral flattening is achieved as a consequence of saturation effects in the initial stage processing. It will be shown in Chapter 12 that this is the case.

Any pitch detection algorithm that claims to be a possible model for human pitch processing should at the very least not disagree with human results for perception of certain pathological signals for which we have psychophysical data. For this reason, it was felt appropriate to process inharmonic signals, as described in Chapter 4, through the pitch detection algorithm. The results are given in Figure 8.15. Part (a) concerns a stimulus consisting of the sum of four tones at 900, 1100, 1300, and 1500 Hz. There is thus a constant spacing of 200 Hz between the tones, but they are offset by 100 Hz from true harmonics. Human subjects tend to hear two pitches for such a signal, one slightly above 200 Hz and one slightly below. Below the original waveform is the derived pitch waveform, and on the right is the output of the pseudo autocorrelation. There are two prominent peaks, of about equal amplitude, just above and just below 5 ms. Part (b) of the Figure shows the pitch estimate obtained by the algorithm for a sequence of four-tone stimuli with the lowest frequency tone rising in 20 ms increments from 700 to 1000 Hz. The estimated pitch frequency gradually increases from a low of 184 Hz to a high of 222 Hz as the offset in frequency from true harmonics changes from -100 Hz to +100 Hz. As the stimulus frequencies pass upward through $(n + 1/2)f_0$, the later peak in the pseudo autocorrelation becomes larger than the earlier one, causing a sudden jump to a longer estimated pitch period. This trend is similar to what has been observed from psychophysical studies of human subjects, as discussed in Chapter 4. An exact duplication of human results is not expected, because no attempt was made in the peripheral model to account for the introduction of additional harmonics below the lowest one present in the signal, the "essential nonlinearity" described in Chapter 4.

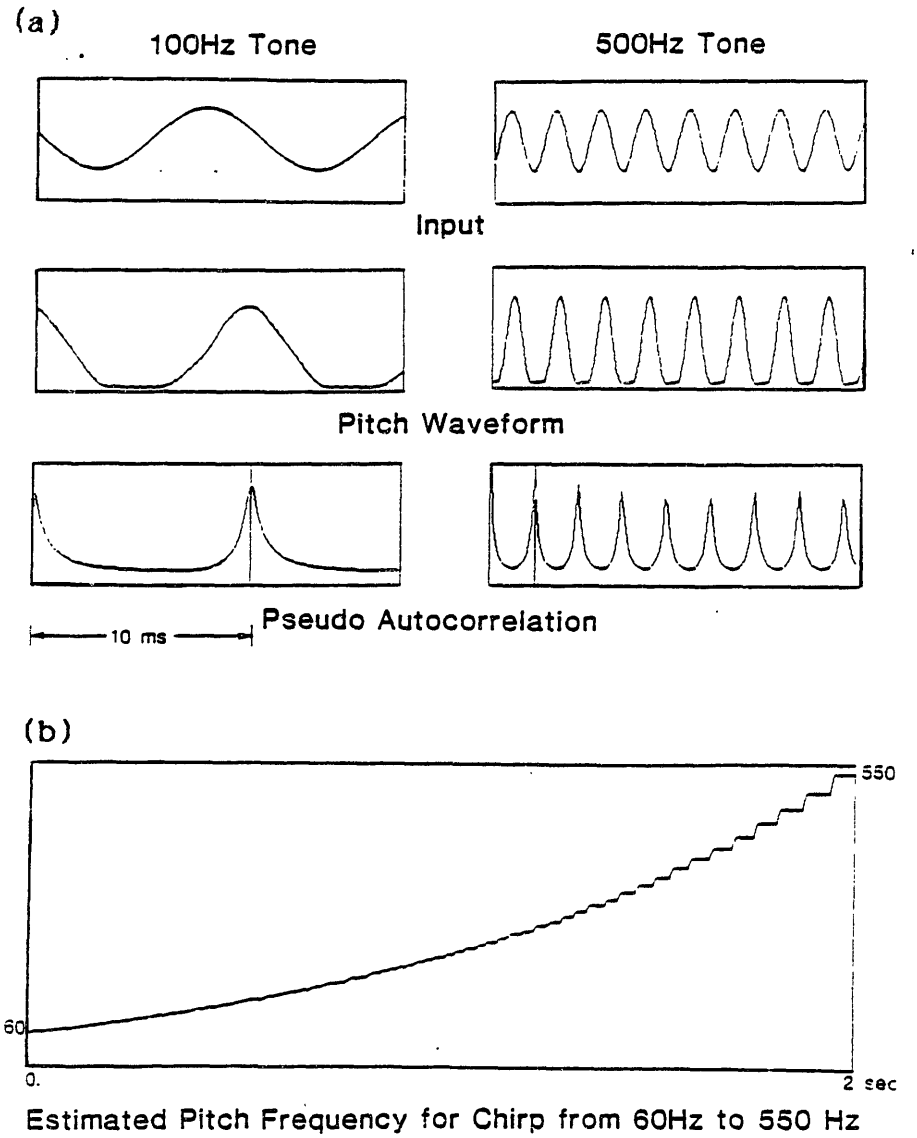

Figure 8.13: Results of processing sine waves through pitch estimator.
a) Original waveform, pitch waveform, and pseudo autocorrelation of 100 Hz tone [left] and 500 Hz tone [right].
b) Pitch track obtained for chirp signal changing in frequency from 60 to 550 Hz over a two second time interval.

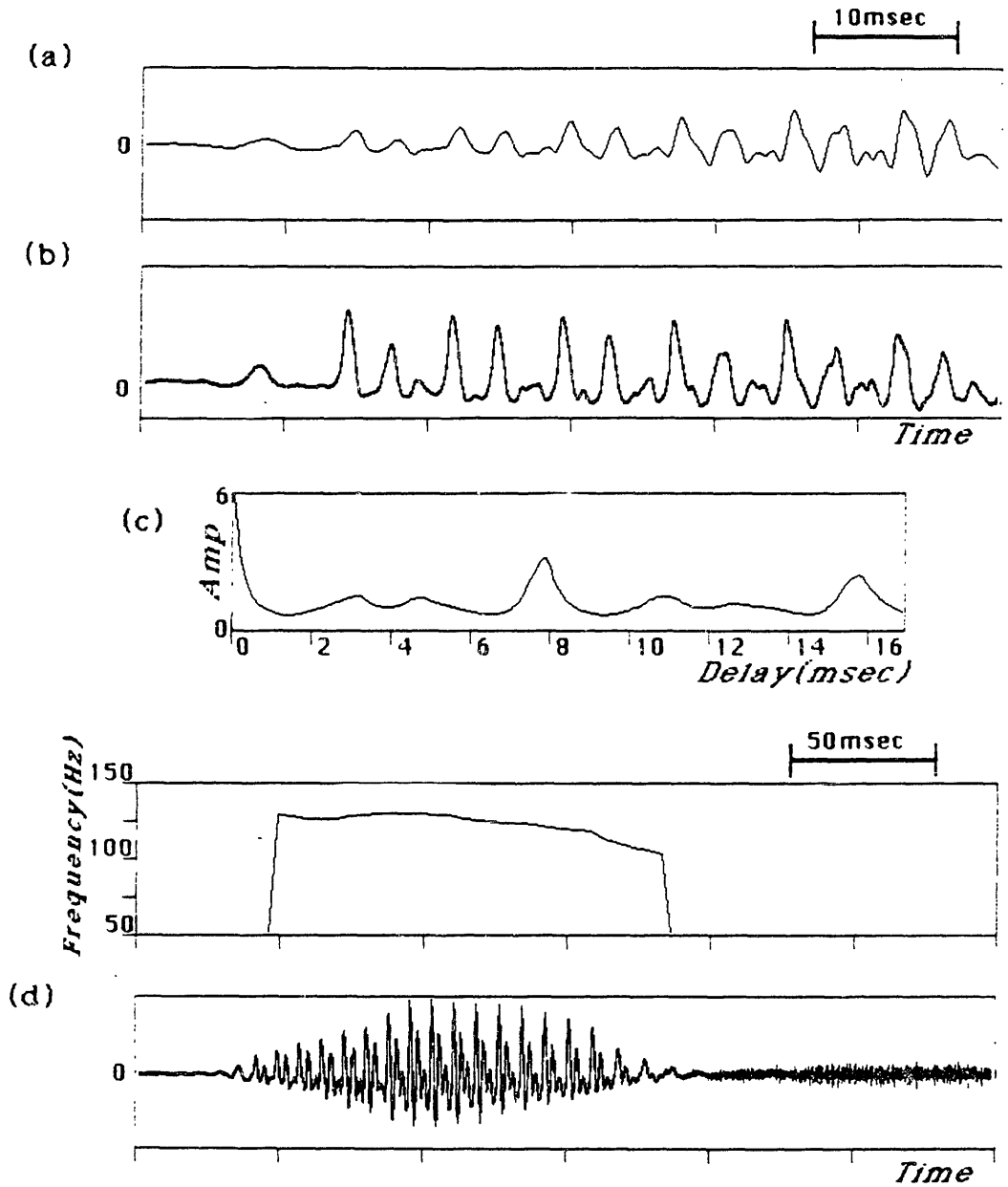

Figure 8.14: Results of processing speech through pitch estimator. The utterance is the word "wish" spoken by a male speaker.
a) Original wavform at vowel onset,
b) Pitch waveform obtained by summing filter outputs and then filtering out the DC component of the sum waveform,
c) Pseudo autocorrelation obtained at vowel onset, after the first two periods have been included under the time window.
d) Resulting pitch track aligned with original waveform.

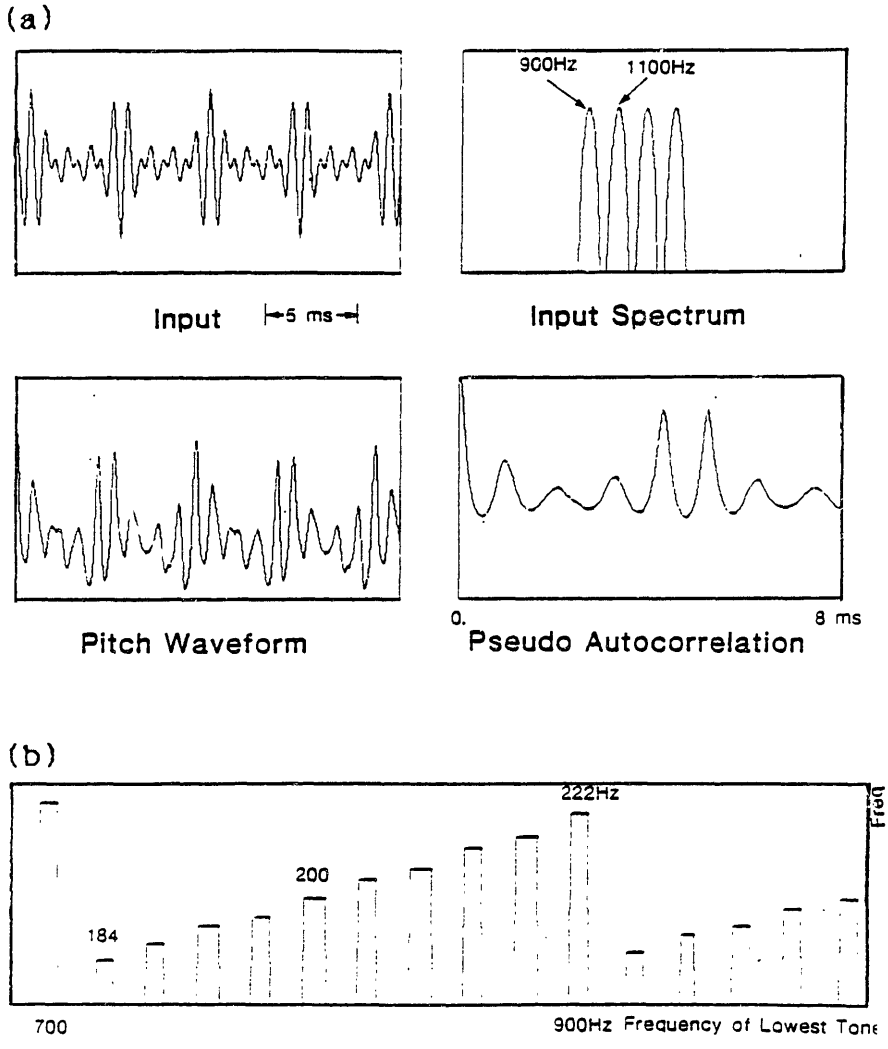

Figure 8.15: Results of processing inharmonic sequences through pitch estimator. a) Results of analyzing a signal consisting of a sum of four tones at 900, 1100, 1300, and 1500 Hz. A listener perceives two fundamental frequencies both offset from the 200 Hz spacing frequency.
b) Estimated pitch frequency obtained for a series of inharmonic sequences, each differing from the previous one by a uniform spectral shift upward of 20 Hz. Thus the first signal consists of components at 700, 900, 1100, and 1300 Hz, and the last signal consists of components at 1000, 1200, 1400, and 1600 Hz.

8.3 Discussion

In this chapter we have attempted to described the implemented computer system in enough detail that a reader could reimplement something quite similar on another computer. We have also illustrated certain features of the system, by applying the analysis methods to simple stimuli, such as sine waves, and to selected examples of natural speech. In Chapter 9, we will show several examples of pseudo spectral analysis of natural and synthetic speech tokens. Chapter 10 will give examples of spectral analysis of some synthetic data which can be related to certain perceptual experiments. In Chapter 11, we will examine the relative importance of various aspects of the system, and, in particular, will consider alternative definitions of synchrony detection that might have been used instead of the one chosen for the thesis. Chapter 12 describes the pitch detection algorithm in more detail, and shows some examples of pitch extraction from difficult data. In the final chapter, we will suggest potential improvements in the model, discuss how the synchrony system could be interpreted in terms of natural neural systems, and hypothesize methods for further processing of the synchrony spectrum aimed towards eventual phonetic recognition.

Chapter 9

Examples of Pseudo Spectral Outputs for Synthetic and Natural Speech

In this chapter the performance of the standard system for pseudo spectral analysis, as described in the previous two chapters, will be illustrated. Pseudo spectra and pseudo spectrograms of a variety of examples of natural and synthetic speech tokens will serve to characterize the system. Comparisons will be made with wide-band spectrograms and narrow-band spectral cross-sections, because these are the most appropriate standard methods for representing the spectral information. The selected examples will cover a number of different issues in speech processing, most of which were discussed in Chapter 6. These issues include formant mergers, rapid formant motion, nasalization, breathiness, high fundamental frequency, presence of noise, and anomalously weak formants.

The following sequence of example types will be illustrated, following a natural progression from simple to complex.

1. Synthetic CV's (Consonant Vowel) and synthetic CVC's,
2. Synthetic CVC's in noisy environment,
3. Natural CVC's spoken by both a male and a female speaker,
4. Natural isolated words and complete sentences spoken by male and female speakers,
5. Pairs of contrastive breathy/non-breathy words in a language where there is a phonetic distinction between the two.

It is difficult to assess the quality of the pseudo spectral analysis without certain biases. It is clear from multiple sources that formant frequencies convey a major portion of the information concerning the identity of voiced sounds, and that formant motions are also important features in the identification of adjacent consonants. Hence, at the very least, a spectral representation that obtains prominent peaks at the formant frequencies is expected.

A common problem in speech recognition is strong differences in overall spectral tilt among diverse speakers or recording conditions. A spectral representation that produces a formant "amplitude" that is somewhat independent of the absolute amount of energy in the signal at that frequency, but rather concentrates on the **relative** energy level, compared to the local environment, would intuitively be advantageous over standard Fourier analysis. Indeed, it is often the case that a weak second formant is realized as a much stronger peak in the pseudo spectrum than in

linear spectral analysis. This type of result can be considered as a feature of the pseudo spectral processing method.

Less clear are certain nonlinear effects that tend to occur in the first formant region, resulting in prominent peaks below F_1 in the pseudo spectrum that were not nearly as prominent in linear spectral analysis. Because the low frequency auditory filters have a bandwidth that is less than the spacing between harmonics of the pitch in typical female speech, all of the individual harmonics below F_1 tend to show up as peaks in the pseudo spectrum. For low vowels, there are typically two additional peaks below the first formant, at the first two harmonics of the pitch. Furthermore, the formant frequency becomes quantized to the nearest available harmonic frequency. Peaks at individual harmonics of the fundamental are not acceptable in a spectral representation that is to be used in a template-matching based speech recognition device. Clearly, the pitch can change over a fairly wide, although not indefinite, range, without disturbing the perceived vowel identity. Unless the template is restricted to a region that begins with the predicted frequency of the first formant peak, some alternative processing method, or some further processing of the pseudo spectrum, such as smoothing in frequency, would be necessary before it could be used as a template.

The pseudo spectrum usually produces prominent peaks at the formant frequencies. Exceptions occur mainly when a formant encroaches too close upon a neighboring formant. Whenever the second formant is within about 300 Hz of the first formant, the peak in the pseudo spectrum at the second formant frequency is drastically reduced in amplitude, relative to its amplitude using standard Fourier analysis. This reduction is a consequence of the fact that a significant amount of energy at the first formant frequency is being passed by the auditory filter centered at the second formant frequency. A similar phenomenon occurs with respect to F_2 and F_3, although in this case typically the third formant is completely absorbed into the second formant, so that only one peak is present in the pseudo spectrum, a very prominent peak essentially at the second formant frequency. Typically when two formants merge, the one lower in frequency dominates, unless, which is unusual in speech, the higher frequency peak is much higher in amplitude than the lower frequency peak. This asymmetry is a consequence of the asymmetry in the auditory filters: the steep slope on the high frequency side prevents energy from intruding from above.

Another aspect of the pseudo spectrogram which can be considered to be a clear feature is that there are never any pitch striations such as occur in typical wide-band spectrograms. These striations are manifested as vertical lines in the spectrogram, and are a consequence of fluctuations in overall amplitude, that depend upon the random placement of the integration window relative to the glottal pulses. Because the definition of the GSD algorithm effects an amplitude normalization, it is transparent to such fluctuations in input signal level.

The pseudo spectral analysis results will be compared for the most part with wide-band spectrograms and narrow-band spectral cross sections. Wide-band analysis [300 Hz wide central lobe on the Hamming window spectrum] is used for the spectrograms because this is the preferred choice for preserving the salient information for spectrogram reading. The pitch striations that appear as vertical lines in the picture do not present a major problem to the visual system. The eye is able to integrate across time and trace the formant frequencies quite accurately. Furthermore, the narrow

time window gives good temporal resolution, making it easier to decode visually crisp onsets such as stop bursts. Wide-band analysis is not however a good choice for isolated spectral cross sections, as discussed in Chapter 6, because there are large variations in spectral shape for two frames that are closely spaced in time. For both the wide and narrow-band spectral analysis, the speech is first lowpass filtered with a 2667 Hz cutoff filter, and 3:1 downsampled from the original 16,000 Hz sampling rate to 5333 Hz.

9.1 Synthetic CV's and CVC's

Examples of wide-band and pseudo spectrograms for a series of synthetic open vowels and diphthongs, preceded by the voiced consonant /b/ are shown in Figure 9.1. These synthetic stimuli were created using the Klatt synthesizer [1980] with specifications as outlined in Table 9.1. The amplitude of voicing and the pitch contour remained the same for all of the stimuli, but varied over time, as indicated in the table. The table also gives the formant frequencies as a function of time for each vowel stimulus. Linear interpolation between the fixed sample values describes the formant contours.

In examining the pseudo spectrograms, it is clear that the prominences at the formant frequencies are preserved, but there are no pitch striations across time such as exist in the wide-band spectrogram. The sharp dip in formant frequencies at vowel onset due to the labial context is preserved in the pseudo spectrogram, and perhaps even enhanced, relative to the wide-band analysis. The third formant peak is much stronger, in general, in the pseudo spectrogram than in the wide-band spectrogram. Such peak enhancement is a consequence of the fact that the pseudo spectral analysis focuses on the amplitude relative to the local environment, rather than the overall energy level.

The first and second formants show up as two isolated bands for all of the stimuli. For the synthetic /bɔ/, the first two formants are quite close in frequency, and the two peaks in the wide-band spectrogram are correspondingly somewhat merged. Although the pseudo spectrogram yields two separated peaks in this case, the prominence of the second formant is considerably reduced relative to, for example, the vowel /*a*/, whose formants are slightly further apart (400 Hz versus 270 Hz). This observation is made clearer with reference to Figure 9.2, which shows cross-sections in the middle of the vowel in /b*a*/ and /bɔ/. The amplitude of F_2 is substantially reduced in the latter case, because a significant amount of energy at the first formant frequency is being passed by the filter centered at F_2, thus reducing the measured synchrony to the F_2 frequency.

Cross sections for the remaining six stimuli are shown in Figure 9.3, compared with narrow-band spectra taken at approximately the same time slice. The pseudo spectra show prominent peaks only at the formant frequencies. The one case when a peak is absent where one is present in the spectrum is for F_3 in the /bɜ^/. The third formant is too close to the second in this case, and therefore too much information at 1500 Hz is passed by the filter centered at the 1850 Hz frequency, resulting in a loss of synchrony. Because of the wider critical band filters here than in the F_1 region, a wider separation between F_2 and F_3 is necessary to obtain a separate peak for F_3.

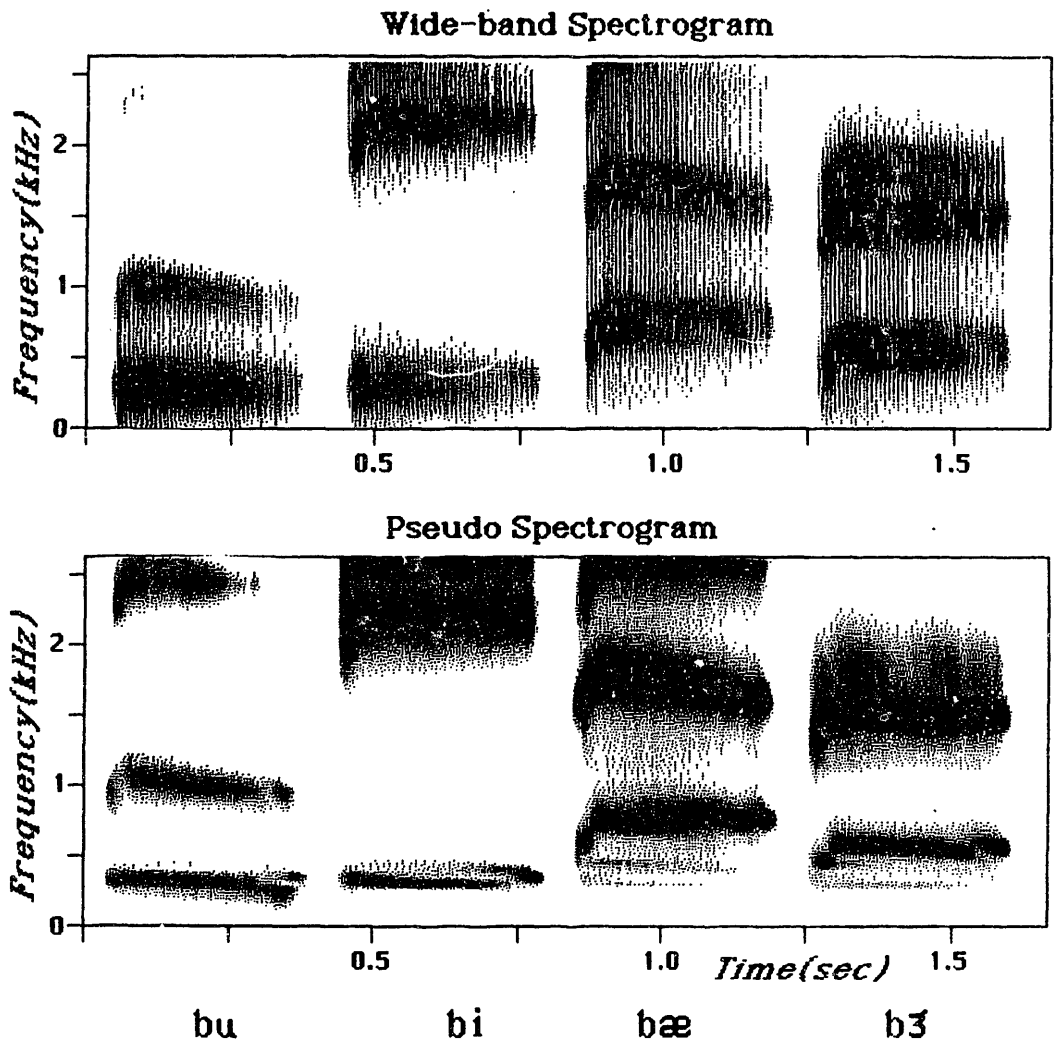

Figure 9.1: Wide-band spectrograms compared with pseudo spectrograms for a series of synthetic CV's.

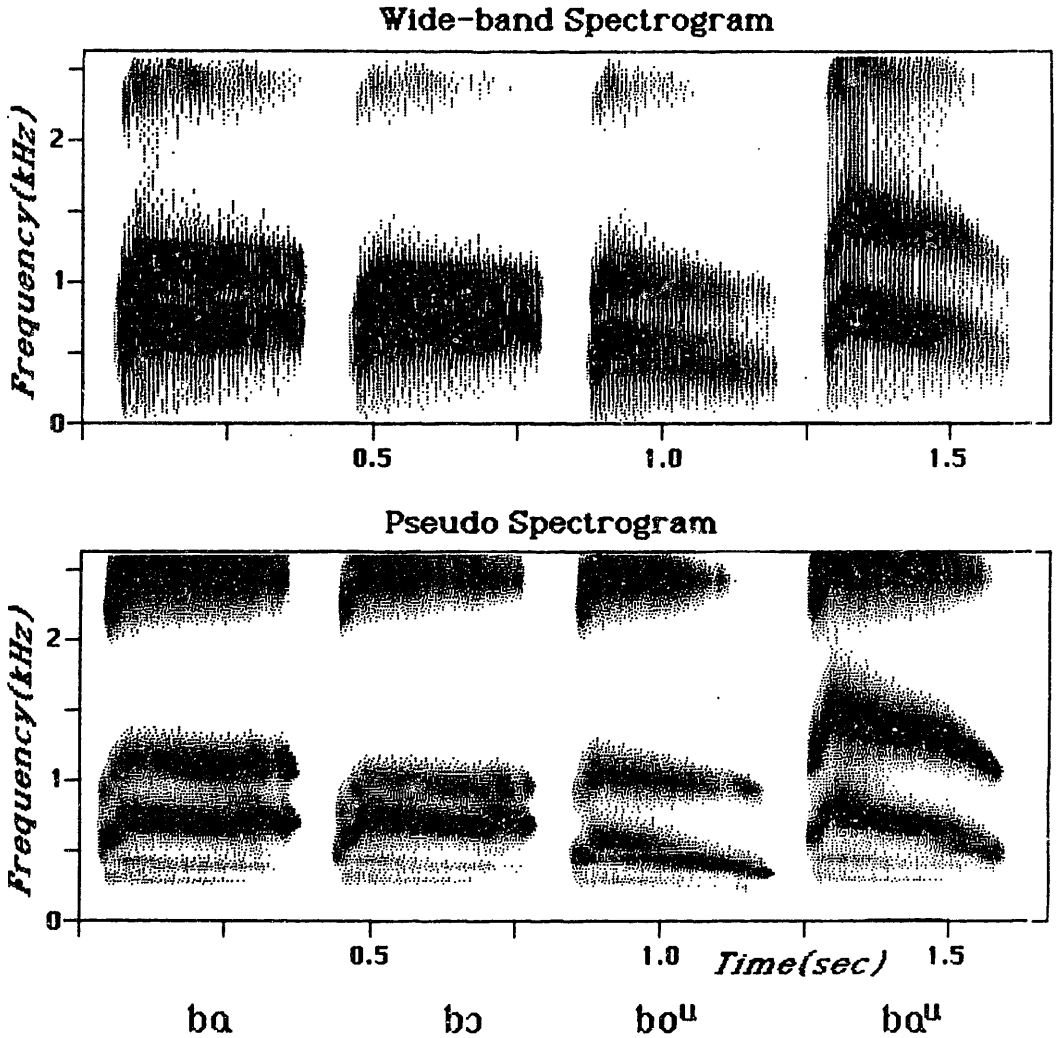

Figure 9.1, continued.

Fixed Settings for Amplitude and Pitch

TIME.	0.0	0.03	0 07	0 38	0.4 sec
AMP	0	0.	70.	62.	0. dB

TIME.	0.0	0.3	0.4 sec
FO	150	130.	100. Hz

Settings for Formants for Individual Tokens

bu

TIME.	0 03	0.1	0.4 sec
F1	250	300.	280. Hz
F2	800.	1000	900. Hz
F3.	2000	2500	2400. Hz

bæ

TIME.	0 03	0.1	0.4
F1:	350	750.	750.
F2:	1400.	1800.	1500
F3:	2000	2700.	2700.

bɔ

TIME.	0.03	0.1	0.4
F1	350.	680	680
F2	600	950.	950.
F3	2000.	2500.	2500.

bi

TIME:	0.03	0.1	0.4
F1:	280.	320.	320
F2:	1800.	2200.	2200.
F3:	2300.	2700	2700.

bɒ

TIME:	0.03	0.1	0.4
F1	300.	550.	350.
F2	700.	1050.	900
F3	2000.	2500	2400.

bɜ

TIME:	0.03	0.1	0 4
F1:	250.	550.	550
F2:	1200.	1500.	1500
F3:	1600.	1850.	1850

bɑ

TIME.	0.03	0.1	0.4
F1.	400.	700	700.
F2	700	1100	1100.
F3	2000.	2500	2500.

bau

TIME:	0.03	0.1	0.3	0.4
F1:	400.	800.	600.	450.
F2:	800	1500	1300	1000.
F3:	2100.	2600.		2500

Table 9.1: Acoustic Parameters for Synthetic CV Syllables. Bandwidths were fixed in all cases at 50, 70, and 110 Hz, respectively, for the first three formants.

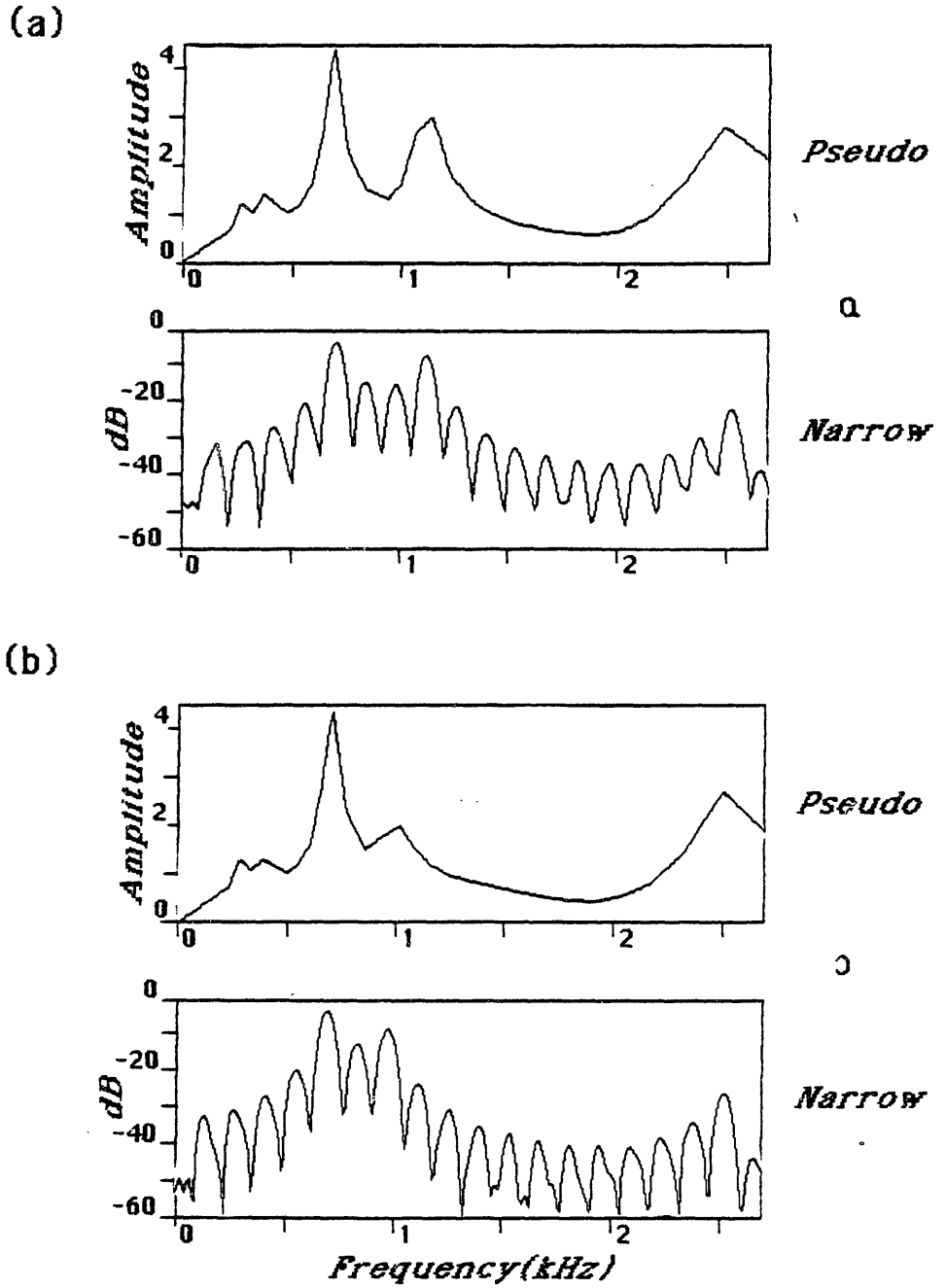

Figure 9.2: Comparison of pseudo spectral analysis of /ɔ/ in "bɔ" with /ɑ/ in "bɑ".

It is usually the case for naturally spoken /r/'s and /ɝ/'s that F_3 does not show up as a separate peak.

Figure 9.4 shows wide-band and pseudo spectrograms for three synthetic CVC syllables, with /b/ and /g/ as the C's, and with the short vowel /I/, /ɛ/, or /ʌ/ as the V. The specifications for these stimuli are given in Table 9.2. In the pseudo spectrogram, the first formant seems to jump discontinuously at the end of the syllable, especially for 'beg' and 'bug'. The upward motion in F_2 at the end is well captured by the pseudo spectral analysis. As in the case of the long vowels, spectral cross sections during the vowel portion are shown in Figure 9.5, with narrow-band analysis as the reference spectrum. Again, the pseudo spectrum shows two prominent peaks at the first and second formant frequencies.

9.2 Synthetic CVC in Noise

It is important to know whether the pseudo spectral analysis process is sensitive to noise added to the signal. To this end, we examined the effect of adding noise to a synthetic vowel stimulus, and the results are given in Figures 9.6 and 9.7. The stimulus, similar to the CVC stimuli above, is the word "bag", with control parameters as indicated in Table 9.3. The noise was created as white noise initially, and then deemphasized to approximate the 6dB per octave falloff that is characteristic of the source and radiation spectrum for voiced speech. The signal-to-noise ratio (SNR) was defined as the ratio of the RMS energy of the deemphasized noise to the RMS energy of the un-preemphasized speech.

Noise was added to the signal in 5dB increments, to obtain 5 tokens varying from 15dB SNR to -5dB SNR. These, together with the clean signal, were processed through pseudo spectral analysis to obtain the pseudo spectrograms in Figure 9.6. Pseudo spectral cross-sections are compared with narrow-band spectra, measured at 200 ms into the vowel of each token, in Figure 9.7. The +5 dB SNR signal sounded unintelligible, although the other stimuli could probably be identified at least as /æ/ if not as "bag". Even the +5 dB SNR signal yielded peaks near the formant frequencies in the pseudo spectrum, in spite of the fact that such peaks are relatively obscure in the narrow-band spectral representation. The remaining stimuli show remarkable consistency of the pseudo spectral shape, in spite of the noise.

9.3 Natural Speech

Synthetic speech is useful for quantifying certain aspects of a processing method for well-specified data. However, it is probably even more important to evaluate the performance of the system on natural speech, which contains complexities that are never completely captured by synthetic stimuli. Figures 9.8 and 9.9 show wide-band and pseudo spectrograms for a series of CVC's spoken by a male and a female speaker respectively. The context for the male tokens is hVd, and for the female dVs. The rapid motion of F_2 is easily captured by the pseudo spectrogram [see, for example, the female dUs]. In the case of the female speaker, however, the first formant rises

Settings for Amplitude and Pitch

TIME:	.04	.06	.08	.15	.38	.40 sec
AMP:	0.	48.	62.	60	59.	0. dB

TIME:	0.0	0.3	0.4	
F0:	150.	130.	100.	Hz

Settings for Formants

TIME:	.04	.09	.33	.40 sec
F1:	350.	700.	700.	400. Hz
F2:	1400.	1650.	1650.	2000. Hz
F3:	2200.	2500.	2500.	2100. Hz

Table 9.3: Acoustic Parameters for Synthetic /æ/-like Stimuli Used in Study of Effects of Noise.

into the vowel in a staircase fashion. This effect is a consequence of the narrow critical band filters in this region, which tend to resolve isolated harmonics for females. Thus the spectral peak is a single harmonic of the pitch, which then jumps discontinuously to the next higher harmonic as the formant resonance frequency rises.

There are also additional peaks below F_1 at multiples of the fundamental frequency of voicing. Such extra peaks below F_1 are characteristic of the system for female speech. It could be viewed as a major problem if, for example, template matching on these pseudo spectra were to be attempted as a recognition strategy. Clearly, the fundamental frequency can vary greatly without changing vowel identity, although such changes would be reflected in a very different peak pattern below F_1. It is possible, however that a template, or alternatively, a spectral weighting function, could be applied to the region between F_1 and F_2, detecting the stable low valley that usually separates the two formants. The pseudo spectral analysis typically suppresses harmonics between F_1 and F_2, even for female speech, and thus spectral patterns in this region may be reliable.

Further data on this effect, including examples of spectral cross sections, are shown in Figure 9.10, an analysis of the word "majority", spoken by a female speaker. Here the individual harmonics below F_1 are quite pronounced on the pseudo spectrogram. Cross sections at the center of four of the voiced phonemes are given below the spectrograms. The /ʌ/ shows no additional harmonics below F_1. However, the /a/, with a high first formant frequency, has two prominent peaks below F_1, but a clear distinct valley between F_1 and F_2. Likewise, the remaining examples, /r/, and /i/, both show extra peaks below F_1, but prominent valleys between F_1 and F_2. The rapid motion of

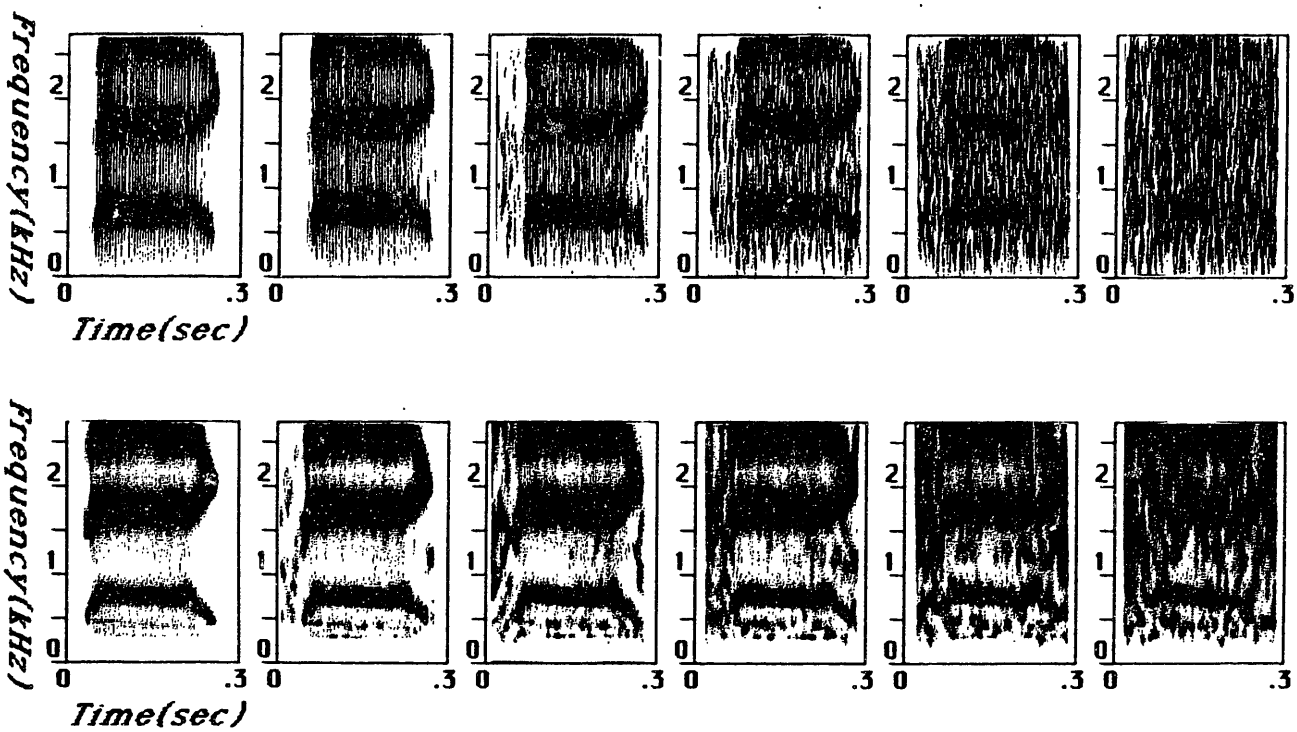

Figure 9.6: Wide-band spectrogram [top] compared with pseudo spectrogram for synthetic word "bag", with white noise added at levels increasing by 5 dB increments. At the far left is the clean speech; the second sample has a 15 dB SNR. The noise level increases by 5dB for each succeeding sample, up to -5dB SNR in the final sample.

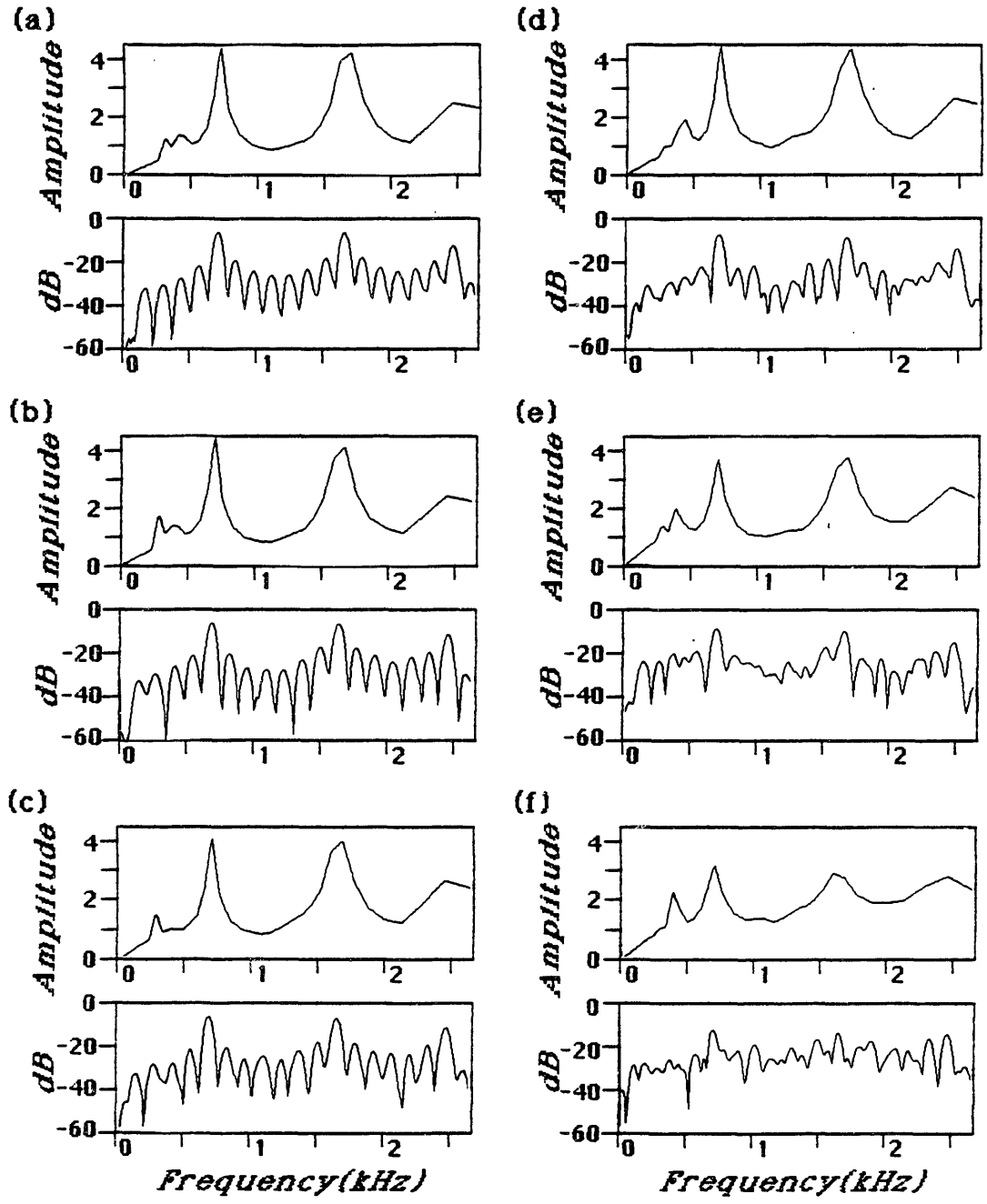

Figure 9.7: (a) through (f):
Pseudo spectra [top] compared with Narrow-band spectra at a time slice 200 ms into the vowel, for each of the samples of noisy synthetic speech described in Figure 9.6. a) clean speech. (b) through (f): Level of noise increasing from 15dB SNR to -5dB SNR in 5dB increments of noise level.

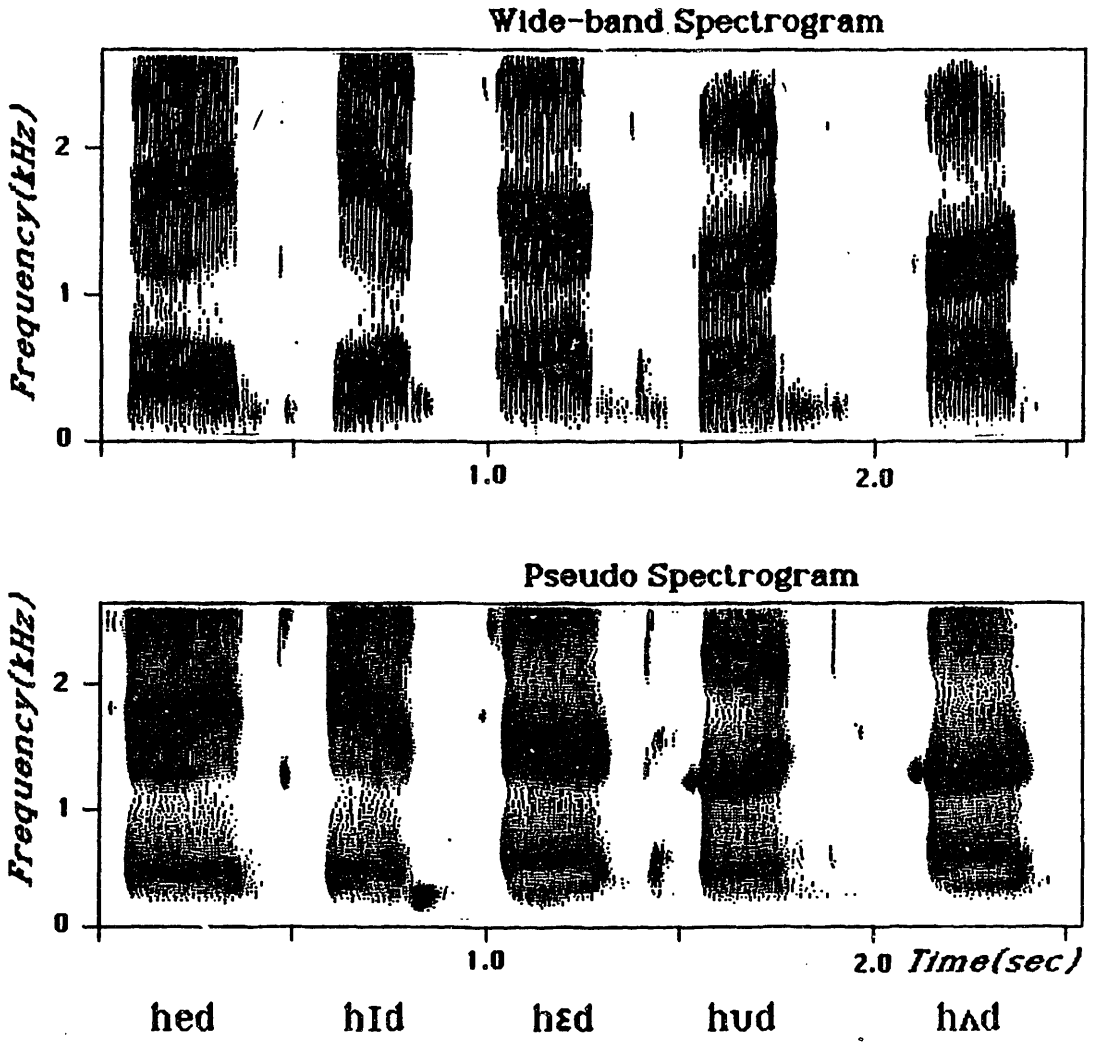

Figure 9.8: Wide-band spectrograms compared with pseudo spectrograms for series of hVd's spoken by a male speaker. The frequency scale is 20% smaller in the pseudo spectrograms.

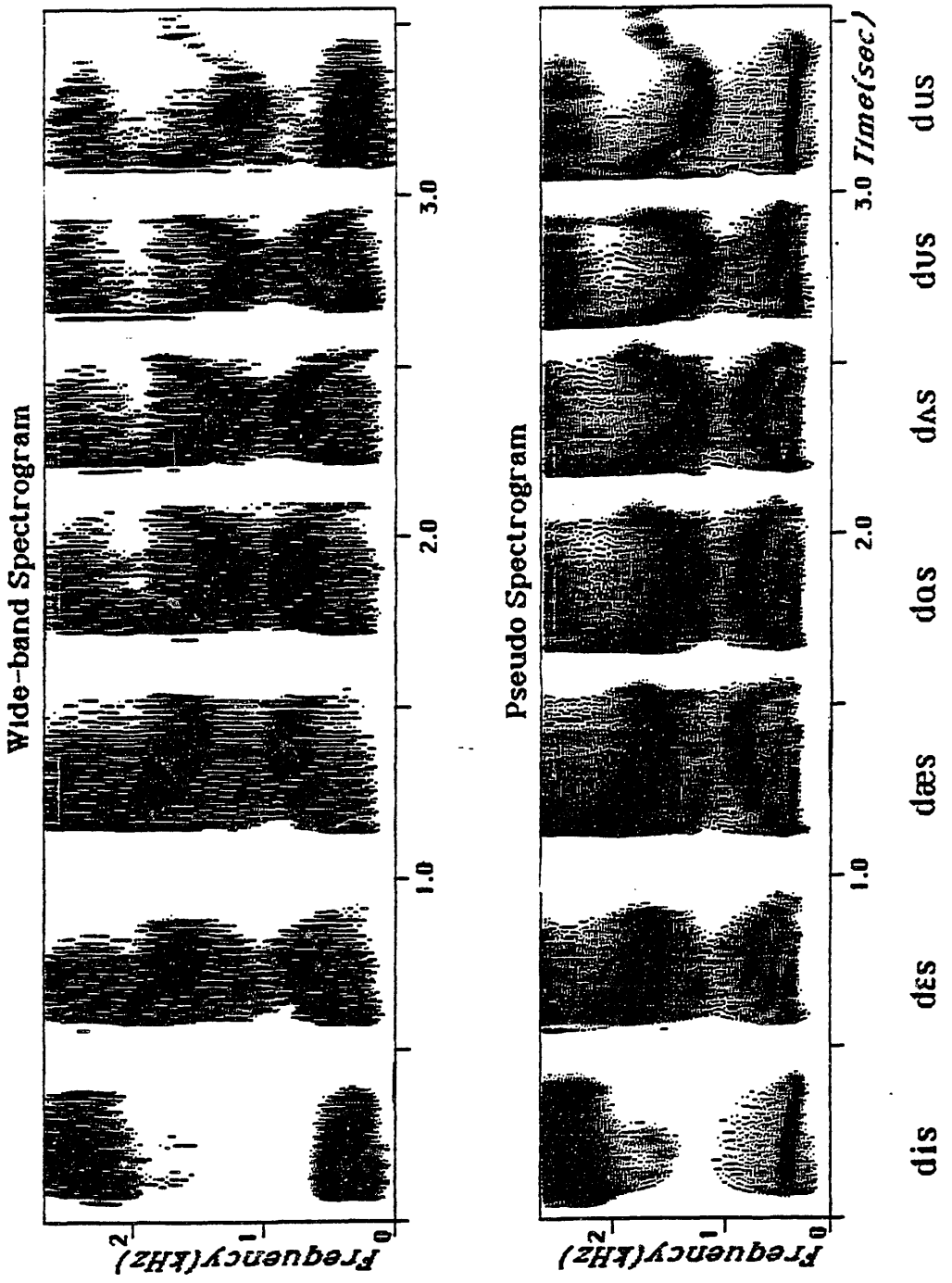

Figure 9.9: Wide-band spectrograms compared with pseudo spectrograms for series of dVs's spoken by a female speaker. The frequency scale is 20% smaller in the pseudo spectrograms.

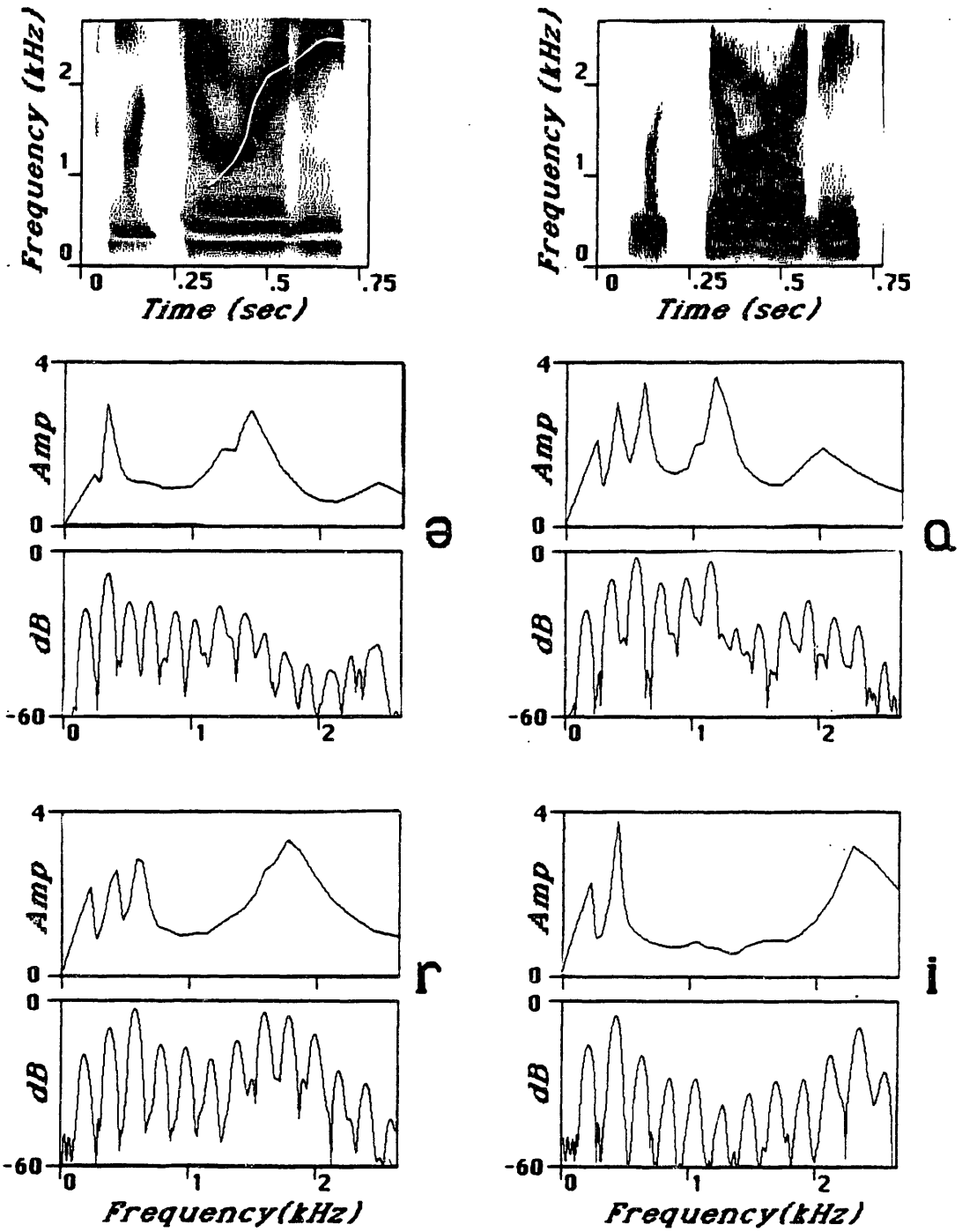

Figure 9.10: Illustration of pseudo spectral analysis of female speech.
a) Wide-band spectrogram compared with pseudo spectrogram for word "majority", spoken by a female speaker, illustrating resolved harmonics in first formant region.
b) Comparison of narrow-band spectra with pseudo spectra at selected time samples in the word.

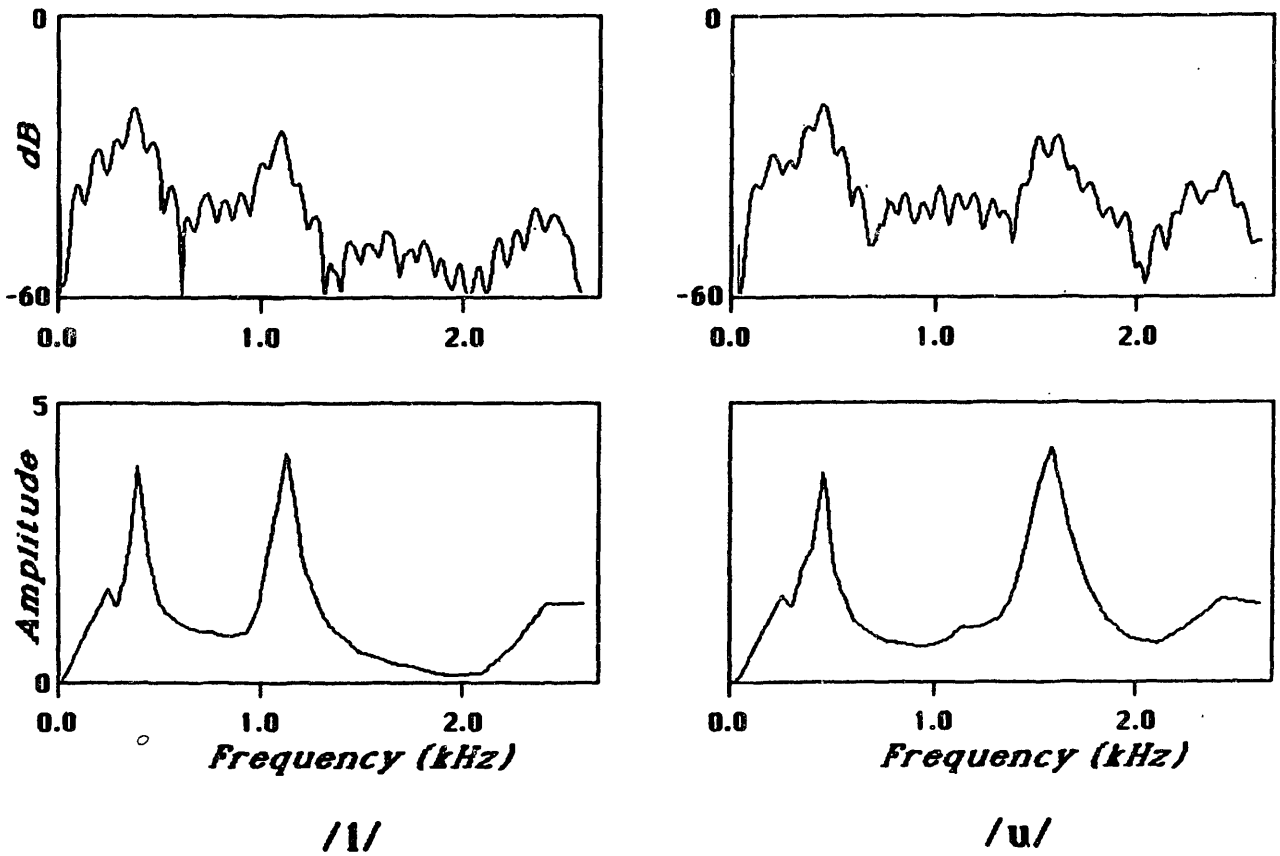

Figure 9.11: Two examples of narrow-band spectra and pseudo spectra for time-slices in the word "blue", spoken by a male speaker.
a) Cross-sections in the /l/, and
b) Cross-sections in the /u/.

F_2 throughout the stressed syllable is captured well by the pseudo spectral analysis procedure.

An example of two cross-sections for male speech is given in Figure 9.11, which shows a time-slice in the /l/ and a time slice in the /u/ of the word "blue". Here the pronounced valleys separating the formant peaks are evident. Ideally, one would like to see spectral representations such as this for all vowels, since the formant frequencies convey most of the identity of the linguistic content. However, the first formant region is often complicated by additional features such as breathiness and nasalization. Such features must be manifested by more complex structures in the F_1 region, whose exact nature remains to be determined.

An example showing how the pseudo spectrum deals with rapid formant motion is given in Figures 9.12 and 9.13. Part (a) of Figure 9.12 shows a wide-band spectrogram of the word "wish", spoken by a male speaker. Part (b) shows a sequence of narrow-band spectral cross-sections, taken at 10 ms intervals in the region between the two vertical bars in part (a). Over this interval, the second formant is rising rapidly in transition from the low frequency appropriate for /w/ to the high frequency appropriate for /I/. The upward movement of F_2 is visible from the envelopes of the sequence of spectra, but there is a substantial amount of intereference from the excitation harmonic structure.

Figure 9.13 shows the results of pseudo spectral analysis of the same word. Part (a) of the Figure shows the pseudo spectrogram along with the wide-band spectrogram. Part (b) shows a series of pseudo-spectral cross-sections taken at the same time intervals as in Figure 9.12. The rapid movement of the peak at the second formant frequency is captured well by the pseudo spectral analysis. The first formant is also prominent, and, there is an additional peak below the first formant, at the second harmonic of the pitch.

Figure 9.14 shows pseudo spectral analysis of the sentence "He ordered peach pie with ice cream", spoken by a male speaker. Part (a) shows the pseudo spectrogram compared with a wide-band spectrogram, and part (b) shows several cross sections during voiced segments, compared with narrow-band spectra. In several of the pseudo spectral cross-sections, there is an extra peak below the first formant peak, usually at the second harmonic of the pitch. This peak, which is probably related to a glottal formant, is also evident in the narrow-band spectrum, although its prominence has been enhanced by the synchrony analysis. The /i/ of "cream", [part vi] has extra peaks between F_1 and F_2. These additional peaks are a consequence of nasalization of the vowel. Again, their prominence has been enhanced in the pseudo spectrum relative to the narrow-band spectrum.

The phenomenon of an extra peak at the second harmonic of the pitch is quite common in pseudo spectral analysis of vowels with a high first formant. This phenomenon is illustrated in Figure 9.15, for the word "topic", spoken by a female speaker. Part (a) of the Figure shows standard spectral analysis, and part (b) shows the pseudo spectrogram and pseudo spectrum in the /a/. Although in this case the second harmonic is weaker than the third harmonic in the narrow-band spectrum [Figure 9.15a], its prominence has been enhanced by the synchrony analysis such that it is almost as large as the first formant peak [Figure 9.15b]. The third harmonic is well reduced in amplitude relative to its neighbors. Probably, the second harmonic is sufficiently remote from the formant

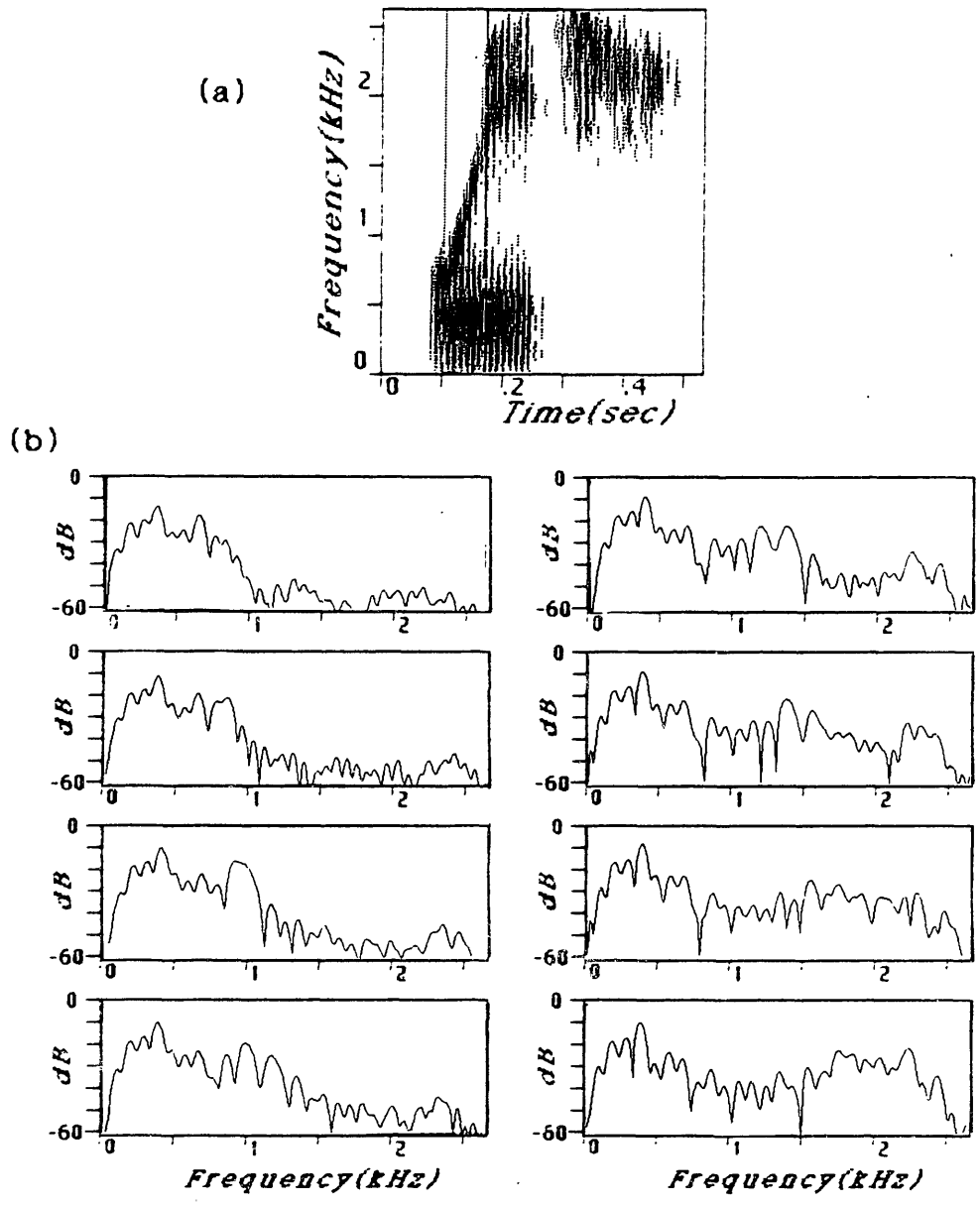

Figure 9.12:
a) Wide-band spectrogram of word "wish", spoken by a male speaker.
b) Series of narrow-band spectra taken at 10 ms intervals between the two vertical bars on the spectrogram in part (a). Time increases from top to bottom and then from left to right.

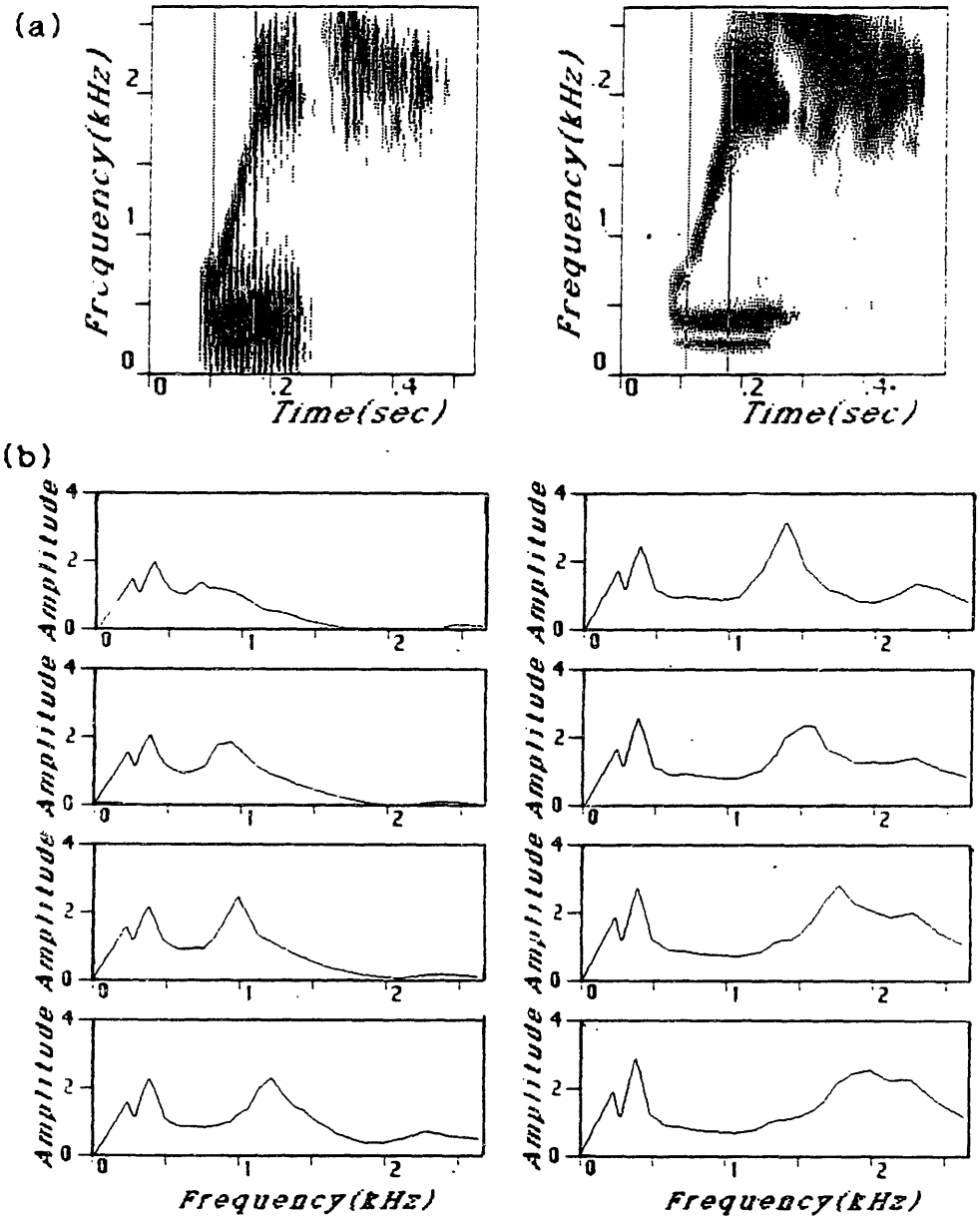

Figure 9.13: Example showing rapid formant motion
a) Wide-band spectrogram compared with pseudo spectrogram for word "wish".
b) Series of pseudo spectral cross-sections taken at 10 ms intervals at time slices between the two vertical bars in the spectrograms. Time increases from top to bottom and then from left to right. This result should be compared with the results for narrow-band analysis given in Figure 9.12.

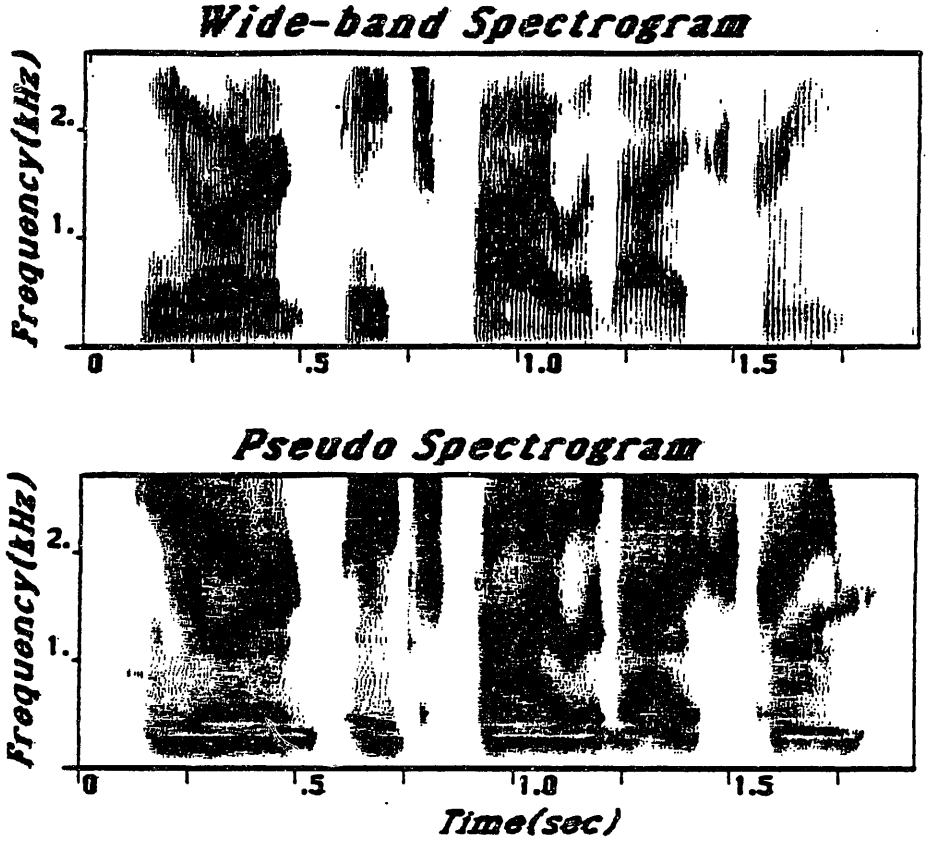

Figure 9.14: Comparison of pseudo spectral analysis of the sentence, "He ordered peach pie with ice cream", spoken by a male speaker, with standard spectral analysis.
a) Wide-band spectrogram [top] compared with pseudo spectrogram of the entire sentence.

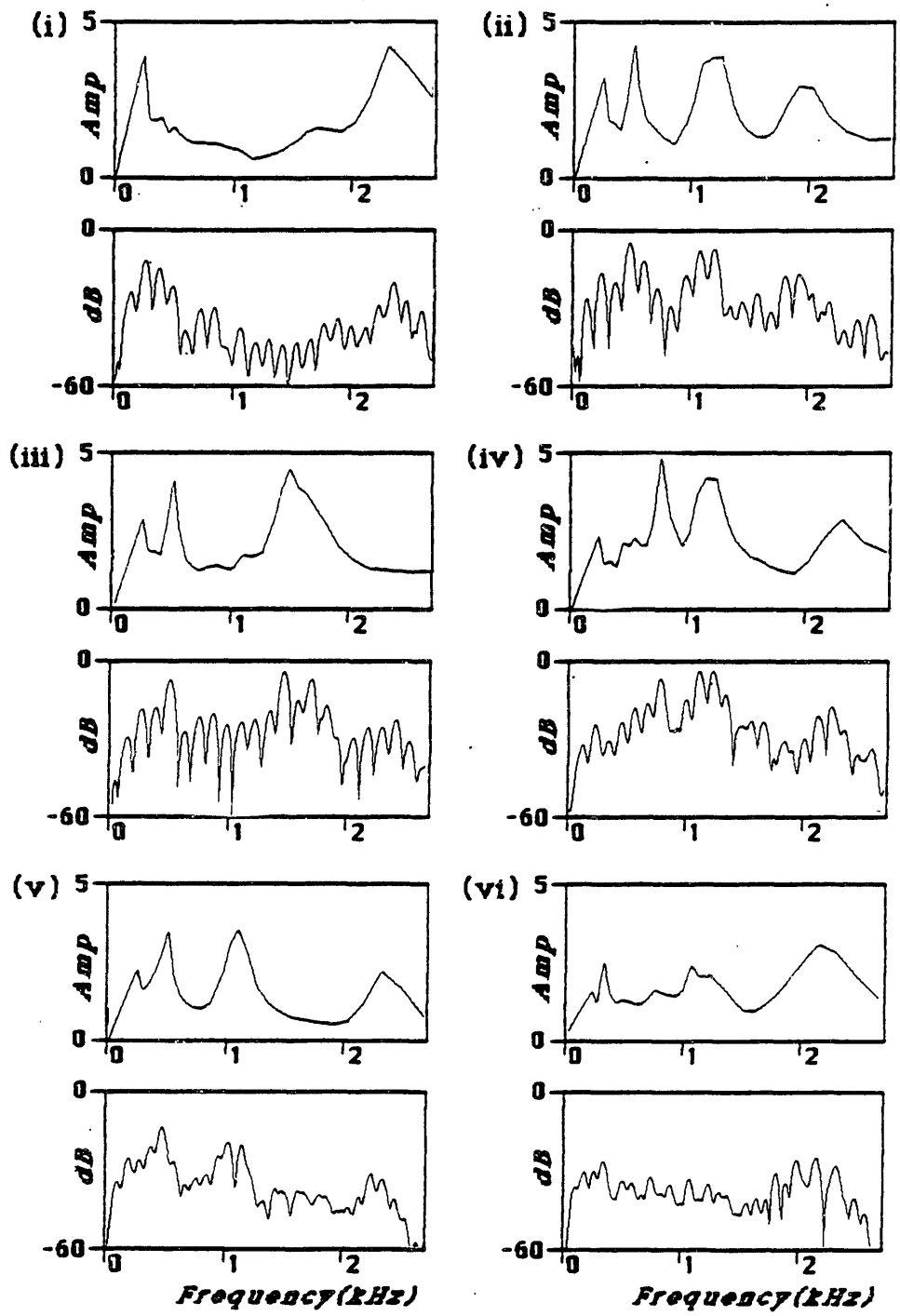

Figure 9.14:
b) Comparison of several pseudo spectral cross-sections with corresponding narrow-band spectrum at same place in time.
i) /i/ in "He". ii) /ɔ/ in "ordered". iii)) /ɜ˞/ in "ordered". iv) /aʊ/ in "pie". v) /w/ in "with". vi) /i/ in "cream".

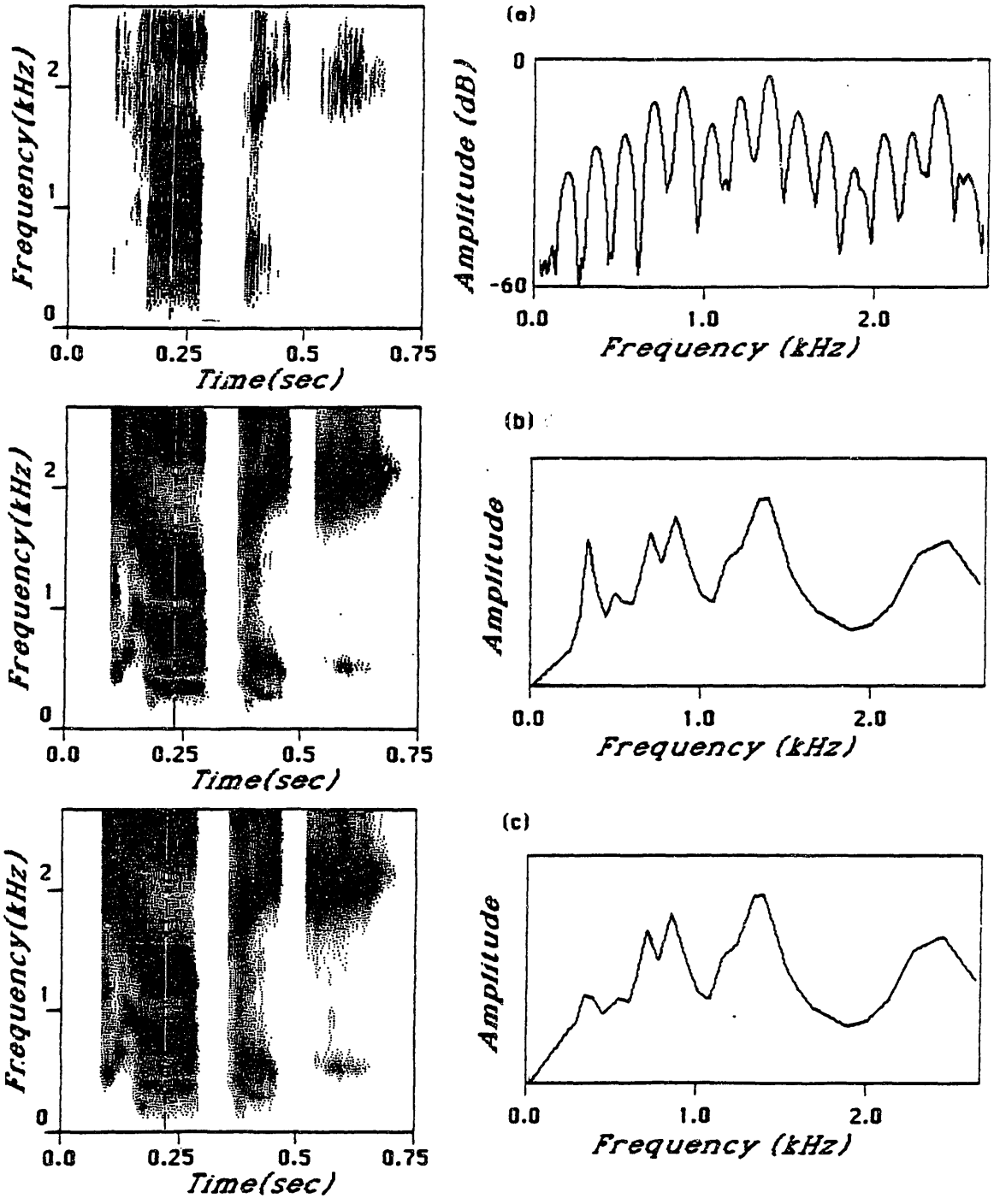

Figure 9.15: Illustration of prominent second harmonic in low vowels in pseudo spectrum.
a) Wide-band spectrogram of word "topic", spoken by a female speaker, and narrow-band spectrum at time slice during the /a/.
b) Pseudo spectrogram for same word as in (a), and pseudo spectrum for same time slice, showing prominent peak at second harmonic of fundamental frequency.
c) Same as in (b), except low frequency filters have been modified as in Figure 9.16.

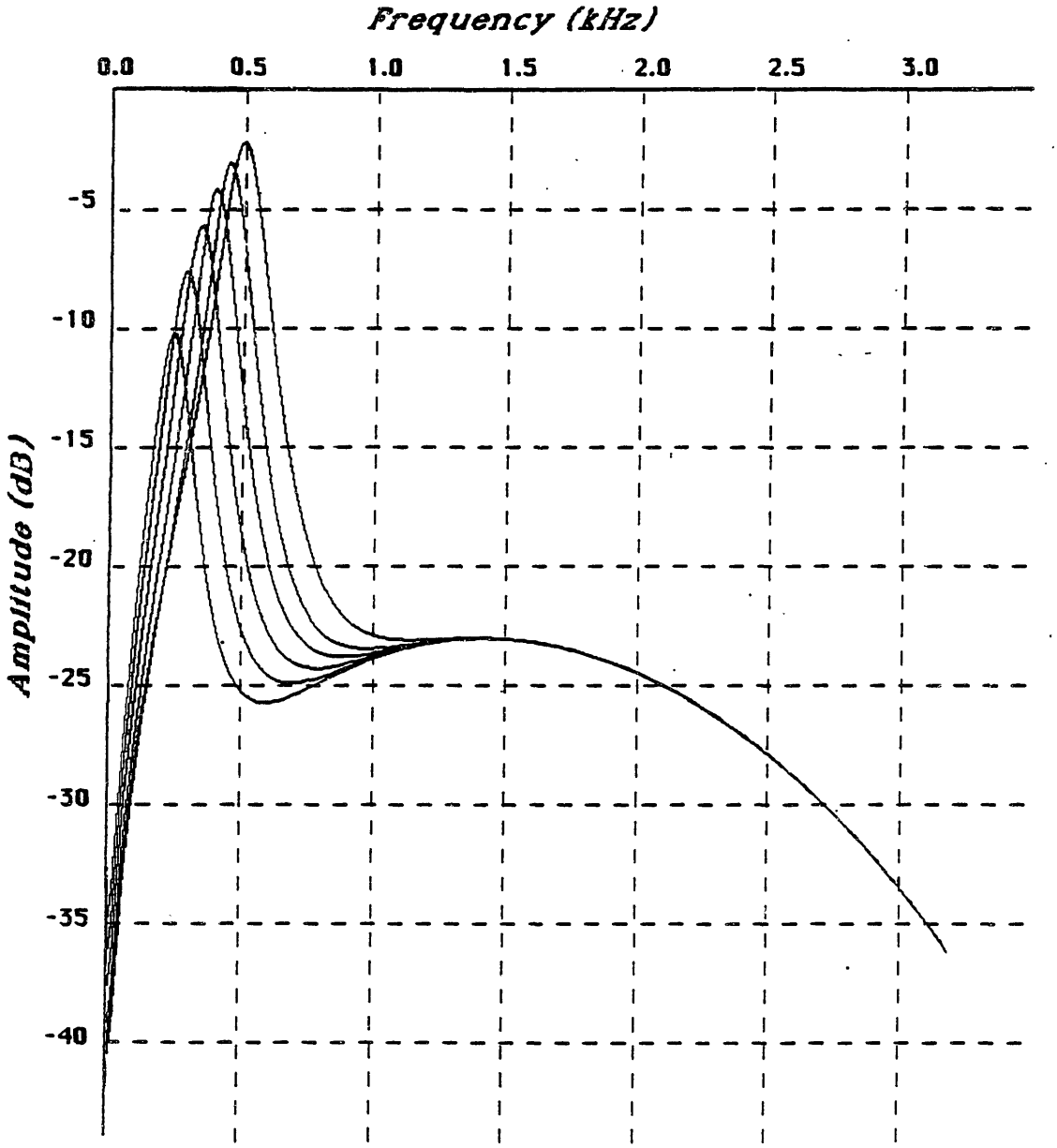

Figure 9.16: Filter characteristics of low frequency filters used for part (c) of Figure 9.9.

that the narrow filter picks up a relatively pure sine wave.

The prominence of the second harmonic can be reduced substantially by adding high-frequency tails to the low frequency filters, as shown in Figure 9.15c. This was accomplished by adding an attenuated, broad-band filtered original waveform to the final filter output, at a level about 15 dB below the peak of the response curve. The resulting filter characteristics are shown in Figure 9.16. This type of strategy is to be preferred over broadening the definition of critical bandwidth, because such tails may actually exist in the auditory system [Kiang, personal communication]. However, it is unlikely that the tails are as high in amplitude as the tails that were introduced for this experiment.

9.4 Breathy Speech

In certain languages, such as Gujarati [Dave, 1977], there is a phonetic distinction between breathy and non-breathy speech. The acoustic correlates of breathy speech are not yet fully characterized. However, there is typically a strong component at the fundamental frequency of voicing in the source spectrum. In addition, the source tends to contain a noise component as well as the periodic component.

Although the frequency of the fundamental is often below the center frequency of the first filter in the pseudo spectrum, a prominent fundamental may be picked up on the tails of the filters in the F_1 region, resulting in a substantial loss of synchrony to the center frequency of the filter. Figure 9.17 shows examples of three contrastive CV sequences in Gujarati, "bi" versus breathy "bi", "bar" versus breathy "bar", and "go" versus breathy "go". The pseudo spectrograms are shown in part (a) of the Figure, and the wide-band spectrograms are given for comparison in part (b). Selected spectral cross-sections at the points indicated by the vertical bars in the spectrograms are shown in Figure 9.18, where comparisons are made between pseudo spectra and narrow-band spectra.

In comparing the two "bi" stimuli, the pseudo spectrum has a very weak first formant peak for the breathy "bi", as contrasted with the non-breathy "bi". In the case of "bar", the situation is not quite as clear. The breathy "bar" does not have additional harmonics below F_1, in contrast to the normal "bar". The first formant amplitude is fairly strong during the first part of the breathy vowel, but becomes much weaker towards the end. In addition, irregularities due to the noisy nature of the source are evident. Both the normal "go" and the breathy "go" have a relatively weak first formant amplitude. However, the normal "go" includes a rather prominent response at the second harmonic of the fundamental, which is missing in the breathy "go". As in the other two cases, there is a trend towards a weakening of the response in the low frequency region, probably as a consequence of the presence of the strong fundamental.

9.5 Discussion

Through selected examples of natural and synthetic speech, we have tried to convey an impression of the characteristics of the pseudo spectrum. Basically, the system maps spectral peaks into

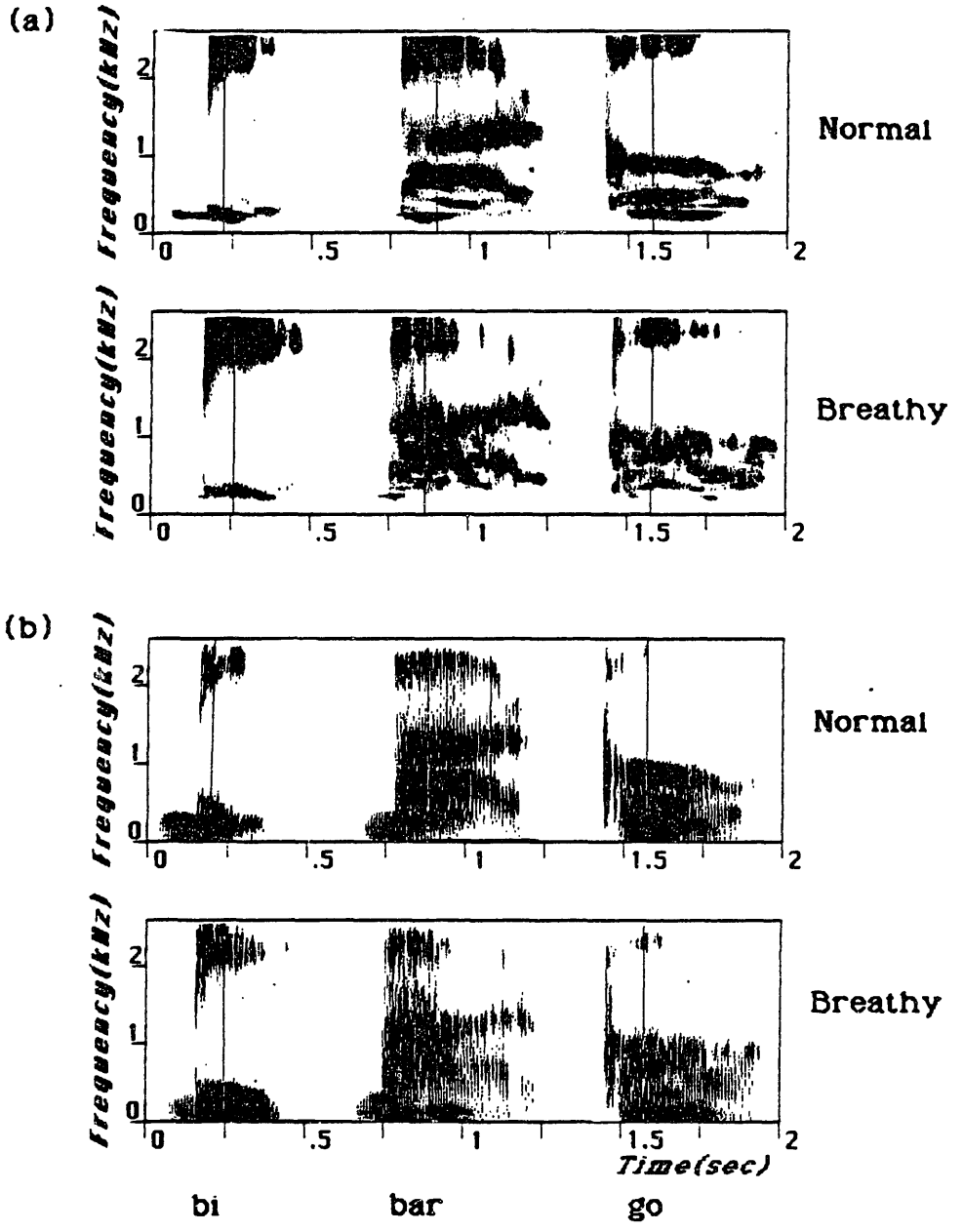

Figure 9.17: Illustration of normal versus breathy speech, a phonetic contrast that exists in the Gujarati language [Dave, 1977].
a) Pseudo spectrogram of normal versus breathy "bi", "bar", and "go", spoken by native Gujarati speakers.
b) Wide-band spectrograms for the same data as in part (a).

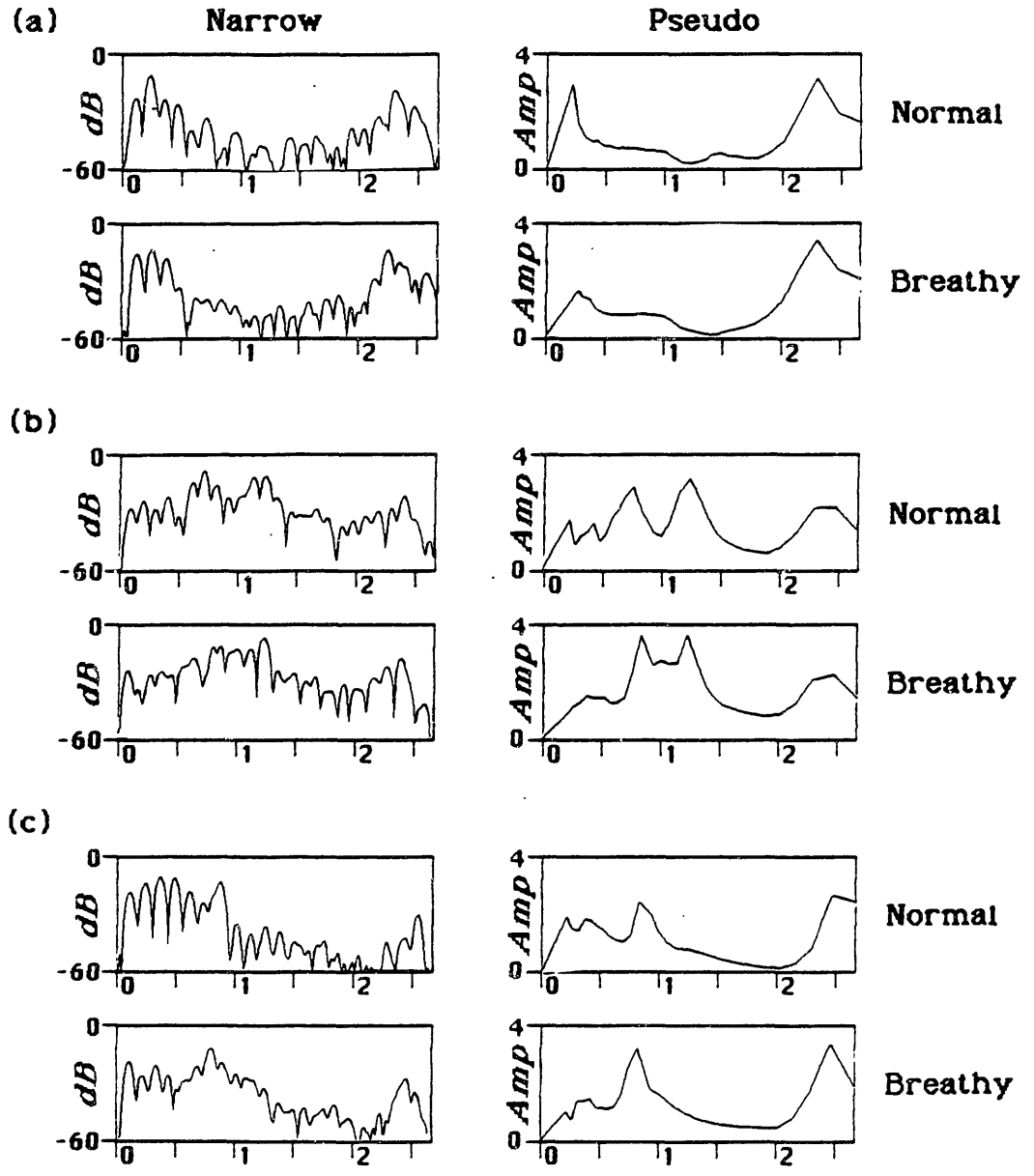

Figure 9.18: Comparison between narrow-band and pseudo spectra for breathy and non-breathy Gujarati vowels at the time slices indicated by the vertical bars in the spectrograms in Figure 9.17.
a) "bi" versus breathy "bi".
b) "bar" versus breathy "bar".
c) "go" versus breathy "go".

pseudo spectral peaks, but the mapping process is highly nonlinear. The level of response to a given peak depends upon its prominence relative to a local environment. The dependence is also strongly related to the detailed shape of the critical band filter centered at that frequency. Thus the presence of a strong peak in the spectrum well below the center frequency of a given filter may affect the response level at that center frequency, depending upon the extent to which information at the frequency of the spectral peak is transmitted by the filter centered above it.

The pseudo spectrum produces a much simpler response characteristic in the F_2 region than in the F_1 region. This result is due in part to the narrow auditory filters in the low frequency region, but also to the fact that speech itself is more complex in the F_1 region. Not only must information about the frequency of the formant be conveyed to some decoding center, but also sufficient information to determine whether the vowel is nasalized, or breathy. It may be no accident that major cues to these two features are carried by the $F1$ region. It is also possible that the sex of the speaker is determined in part from the presence or absence of distinct harmonics in the first formant region.

There are several ways in which the pseudo spectrum could be modified so as to reduce the response to the individual harmonics in the first formant region. One possibility, certainly, is to manipulate the filter characteristics to be more broad-band. This is unattractive, because there is little evidence to support such a modification in the data on auditory filters. In addition, the resulting loss of synchrony would affect the formant itself as well as the "unwanted" harmonics. Another possibility is that the model that implies a measure of perfect synchrony to the center frequency of the filter is unrealistically accurate. If some noise were allowed in the delay time that was used for the difference computation, then the measure of synchrony would become somewhat fuzzier, which might reduce the sensitivity of filters tuned to individual harmonics below the first formant frequency.

On the other hand, a representation of the first formant region by a series of distinct harmonics at this stage of analysis may not be an "incorrect" result. It is possible that a later stage extracts the individual peaks and decodes them as being harmonically related. They may then be combined using some weighting scheme to produce an estimate for an underlying formant frequency, not necessarily at the frequency of any one harmonic. Another possibility is the computation of a center-of-mass in the first formant region to represent the underlying formant peak. Only after attempts have been made to identify phonemes based on some as yet unspecified further processing of the pseudo spectrum will it become clear whether the presence of such individual harmonics poses major difficulties to speech recognition. This first formant region remains somewhat puzzling, and is clearly an area for further research.

Chapter 10

Experiments with Synthetic Stimuli and Relationship to Psychophysics

10.1 Introduction

In this chapter we will examine the pseudo spectrum obtained from a variety of different synthetic stimuli that are varied systematically, in order to better quantify some aspects of the system. The advantage of using synthetic stimuli over natural stimuli is that parameters can be manipulated at will, and the spectral content of the stimulus is then defined explicitly. We will show how the pseudo spectrum changes as a function of changes along a continuum in some dimension such as the frequency or amplitude of a given formant. Whenever possible, the results will be discussed in the light of available psychophysical data from the literature.

We begin with a series of steady state vowel stimuli, generated using the Klatt parallel synthesizer [1980]. The effect of changing the amplitude of F_2 is examined, with all other parameters of the synthesizer fixed. Results are given for two vowel series; one with the formants widely spaced, and the other with F_1 and F_2 differing by only 300 Hz. The results are related to the study by Chistovich et al [1979] on the perception of vowel-like stimuli with differing formant amplitude relationships. This study suggests that there are some distinct differences between the perception of vowels with formants spaced within a critical distance and vowels with formants well separated.

The next section examines the pseudo spectrum for a series of synthetic /æ/-like stimuli, each of which represents some small perturbation from a reference /æ/ stimulus. The results are interpreted with reference to a study on vowel perception by Carlson, Granstrom and Klatt [1979], which makes use of a similar set of /æ/-like stimuli. The perceptual study evaluated perceptual salience of selected attributes of the vowel, such as formant frequency, formant bandwidth, phase characteristics, etc.

The final section concerns another set of stimuli for which perceptual data are available. These stimuli were used for a study of perception of vowel nasalization by Hawkins and Stevens [in press]. The F_1 region is represented by a pole-zero-pole complex, in accord with models for nasalization which include additional resonances and antiresonances due to the opening of the parallel nasal branch. The relative spacings of the poles and zero were manipulated in a systematic way to increase the percept of nasalization. As in the previous case, the results of pseudo spectral analysis of these stimuli are related to the perceptual results.

10.2 Effects of Variations of Relative Amplitudes and Frequencies of F_1 and F_2

In this section we describe the results of manipulating relative formant amplitudes and frequencies for a number of stimulus conditions. There are two main points to this experiment. There is the obvious one of quantifying the relationship between absolute formant amplitudes, as measured using Fourier analysis, and formant amplitudes as defined by the pseudo spectrum. The other is to show how the presence of a proximal formant can strongly bias the amplitude of a given formant in the pseudo spectrum.

The results for a series of stimuli for the synthetic vowel /ʌ/, with formant frequencies at 450, 1450, and 2450 Hz, are shown in Figure 10.1. For these stimuli, F_0 was held fixed at 100 Hz, and the amplitude of voicing was also constant for the duration of the stimulus. The amplitudes of the first and third formants were held fixed, while the amplitude of F_2 was varied in 5dB increments. Part (a) of the Figure shows wide-band spectrograms compared with pseudo spectrograms for the vowel series. In the first vowel of the series, the amplitude of the second formant is very weak in the wide-band spectrogram; nonetheless, it shows up as a prominent dark band in the pseudo spectrogram. The amplitude increases are expressed in the pseudo spectrogram by a widening of the dark band at the second formant resonance frequency. The bands at F_1 and F_3 remain essentially unchanged until the very last vowel of the stimulus, when F_3 becomes considerably weaker.

Cross-sections, taken at the time slice indicated by the vertical bar in the spectrograms, are shown in part (b) of the Figure. There is clearly a monotonic, although nonlinear, relationship between true amplitude of the second formant and amplitude in the pseudo spectrum. The amplitude of the second formant in the pseudo spectrum increases by approximately 50% for a twenty dB increase in the amplitude in the log spectrum. Since 20 dB is about a factor of 10 change in the magnitude on a linear scale, the pseudo spectrum shows a substantially reduced sensitivity to localized level changes.

When I listened to the series, the last stimulus seemed distinct from the rest, sounding more like an /ɝ/ than an /ʌ/. The main difference between the pseudo spectrum for this stimulus and the next-to-last stimulus is the substantial reduction in the amplitude of F_3 at 2450 Hz. This reduction comes about because the energy at the second formant frequency is strong enough to be picked up by the filter centered at the third formant frequency. The presence of this energy distal to the center frequency, f_c, causes a loss in synchrony at f_c, and a corresponding loss in amplitude of the peak. The phone /ɝ/ has a very low third formant frequency, which is quite close to the second formant frequency. Thus there is typically a large concentration of energy near 1500 Hz. Pseudo spectral analysis of /ɝ/ consistently produces a single prominent peak for the two formants. The prominent peak at the F_2 frequency in the last stimulus can presumably assume the role of representing both second and third formant frequencies, because the amplitude of the peak at F_3 is so low.

Figure 10.2 shows a similar study for a steady state vowel with two formant frequencies at 700, and 1000 Hz, approximating an /a/. The amplitude of the second formant is varied in 6 dB

increments, such that F_2 ranges from 12 dB below to 12 dB above F_1. Of the five stimuli, the last two sound distinctly nasalized to me, and the middle one sounds only slightly nasalized. As in the preceding example, pseudo spectrograms are compared with wide-band spectrograms in part (a), and pseudo spectra are compared with narrow-band spectra in part (b). The speech is not preemphasized for the narrow-band spectra, so that the relative amplitudes of the two formants are authentic for the original waveform.

The changes in the amplitude of the second formant peak in the pseudo spectrum are much more pronounced than in the preceding example. In the weakest vowel of the series, F_2 is almost nonexistent as a peak. It is not until the last two vowels of the series that the amplitude of F_2 in the pseudo spectrum becomes greater than the amplitude of F_1. This simple test suggests a perceptual correlate between the relative amplitudes of the first two formants in the pseudo spectrum and nasalization of the vowel. We will return to this topic later in this chapter in the section on synthetic nasalized vowels.

The changes in formant amplitudes in the pseudo spectrum for this stimulus are generally more pronounced than for the above /ʌ/ stimulus, when the formants were better separated in frequency. The pseudo spectrum effectively measures the amplitude of a peak relative to the local environment. If the local environment includes another spectral prominence, then the response to the peak shows greater sensitivity to amplitude variations. The influence is much stronger if the interfering peak is below the given peak than if it is above. Such asymmetry is tied to the asymmetry in the tuning curves.

A third study was devised to show the effect of the distance between F_1 and F_2 on the corresponding amplitudes. For this experiment the Klatt cascade formant synthesizer was used, with bandwidths of the two formants at 50 and 70 Hz respectively. The frequencies of the two formants were set according to Table 10.1. These values were selected because they correspond to center frequencies of the filters of the model. It was felt that an additional variable of distance from formant frequency to filter f_c would cause undue complications in the interpretation of the results. The relative amplitudes of the formants are not fixed, because that would result in an over-specified system. As the separation between F_1 and F_2 grows, the amplitude of the peak at F_2 is reduced, because the peak is riding on a 12dB/octave slope of the F_1 resonance.

Narrow-band spectral cross-sections are compared with pseudo spectra during the middle of each of these steady state vowels in Figure 10.3. In the first vowel, F_2 is only 147 Hz away from F_1. In neither the narrow-band nor the pseudo spectrum does F_2 show up as a separate peak. The peak at F_1 in the pseudo spectrum is considerably sharper than in the narrow-band spectrum; mostly because of a reduction of the levels in the frequency region just above F_1. This effect is a consequence of the presence of substantial energy below f_c in the filters tuned to the second formant. The second vowel has a 200 Hz separation between the two formants, and again there is no distinct peak for F_2. However, at this point there is a clear shoulder above the first formant peak. In the third vowel of the series, the two formants are spaced by 300 Hz, and both appear as peaks in the pseudo spectrum. The amplitude of F_2 in the pseudo spectrum is not significantly less than the amplitude of F_1. In the fourth vowel of the series, the two formants are spaced by

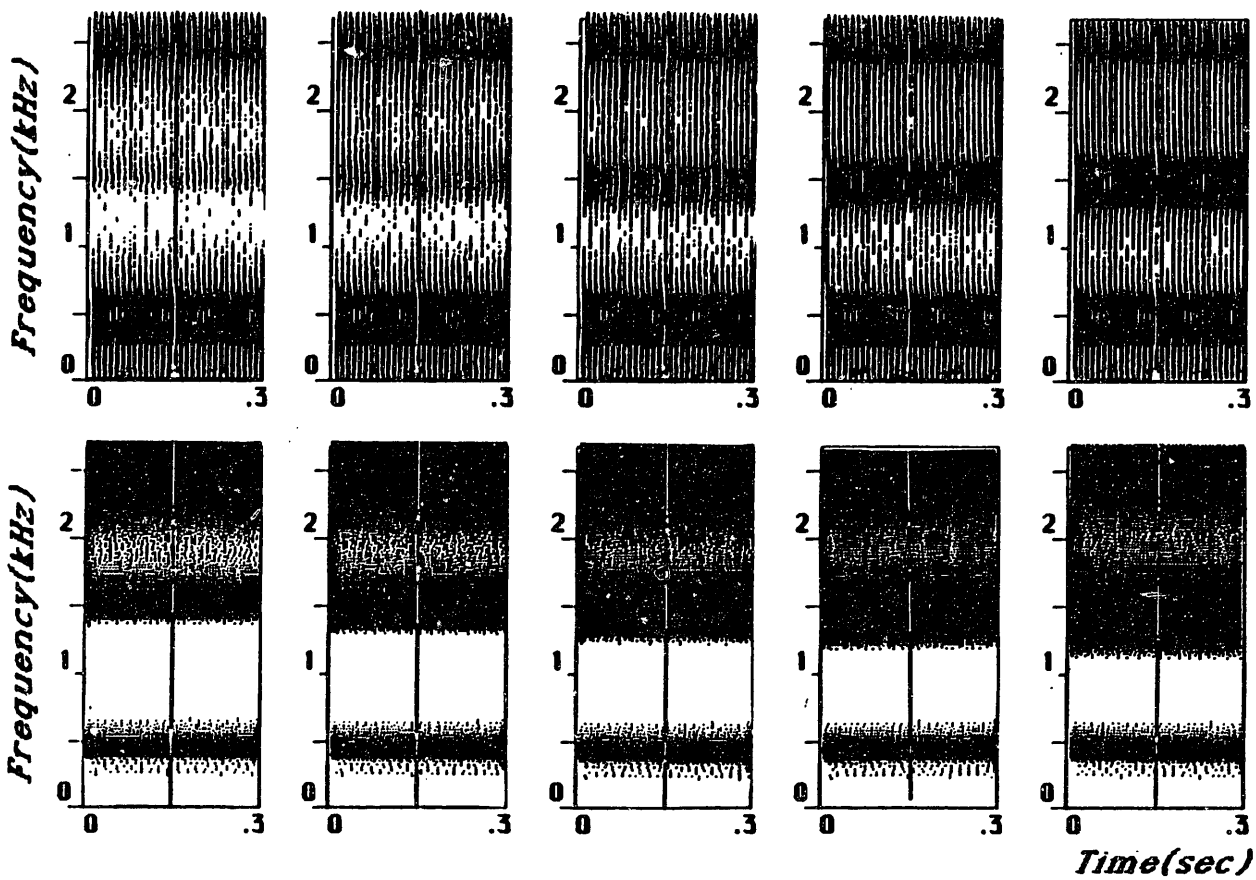

Figure 10.1: Results of modifications in amplitude of the second formant for a vowel with formants well spaced in frequency.
a) Wide-band spectrograms compared with pseudo spectrograms for a series of steady state stimuli with formant frequencies at 450 Hz, 1450 Hz,and 2450 Hz, simulating the neutral vowel. Each vowel in the series differs from the previous by an increase in the amplitude of the second formant peak by 5 dB. The stimuli were created using the Klatt parallel synthesizer [1980].

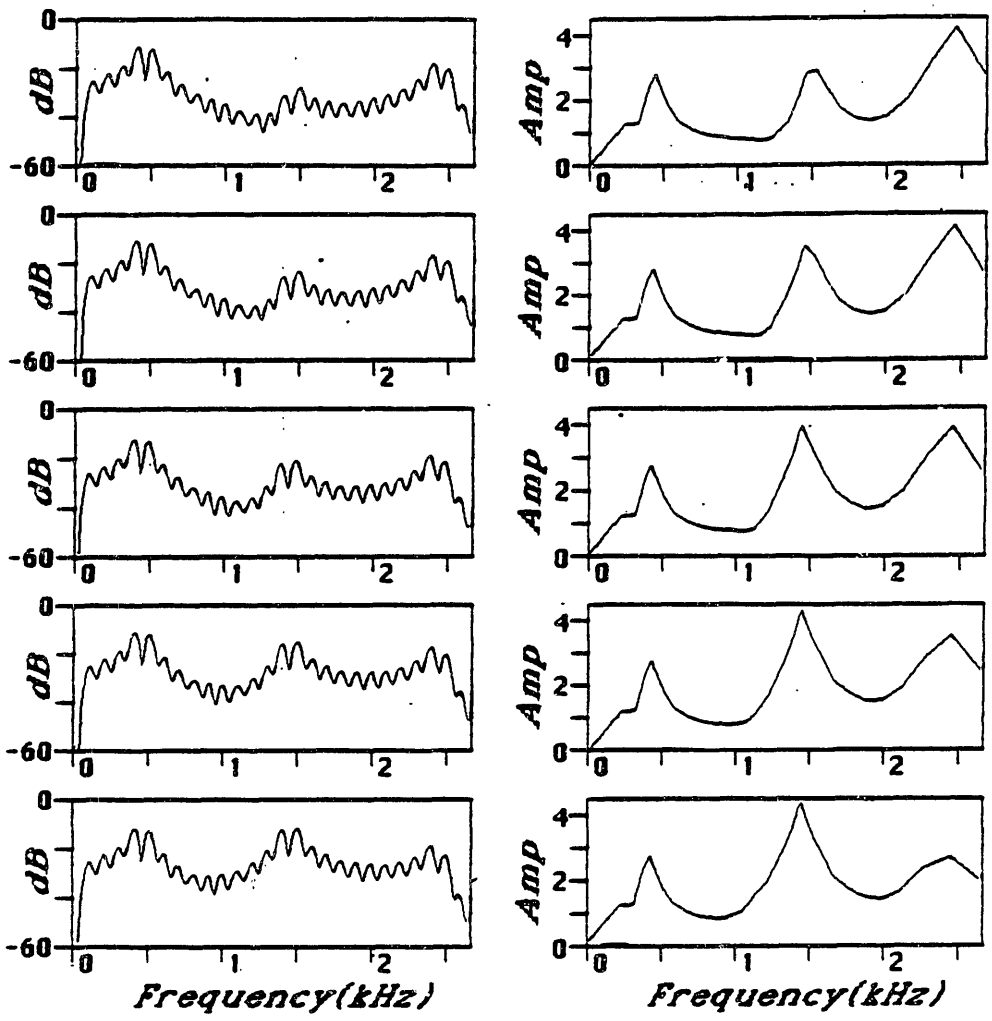

Figure 10.1:
b) Narrow-band spectral cross sections compared with pseudo spectra taken at the midpoint of each vowel in the series of part (a).

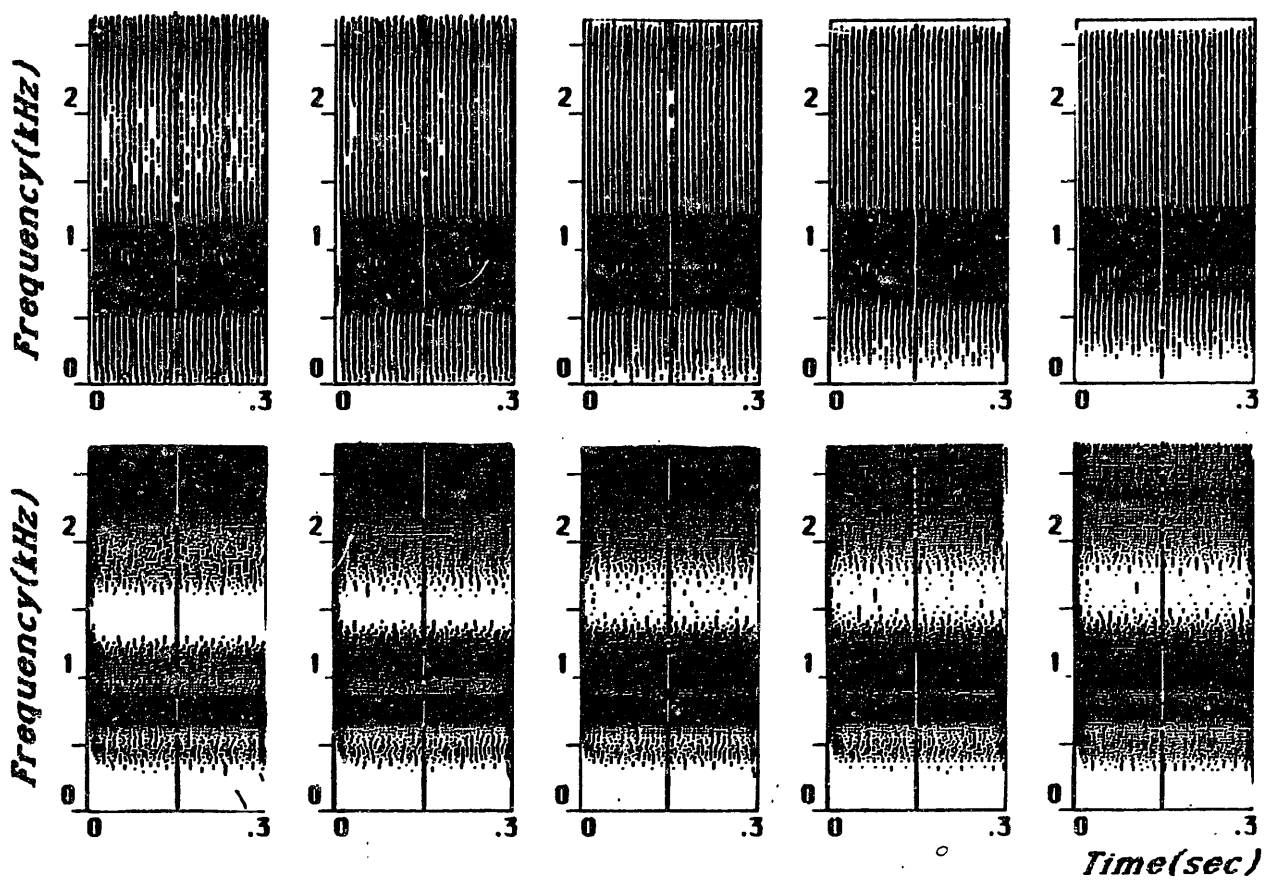

Figure 10.2: Results for modifications in the amplitude of the second formant when the first two formants are close in frequency.
a) Wide-band spectrograms compared with pseudo spectrograms for a series of steady-state vowel stimuli with formant frequencies at 700, 1000, and 2500 Hz. The amplitude of F_2 is increased in 6dB increments from 12 dB below F_1 to 12dB above F_1.

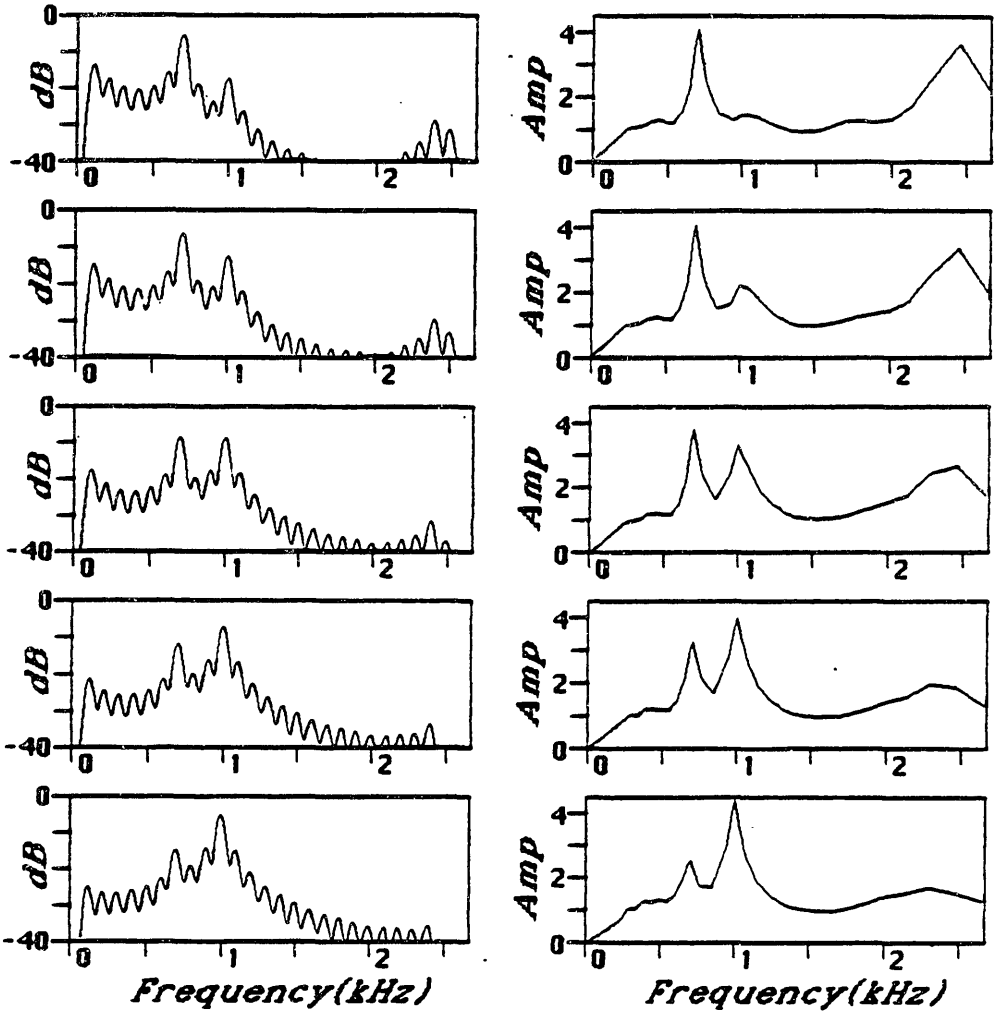

Figure 10.2:
b) Narrow-band spectra compared with pseudo spectra measured at the center of each of the vowels in Figure 10.2a.

	Stimulus No.	F1	F2
F0: 100 Hz	1	695	842 Hz
B1: 50Hz	2	640	842
B2: 70Hz	3	640	941
B3: 110Hz	4	592	941
F3: 2450 Hz	5	592	1000
	6	533	1000

Table 10.1: Acoustic Parameters for Stimuli Analyzed in Figure 10.3.

349 Hz. Although there is a peak at F_2, its amplitude is greatly reduced relative to the amplitude of F_1. F_2 begins to regain prominence with the fifth vowel, with a spacing of 408 Hz. The sixth vowel shows two distinct equal amplitude peaks, with a deep valley between them. The formants are spaced by 467 Hz at this point, which represents an approximate lower limit for separation to avoid the interference of the first formant peak in the response at F_2.

It may be instructive to compare the results of these three experiments with perceptual studies by Chistovich et al [1979]. In one experiment, these researchers generated a set of vowel stimuli with F_1 and F_2 differing in frequency by 350 Hz. In the series S_1, the first formant was 5 dB higher than the second, and in the series S_2, the opposite was true. Listeners were asked to adjust the frequency of a single formant in a stimulus S_0, until they obtained a best match to either S_1 or S_2. Listeners tended to place the single formant somewhere between the two formants, closer to F_1 for the S_1 stimuli than for the S_2 stimuli. Listeners were able to set the single formant frequency in a consistent way, suggesting that the two-formant stimulus could be well represented, at least with regard to phonetic content, by a single formant. The authors were careful to point out, however, that the single formant stimulus was not qualitatively identical to either of the two formant stimuli.

An attempt to adjust a single formant frequency so as to obtain a best match in the pseudo spectrum to a two formant stimulus would probably yield a different sort of result. The resolution of the pseudo spectral analysis is more than adequate to separate out two distinct peaks for two formants spaced by 350 Hz. It would be anticipated that either a single peak could not be adjusted to obtain an adequate match, or, if the imbalance in formant amplitudes was sufficiently great, the single peak would be set at the frequency of the larger peak. However, it is possible that the listeners are dealing with a spectral representation at a higher level than the level corresponding to the pseudo spectrum. At this hypothesized higher level, information in the pseudo spectrum could be further processed, through, for example, a center-of-mass computation. The listeners might then set the single formant peak at the center-of-mass of the two isolated peaks in the pseudo spectrum.

Another complication is that the stimuli generated by Chistovich et al were excited by noise

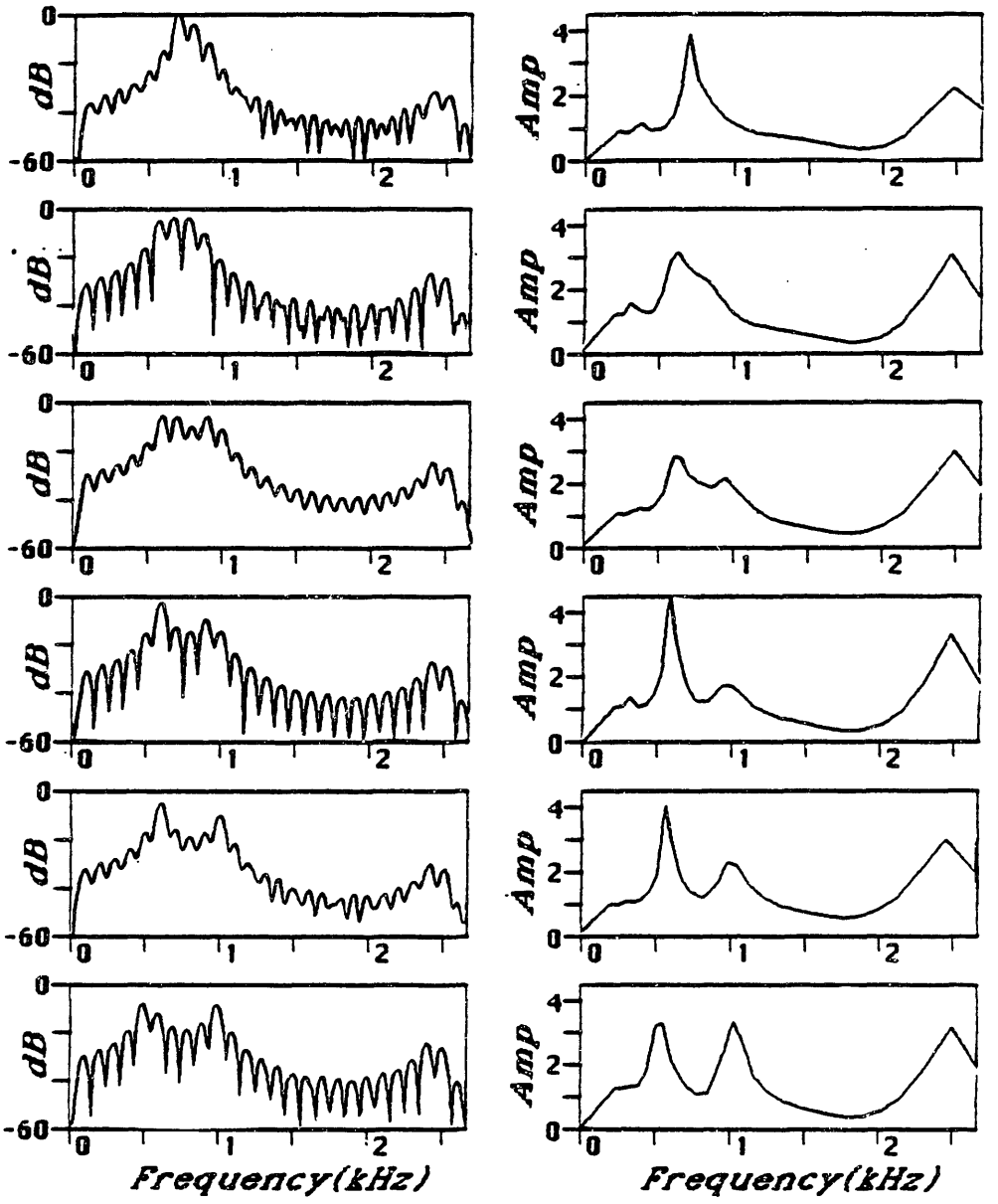

Figure 10.3: Narrow-band spectra compared with pseudo spectra taken at the middle of a series of steady state vowels with formant frequencies as in Table 10.1.

rather than by a periodic source. An ideal random white noise sequence has a flat spectrum, but its duration is infinite. Any restricted time segment of the white noise will have randomly placed spectral prominences that will fluctuate from frame to frame. The variations in frequency characteristics of the noise source spectrum interact with the formant resonances in a complex way. Figure 10.4 shows an example of a noise-excited stimulus, with F_1 and F_2 at 700 and 1050 Hz respectively. Part (a) shows a wide-band spectrogram compared with a pseudo spectrogram, and part (b) shows narrow-band spectra and pseudo spectra at the two time-slices indicated in part (a). Due to differences in the excitation spectrum, the pseudo spectrum does not always resolve the two peaks. It may be easier for a listener to adjust a single formant to "match" this stimulus than one that is harmonically excited.

10.3 Results for a Series of Synthetic /æ/-like Stimuli Differing in a Single Dimension from a Reference /æ/

In this section we examine pseudo spectral analysis of a number of synthetic /æ/-like stimuli, each representing a small perturbation from a canonic reference /æ/. The stimuli were reconstructed as much as possible according to the specifications given in the experiment by Carlson et al. [1979] on perceptual studies for a similar set of stimuli. The purpose of the perceptual study was to evaluate which aspects of the signal spectrum the listener is most sensitive to. In each case, listeners were asked to rank on a scale from 1 to 10 the quality difference between the perturbed vowel and the reference /æ/. Some general results were that listeners were much more sensitive to changes in formant frequencies than in formant bandwidths, and were more sensitive to notches in the spectrum introduced near DC than notches between the first and second formants. Listeners were also surprisingly sensitive to phase characteristics, a result that had not been predicted. In most cases, the differences between the reference and the perturbed vowel were not phonemic; i.e., both vowels sounded like /æ/, but they were qualitatively distinct.

It should be instructive to compare the changes that occur in the pseudo spectrum with the corresponding perceptual differences for various /æ/-like stimuli. However, it is difficult to assess in any precise way the meanings of such comparisons. If, for example, the pseudo spectrum changed little when the perceived difference was great, it could be argued that the perception of this particular spectral change was carried by some other mechanism, such as processing for pitch, roughness, or loudness. On the other hand, if the perceived distance was small and the pseudo spectrum changed substantially, it could be argued that higher level processing of the pseudo spectrum weighted minimally the aspects that had changed. Of course it would be more convincing if the pseudo spectrum changed in accord with the perceptual results, but, perhaps unfortunately, contradictory results do not necessarily **prove** that the proposed central processing strategy is incorrect.

Given the above caveats, we now describe the results for a number of different experiments. Our experiments concern the following categories of change:

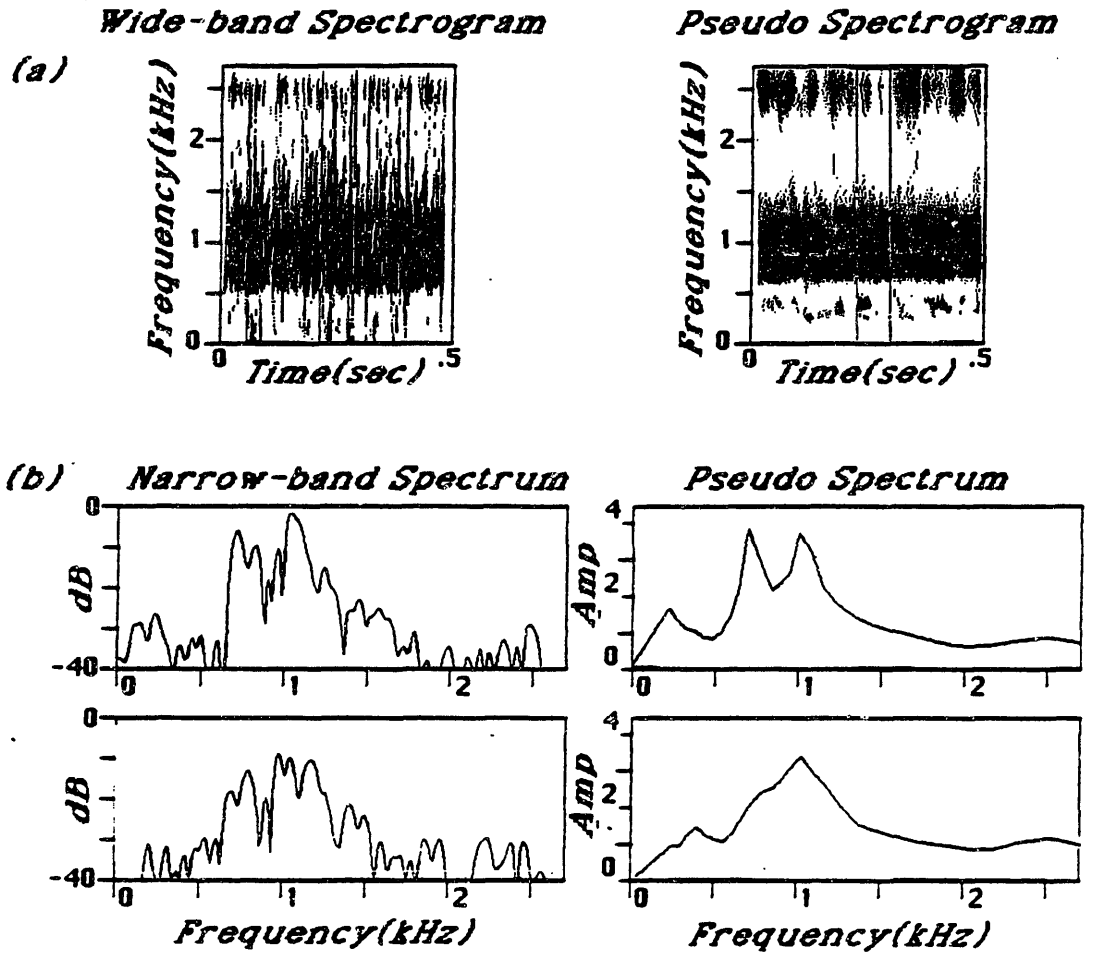

Figure 10.4: Example of pseudo-spectral analysis of noise-excited synthetic speech, with formant frequencies at 700 and 1050 Hz.
a) Wide-band spectrogram compared with pseudo spectrogram.
b) Examples of narrow-band spectra compared with pseudo spectra for the two time slices indicated in part (a) of the Figure.

Time-varying Control Parameters

Time	0	70	140	300ms	
F0	110	125	125	100Hz	
Time	0	20	200	295	300ms
Amp	48	60	60	54	0dB

Constant Control Parameters:

F1-F5:	700	1800	2500	3300	3700Hz
B1-B5:	60	140	150	200	250 Hz

Table 10.2: Acoustic Parameters for /æ/-like Stimuli Used for Experiment 2.

1. Overall amplitude,
2. Phase characteristics,
3. Relative formant frequencies and bandwidths, and
4. Spectral notches

The control parameters for the reference /æ/ were as specified in Table 10.2. In the case of the phase study, however, the overall amplitude and fundamental frequency were kept constant, because it seemed difficult to conceive of a way of implementing a truly random phase spectrum with changing harmonic frequencies. Instead, the N harmonics of a 125 Hz fundamental were initialized to the specified phase characteristic [minimum, maximum, zero, or random]. Such phase initialization, together with amplitude information, fully characterizes the sine wave at each harmonic frequency, since the signal is steady state. In all cases, the spectral and pseudo spectral cross sections are taken, arbitrarily, at the 150 ms time slice, the exact middle of the vowel.

10.3.1 Overall Amplitude

The effect of overall signal level on the pseudo spectrum is shown in Figure 10.5. A series of five stimuli were generated using the reference /æ/ acoustic parameters, with overall energy decreasing in 3dB units down to a minimum of -12 dB. Thus the amplitude of the weakest signal is down by

a factor of four relative to the reference amplitude. It is clear from the Figure that, at least over these ranges, overall level has little influence on the pseudo spectrum. This result is expected, since the GSD algorithm includes an energy normalization step.

Overall amplitude is probably one perceptual result where it is obvious that the perception may not be tied to changes in the pseudo spectrum. Loudness perception seems to be a complex process, and it is unreasonable to tie it to the representation of spectral shape. It is intuitively the case that softly spoken speech and loud speech are represented by similar perceptions of phonetic content. Thus, although the perceptual study indicated a sensitivity to overall level, it seems preferable that the pseudo spectrum exhibit a relatively reduced sensitivity to signal level.

10.3.2 Phase Characteristics

In this section we examine the effects of modifications only to the phase spectrum on the shape of the pseudo spectrum. The study includes the following phase conditions:

1. Minimum Phase,
2. Maximum Phase,
3. Zero Phase, and
4. Random Phase.

Figure 10.6 shows the results for the differing phase characteristics. Part (a) of the Figure shows a section of the waveform for each of the different phase characteristics, and part (b) shows a comparison among pseudo spectra measured at vowel center. The narrow-band spectra were all essentially identical, and therefore are not included in the Figure. The random phase signal is highly dispersed in time. Although it is periodic, the periodicity is not immediately evident, because there is no rise time or decay time with each period. The zero phase signal exhibits a highly peaked symmetrical shape in each period. The maximum phase signal would be equivalent to playing the speech backward in time, if the speech were perfectly steady-state.

Perhaps surprisingly, there are substantial differences among all four conditions in the pseudo spectrum. These differences are most pronounced in the region below the first formant. There is no reason to assume that the very nonlinear processing involved to obtain the pseudo spectrum should not be sensitive to differing phase characteristics. The highly peaked waveform associated with zero phase probably behaves differently through the peripheral model than the diffuse random phase signal. Furthermore, the GSD algorithm is nonlinear, and should not in general be expected to obtain an identical result if the phases of the component sine waves were altered.

The perceptual results of Carlson et al. [1979] showed a substantial sensitivity to phase. It was one of the most perceptually salient modifications from the standard. The random phase signal in fact was the "curve setter" for the experiment: the authors normalized results for all experiments so that the random phase signal received a score of 10, the highest score allowed.

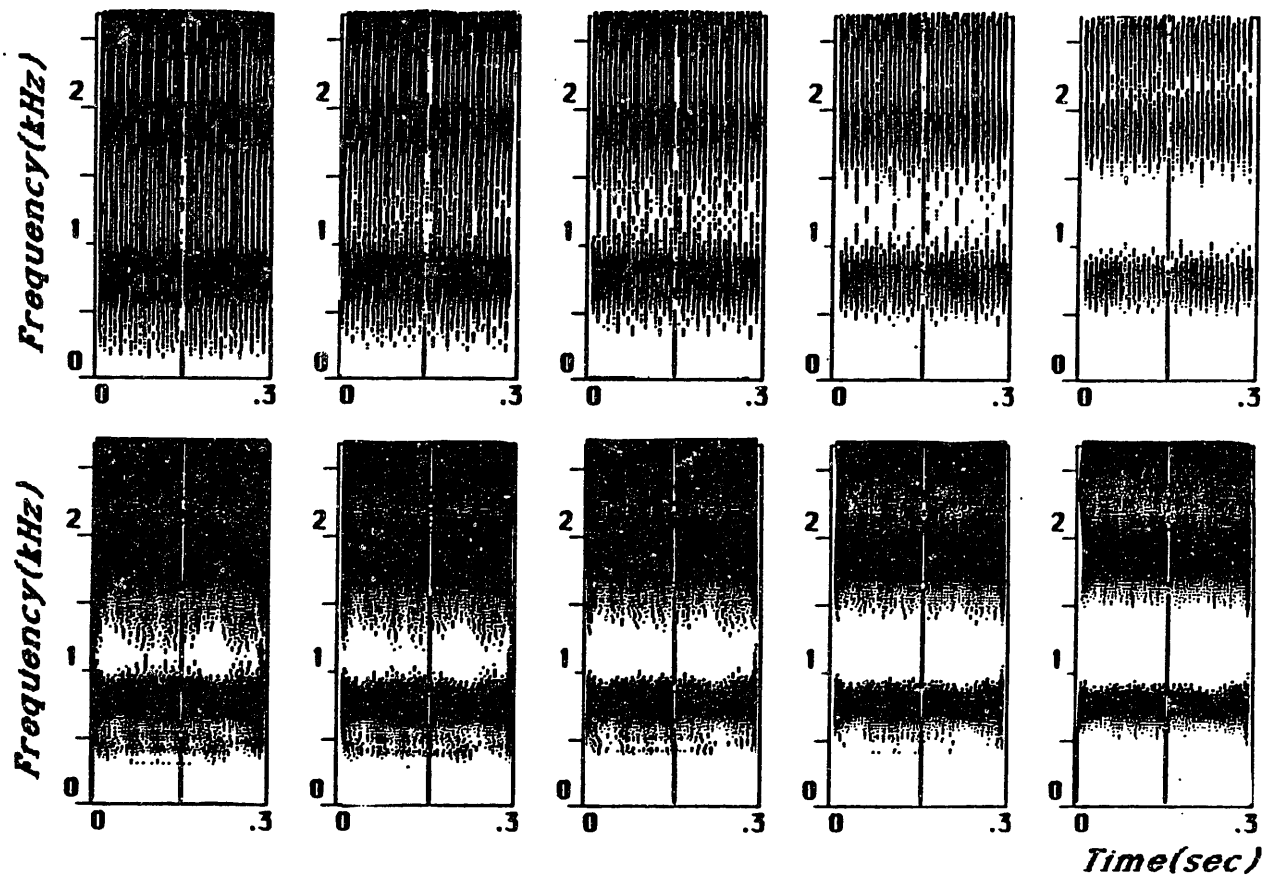

Figure 10.5: Examples showing the effect of overall signal level on the pseudo spectral analysis.
a) Wide-band spectrograms [top] compared with pseudo spectrograms for a series of synthetic /æ/-like stimuli, with overall level decreasing by 3dB increments. The control parameters for synthesis are as in Table 10.2.

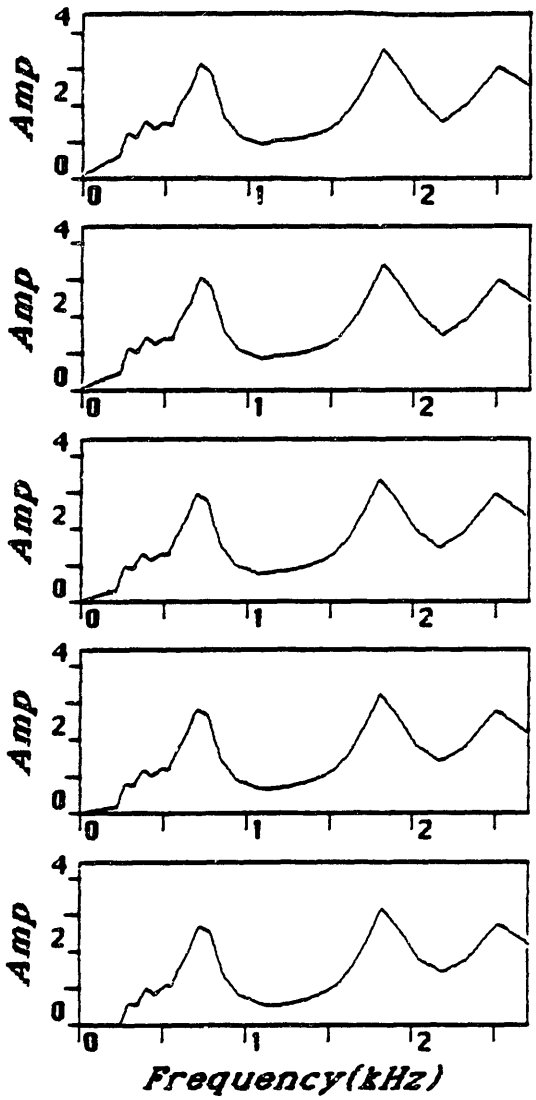

Figure 10.5: b) Pseudo spectra measured at the center of each of the stimuli in part (a) of the Figure. Amplitude decreases from top to bottom.

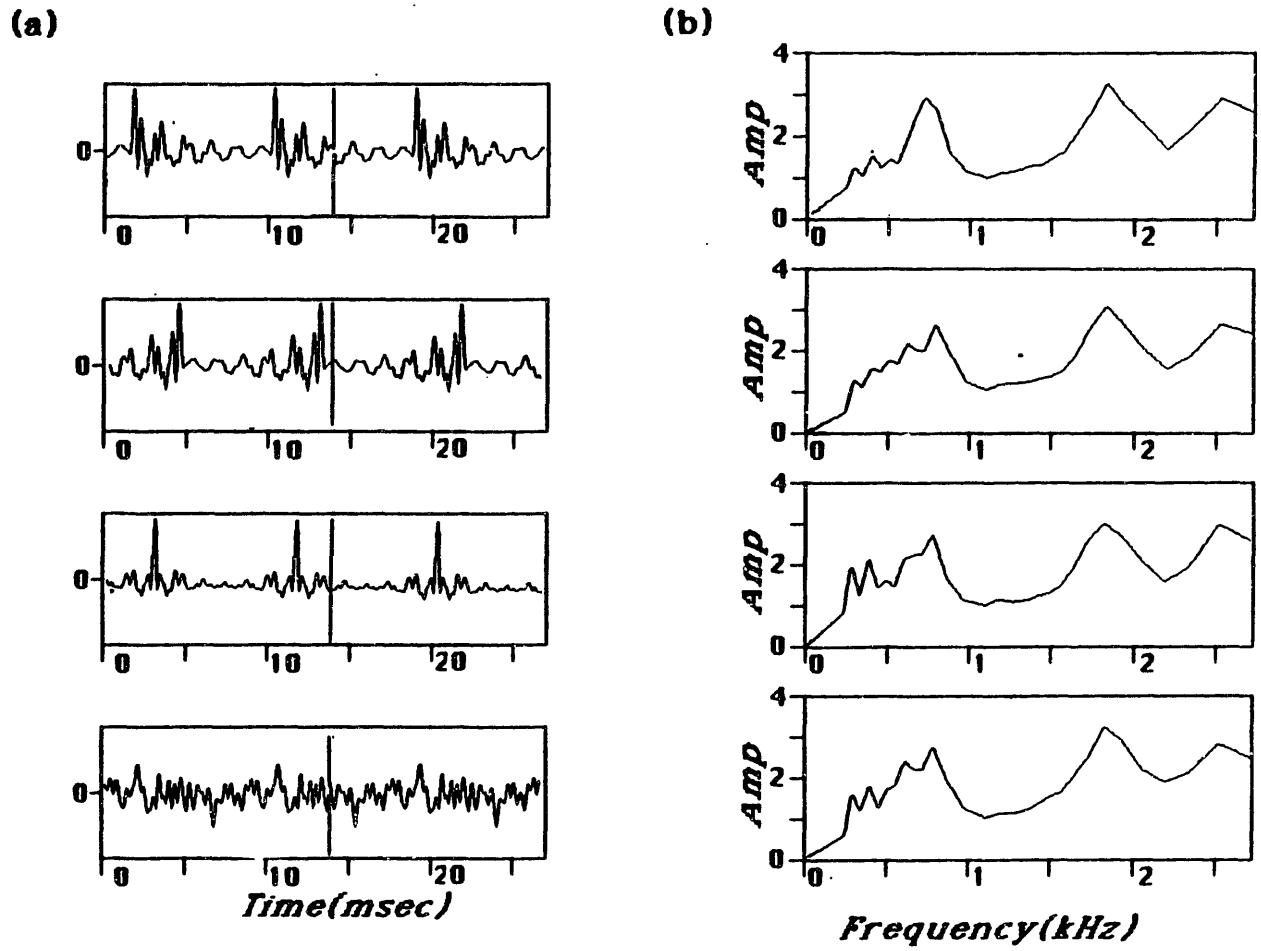

Figure 10.6: Examples illustrating the effect of phase on the pseudo spectral analysis.
a) Portions of the original waveform for a series of synthetic /æ/-like stimuli with different phase characteristics. The phase characteristics from top to bottom are minimum, maximum, cosine [zero], and random.
b) Pseudo spectra computed at the center of each of the vowels as in part (a). Narrow-band spectra are not included because they are all identical.

Given such normalization, the zero phase signal and maximum phase signal received scores of 5.9 and 6.9 respectively. These results came as a surprise to the authors, because it had been generally believed that the ear was relatively insensitive to phase.

It is not necessarily the case that perceived differences due to phase changes are perceived by the same mechanism as that which is related to the pseudo spectrum. However, it is tempting to equate the changes in the low frequency region with phase to the perceptual sensitivity to phase. Since the low frequency region also carries such percepts as nasalization and breathiness, it is likely that changes in this region are perceptually more significant than changes elsewhere in frequency. It is thus possible that inherent phase differences in the signal are heard as amplitude differences in the low frequency region of the spectral shape representation.

10.3.3 Relative Formant Frequencies and Bandwidths

Results for changes in the frequencies of the first two formants are shown in Figure 10.7. Part (a) shows the pseudo spectrum compared with the narrow-band spectrum for the reference /æ/, part (b) shows the results for an 8% increase in the frequencies of the two formants, and part (c) shows the results for an 8% decrease in formant frequencies. The vertical bar is located at the 700 Hz first formant frequency of the standard. It is clear from the Figure that the first formant has not only shifted in frequency but also become substantially sharper. The reason is that the formant has shifted from a frequency that is between two harmonics to a frequency that is much closer to a harmonic frequency, in both cases.

The change in shape of the peak that takes place in the F_1 region is not paralleled by a similar change in shape in the F_2 region, in spite of the fact that the same phenomenon of a shift from interharmonic to harmonic site takes place. The reason is simply that the peripheral filters are broader in the F_2 region, and hence the advantage of nearly isolating a single harmonic does not occur. Nonetheless, there are clear frequency shifts for the second formant in the pseudo spectrum.

Results for the bandwidth experiments are given in Figure 10.8. A 40% increase or decrease of the bandwidth of F_1 results in essentially no change in the pseudo spectrum, as shown in part (a) of the Figure. Actually, a 40% change in bandwidth is realized as only a slight increase in the amplitude of the peak of the narrow-band spectrum, and a slightly accelerated decrease in amplitude as one moves away from the peak, as can be seen from the narrow-band spectra in the Figure.

Results for 40% changes in the bandwidth of F_2 are given in Figure 10.8b. The pseudo spectrum is much more sensitive to bandwidth changes in F_2 than in F_1. There is a slight decrease in the level of the peak at F_2 for the decrease in bandwidth, and there is a substantial sharpening of the peak for an increase in bandwidth. These effects are critically related to the shapes of the filters in the peripheral model.

The perceptual study indicated a relatively strong sensitivity to changes in formant frequencies, which was contrasted with an almost complete insensitivity to changes in formant bandwidths, at least to the degree to which these changes were introduced in the stimuli. For an 8% change in

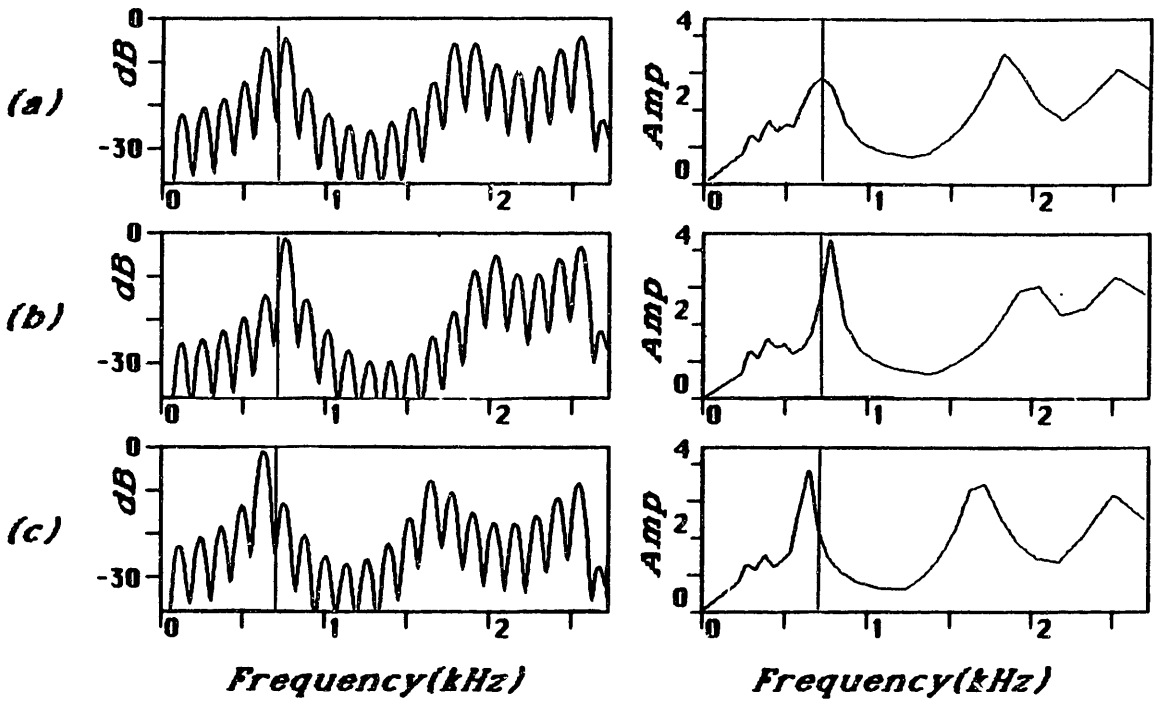

Figure 10.7: Examples illustrating the effects of changes in formant frequencies on the pseudo spectrum.
a) Narrow-band spectrum [left] and pseudo spectrum for the reference /æ/.
b) Narrow-band spectrum and pseudo spectrum for a vowel with an 8% increase in the frequencies of both F_1 and F_2.
c) Narrow-band spectrum and pseudo spectrum for a vowel with an 8% decrease in formant frequencies.

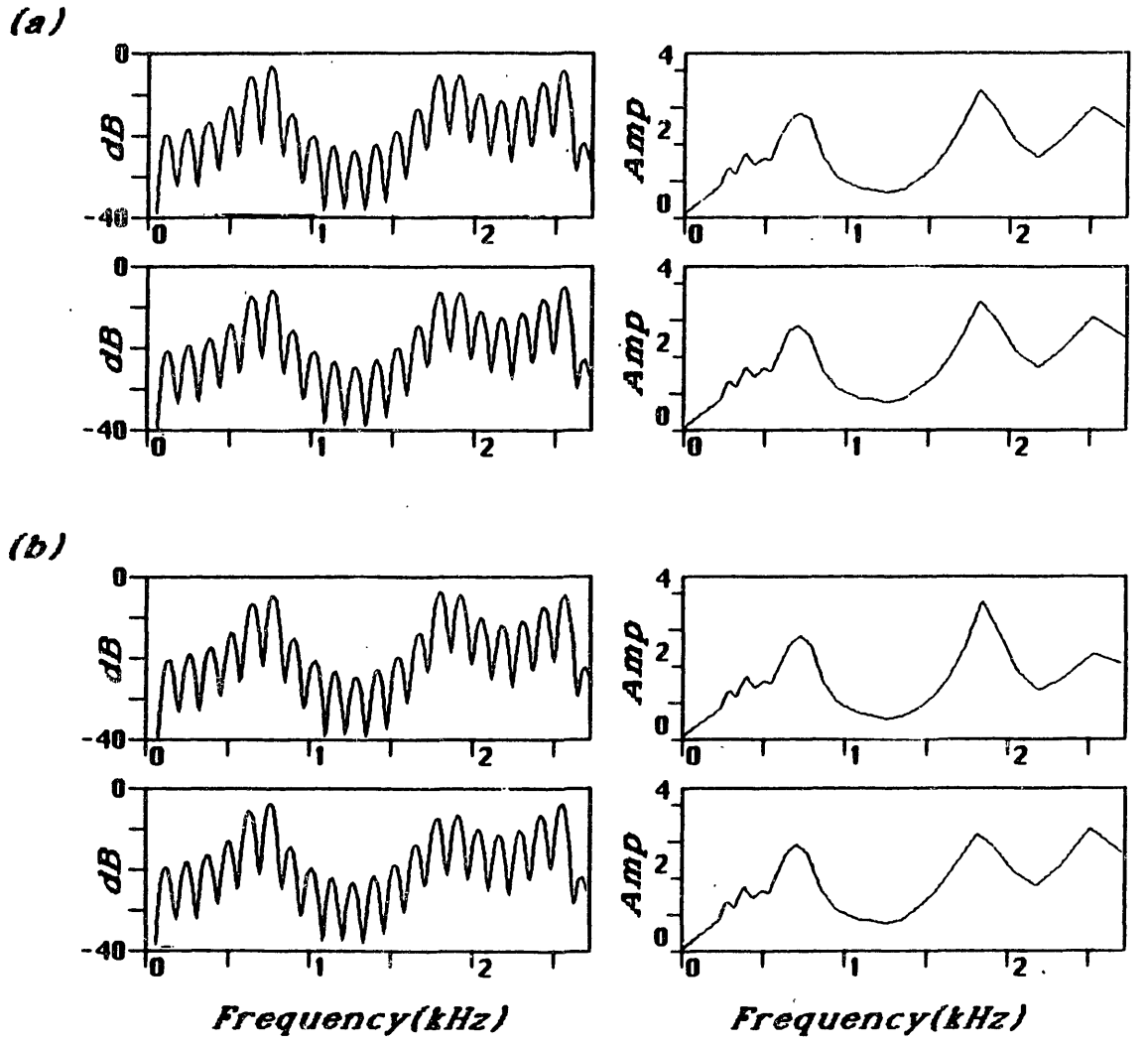

Figure 10.8: Effects of modifications of formant bandwidths on the pseudo spectral analysis.
a) Narrow-band spectra compared with pseudo spectra for /æ/-like stimuli differing from the standard only in the bandwidth of F_1. [Top] 40% increase in bandwidth of F_1
[Bottom] 40% decrease in bandwidth of F_1.
b) Same as in part (a) of the Figure, except that the bandwidth of F_2 is increased [Top] by 40%, andd decreased [Bottom] by 40%.

the frequencies of both F_1 and F_2, the perceived change was ranked at about 4.5, regardless of the direction of change. A 40% change in bandwidth of either formant, in either direction, resulted in a subjective distance of only slightly greater than zero.

Perceptual results seem to be in agreement with pseudo spectral results, except for the relatively large change in the prominence of the peak at F_2 with 40% changes in the formant bandwidth. This may suggest that the filter characteristics near the frequency region of the second formant are incorrect in the model. However, there are certain situations where it is advantageous to show a sensitivity to the bandwidth of the peak near the second formant frequency. For example, as mentioned previously, an /ɝ/ is almost always characterized by a very prominent single peak in the pseudo spectrum representing a merger of the second and third formants. The double pole could be conceptualized as approximating a single pole with a broader bandwidth. A strong contrast between the two peak and single peak situations is desirable. A similar situation may hold for rounded front vowels, which typically also have the two formants close together, but shifted upward in frequency, relative to /ɝ/.

The sensitivity of the pseudo spectrum to the frequency of the first formant was enhanced by the changing relationship between the formant frequency and the harmonic structure. It would be interesting if a perceptual experiment could be devised to test whether vowel quality is substantially different when the formant frequency straddles two harmonics versus when it coincides with a single harmonic. Such a test is difficult because the necessity of changing either pitch or F_1 complicates the perceptual correlates that one is trying to measure.

10.3.4 Spectral Notches

In this section we examine the effects of notches in the spectrum on the pseudo spectral analysis, as well as the effect of an amplitude boost in the low frequency region. Figure 10.9 shows a series of stimuli with a notch at 1250 Hz, and a notch width of 0, 300, 500, and 700 Hz, respectively. Although the valley between the two formant peaks does become somewhat deeper in the pseudo spectrum, the locations of the peaks do not change until the 700 Hz notch bandwidth stimulus. This stimulus sounded more like an /ɛ/ than an /æ/ to me, whereas the other stimuli all seemed very similar to the original /æ/.

Figure 10.10 shows the results of some manipulations in the low frequencies. Part (a) shows the pseudo spectrum for the standard /æ/ for reference. Part (b) shows the result when a notch is introduced in the spectrum below 300 Hz, and part (c) shows the result when the low frequency spectrum is boosted in amplitude. Since the lowest filter f_c in the pseudo spectrum is at 228 Hz, the region where the changes are most manifest is absent from the representation. Nonetheless, there are some obvious effects on the spectral region just above 300 Hz. In particular, the second and third harmonics of the fundamental are stronger when the notch is present, and are absent altogether when the low frequencies are boosted. These manifestations are a consequence of a masking-like phenomenon that takes place when energy is present in the very low frequencies. The lack of harmonics below F_1 for the low-frequency boosted speech is consistent with the results for

breathy Gujarati vowels reported on in Chapter 9.

The perceptual studies indicated a lack of sensitivity to a notch in the spectrum between the first and second formants, until the width of that notch became great enough so as to encroach upon the formant regions. Such results are consistent with the pseudo spectral analysis, particularly if we hypothesize that frequencies of peaks are more significant than detailed shapes. On the other hand, a notch near DC was perceptually quite salient. In spite of the fact that the changes are introduced in a region that is outside of the main concentrations used for the pseudo spectrum, there are some effects on the low frequency region of the pseudo spectrum when the spectral region below 200 Hz is modified. If we hypothesize an enhanced sensitivity of a central processor to the F_1 region, then perhaps the observed changes would be perceptually significant.

To summarize the results of the above experiments, it seems that the proposed pseudo spectral analysis method is not inconsistent with the perceptual experiments of Carlson et al. In particular, the results suggest that a more central processor must pay attention to the details of the shape of the pseudo spectrum in the first formant region in order to detect certain voice quality aspects of the signal. In contrast, the frequency location and relative prominence of the peak in the F_2 region may be sufficient information to characterize the spectrum here.

10.4 Results of Analysis of a Series of Synthetic Vowels Varying on a Nasal-nonnasal Continuum

The third experiment concerns a set of five series of synthetic vowel stimuli, each of which represents a continuum from non-nasal to nasal vowel. The five vowels are /i/, /e/, /a /, /o/, and /u/, each of which is realized in a dental context [preceded by /t/]. The stimuli were generated by S. Hawkins and K. Stevens [1984] as part of a study of perception of nasalization. The intent was to model nasalization acoustically, rather than through an articulation model, so that the spectral characteristics could be manipulated directly and well understood.

The stimuli were created using the Klatt cascade synthesizer [1980], with formant frequencies specified directly by pole locations, and an additional pole-zero pair in the first formant region used to simulate nasalization. Figure 10.11 shows the target frequency of the first formant, F1, the nasal zero, FNZ, and the nasal pole, FNP, for each token in the series. Nasalization was introduced in each stimulus 40 ms after vowel onset. The frequency of the extra pole and zero always began at 400 Hz at 40 ms after vowel onset, and then diverged linearly for the next 40 ms, at which time the target frequencies for that stimulus were reached. The specifications then remained constant, at the target frequencies, until the end of the vowel. For the nonnasal token, Stimulus No. 1 in each case, the extra nasal pole and zero stayed at 400 Hz throughout the vowel, superimposed so as to cancel out. Increasing nasalization was realized by spreading the target pole and zero along two diverging lines in frequency. In some cases the target first formant frequency was also varied along a straight line. The bandwidths of the nasal pole and zero were held constant at 100 Hz.

The results of the perceptual studies are shown in Figure 10.12, for a variety of different listener groups. For the most part, each series can be grouped into a number of tokens that are mostly

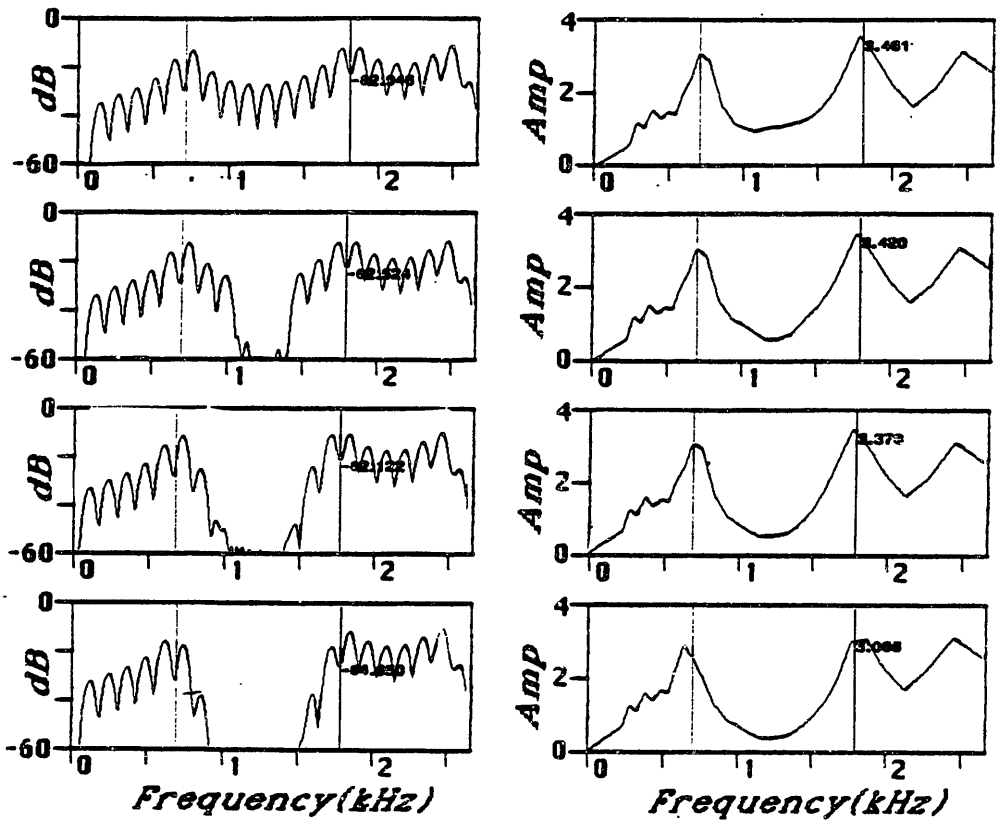

Figure 10.9: Effects of spectral notches between F_1 and F_2 on pseudo spectrum. In all cases, narrow-band spectra [left] are compared with pseudo spectra [right]. The notch is centered at 1250 Hz. Vertical bars mark the frequencies of the two formants.

a) Notch width = 0 Hz [no notch]

b) Notch width = 300 Hz.

c) Notch width = 500 Hz.

d) Notch width = 700 Hz.

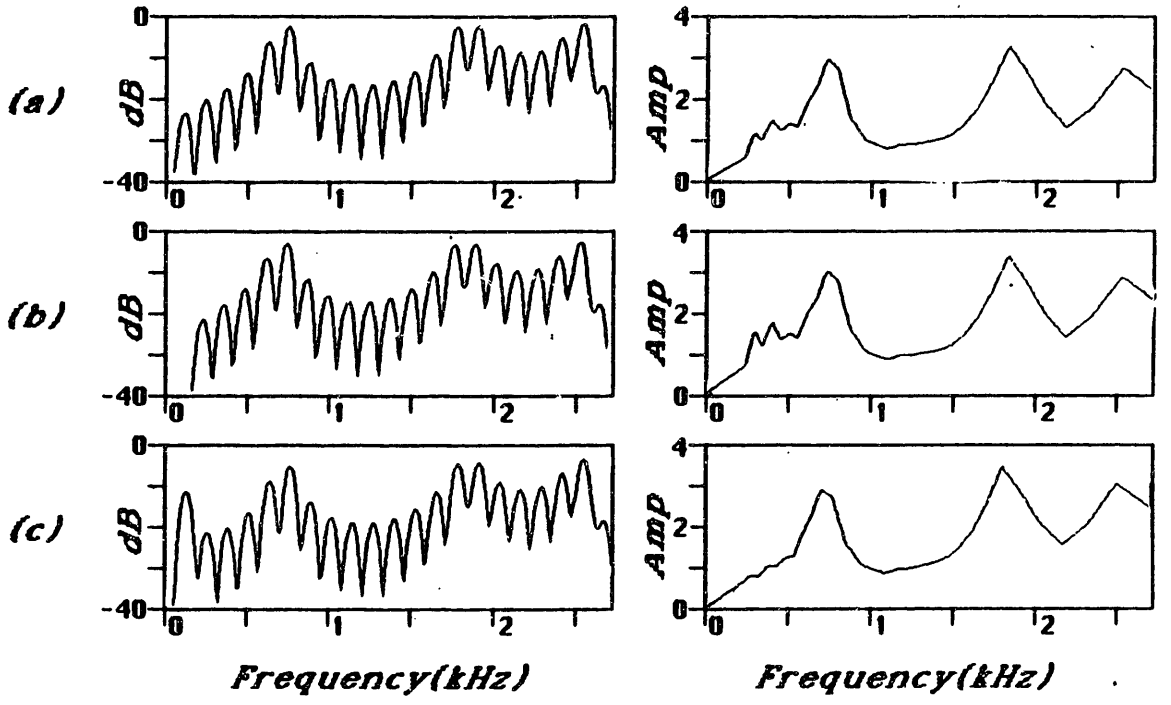

Figure 10.10: Effects of modifications near DC on the pseudo spectrum.
a) Narrow-band spectrum [left] compared with pseudo spectrum [right] for reference /æ/.
b) Same, for /æ/ with spectral notch from 0 to 300 Hz frequency.
c) Same, for /æ/ with amplitude boost in low frequency region.

judged nonnasalized, followed by a number of ambiguous tokens, and finishing with a number of mostly nasalized tokens. The crossover point is different, however, for different vowels. For example, tokens 3 through 5 of /u/ are ambiguous, whereas these are judged mostly nonnasalized for /o/.

Figures 10.13 through 10.17 show the results of a computer analysis of the synthetic tokens, where the pseudo spectral analysis [right] is compared with cepstral analysis [left] and LPC analysis [middle]. These two methods were selected for comparison because they represent the current standard methods of spectral processing. Each figure concerns a single vowel series, and each spectrum was taken at the same fixed place near the end of the vowel. The series represents increasing nasalization from top to bottom, and thus the top spectrum represents the nonnasalized stimulus, and the bottom spectrum the most nasalized stimulus of the series. The panel at the top left of the Figure shows the actual pole-zero-pole specifications for that vowel, taken from Figure 10.11, and the panel at the top right shows the results of the perceptual studies, taken from Figure 10.12.

For the cepstral and LPC analyses, the original waveform was digitized at 16 kHz, and then lowpass filtered and downsampled to 5333 Hz. Both methods used a 25 ms Hamming window. The LPC was an 8th order model, and the cepstral analysis was computed by applying a low time lifter to the log spectrum of the windowed speech, with a cutoff time of 2 ms. The pseudo spectrum was computed directly from the 16 kHz waveform, using the standard method as described in Chapter 8.

It is instructive to examine the pseudo spectral analysis from two perspectives. On the one hand, we can examine whether the pseudo spectrum can recover the underlying pole-zero-pole complex in the F_1 region. On the other hand, it is of interest to see whether the perceptual groupings of the series into nonnasal, ambiguous, and nasal bear any obvious relationship to natural groupings of the pseudo spectrum. A model for the relevant acoustic correlates of the perception of nasalization might also emerge from the study.

Figure 10.13 shows the series for /to/. In this case, the LPC analysis yields two peaks, with one at the second formant and the other drifting up slowly from the 400 Hz first formant frequency to about 500 Hz, at the bottom, where it becomes a shoulder on the second formant frequency. The cepstral analysis also yields a peak for F_2, but the peak at F_1 is lost during the middle tokens of the series, and regained at the end when the extra nasal pole has merged with the second formant. In addition, there is a low amplitude peak at about 200 Hz, the second harmonic of the pitch. The pseudo spectral analysis maintains a steady peak at the first formant frequency, and a valley begins to appear at the nasal zero frequency [around 500 Hz] at about the sixth stimulus. The nasal pole can also be traced by eye, starting at about the fifth stimulus, although it is usually not present as a distinct peak. The peak at the 800 Hz second formant frequency is maintained throughout.

The perceptual response to the /to/ series, shown on the upper right hand corner of the Figure, gives a fairly sharp boundary between nasal and nonnasal. The first five tokens can be considered nonnasal, and the last three are essentially nasal. Only the sixth token is in question. For the series of pseudo spectra, the first four have a sharp high peak at the first formant frequency, and a weak peak at the second formant frequency, with a shallow valley between them. For the last

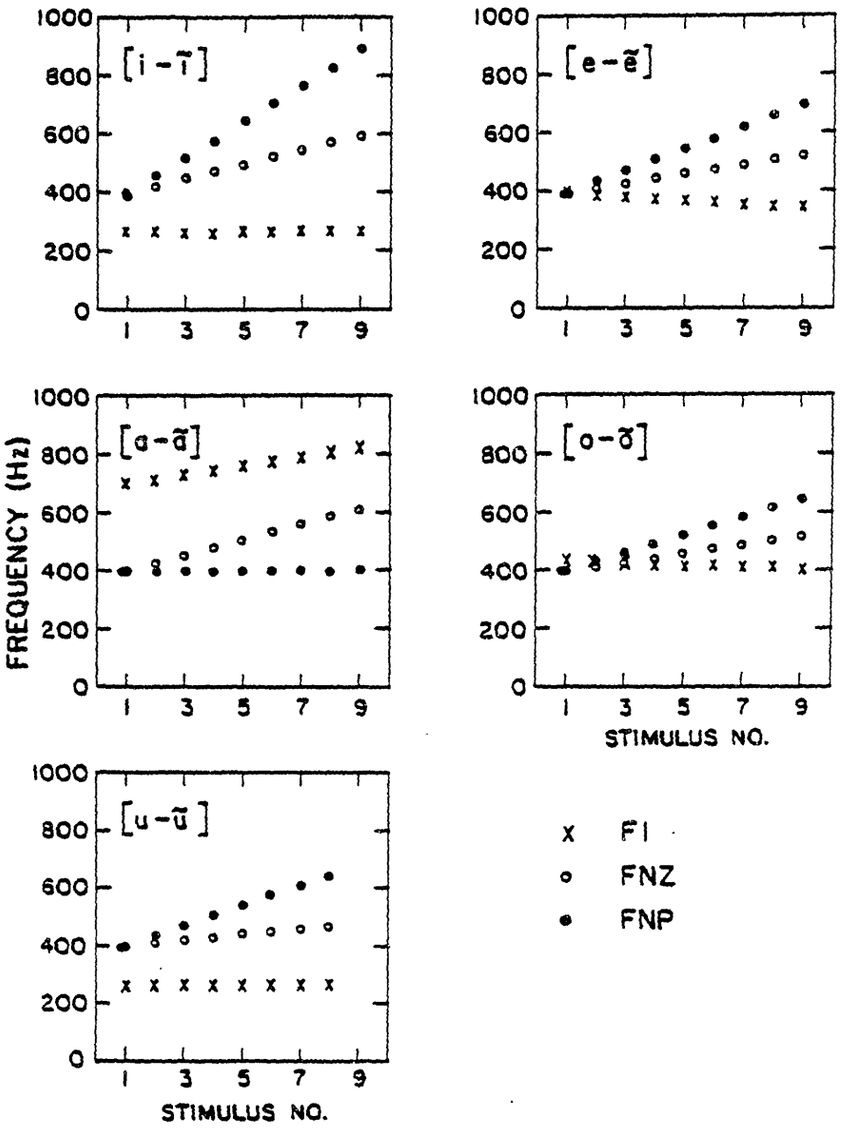

Figure 10.11: Target frequencies of pole for F_1 and pole-zero complex for nasalization for each of the stimuli used in the nasalized vowel study by Hawkins and Stevens. The second formant frequency is omitted from the plots. [From Hawkins and Stevens, in press]

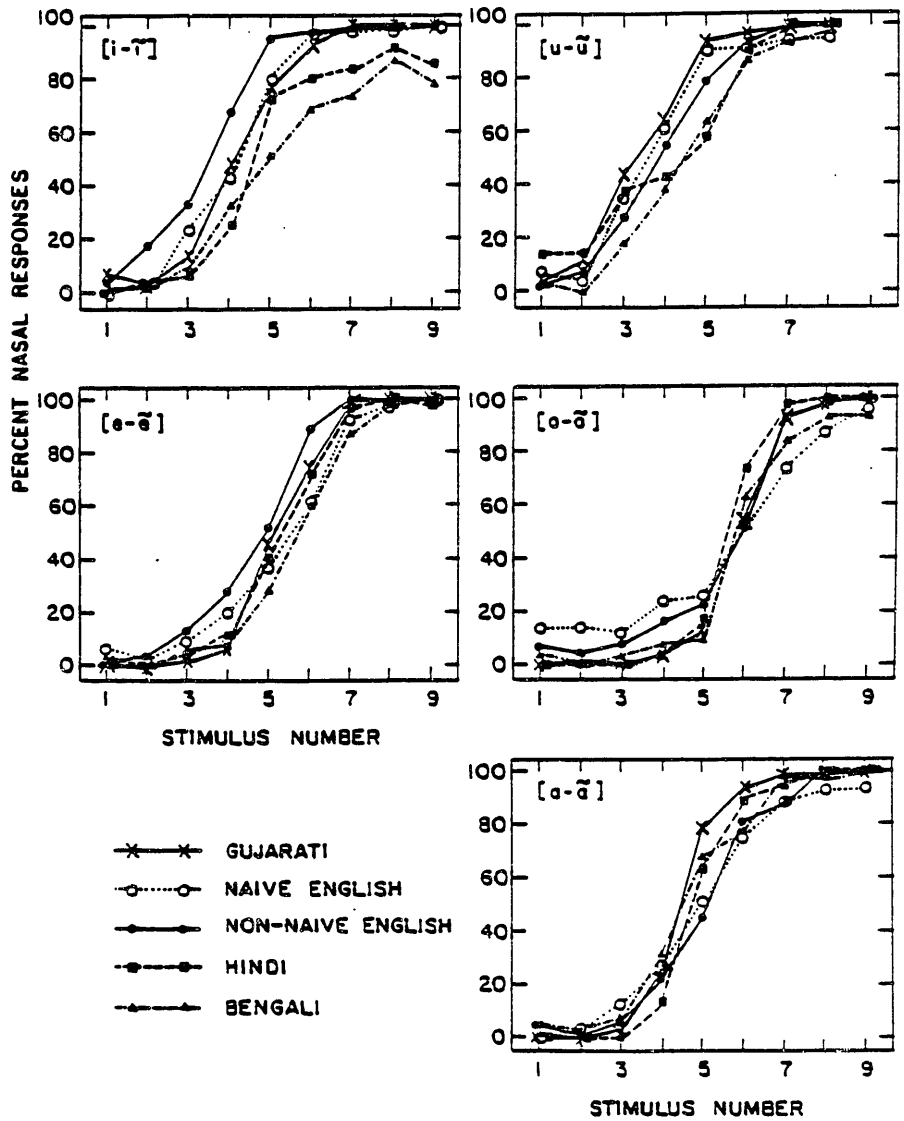

Figure 10.12: Results of a perceptual study of the vowel stimuli described in Figure 10.11. Each subject was asked to identify each stimulus as nasalized or non-nasalized. The data are separated into five distinct subject classes. Each vowel series shows a consistent trend towards increased nasalization with increased separation of the pole-zero-pole complex in the first formant frequency region. [From Hawkins and Stevens, in press.]

three tokens, the second formant is only slightly reduced in amplitude relative to the first, and the valley at the zero is evident. Tokens 5 and 6 can be grouped into a separate class with neither a prominent first formant nor a valley for the zero. If we hypothesize that the narrow prominent peak at F_1 represents nonnasalization, then probably the first four tokens fall in that class, and the fifth is somewhat dubious. In the case of LPC the changes from frame to frame are more gradual. If we use a definition of nasalized as having the second formant exceed the first formant in amplitude, then the first three are nonnasalized, the next three are in question, and the last three are nasalized. The cepstral analysis does not yield a clear peak for F_1 during several of the intermediate tokens, and it is difficult to conceive of a definition of nasalization for this vowel series.

The /ta / series is shown in Figure 10.14. Here again, the pseudo spectrum shows a prominent peak for F_1, with an amplitude well above that of F_2, for the first three, and possibly fourth, of the series. The valley for the zero begins to appear early in the series, and is distinct by the fourth or fifth token. The nasal pole also begins to show up early, but is prominent starting at the sixth token. LPC analysis shows only two peaks, one for each formant, in the first five tokens of the series. The remaining four show a peak for the nasal pole, but, as might be expected for an all-pole analysis, a weak indication of the valley for the zero. The cepstral analysis is quite complicated in appearance, with an extra peak at 200 Hz cluttering up the spectrum. The zero is prominent beginning with about the fifth token, as for pseudo spectral analysis.

Perceptually, the first three tokens are nonnasalized, and the last four are nasalized. This result fits well with a hypothesis of prominence of the first formant relative to the second, with respect to pseudo spectral analysis. The contrast is definitely more pronounced in the case of the pseudo spectrum than in the other two analysis methods.

The vowel series /tu/ is shown in Figure 10.15. Here, the pseudo spectral analysis traces clearly the pole-zero-pole progression in the F_1 region, with a deep valley apparent for the last three or four tokens. As for /to/ the amplitude of the first formant is much greater than that of the second for the nonnasalized tokens, and, eventually, the second formant becomes almost equal in amplitude to the first by about the sixth token. LPC analysis yields only two peaks for each token, with the bandwidth of F_1 changing somewhat dramatically from token number 4 to token number 5. The cepstral analysis is able to recover the pole-zero-pole complex, although the peak at the first formant is rather broad.

The perceptual results for /tu/ show only the first two tokens as definitely nonnasal, and the last three or four as definitely nasal. Thus the perceptual boundary seems to be at odds with both the pseudo spectral and the LPC analysis methods. The cepstral method, on the other hand, seems to show a natural grouping of the first two as containing no valley, and the last four as containing a prominent valley, with the remaining ones ambiguous.

The /te/ series is shown in Figure 10.16. Here, the pseudo spectral analysis shows a single pronounced peak in the F_1 region for the first three tokens, and a distinct diffusion of energy for the last three tokens. The middle three lie somewhere in between. This grouping corresponds well with the listener tests, and would imply that a pronounced spectral prominence in the F_1 region corresponds to a percept of nonnasalized. In addition the apparent formant frequency shifts from

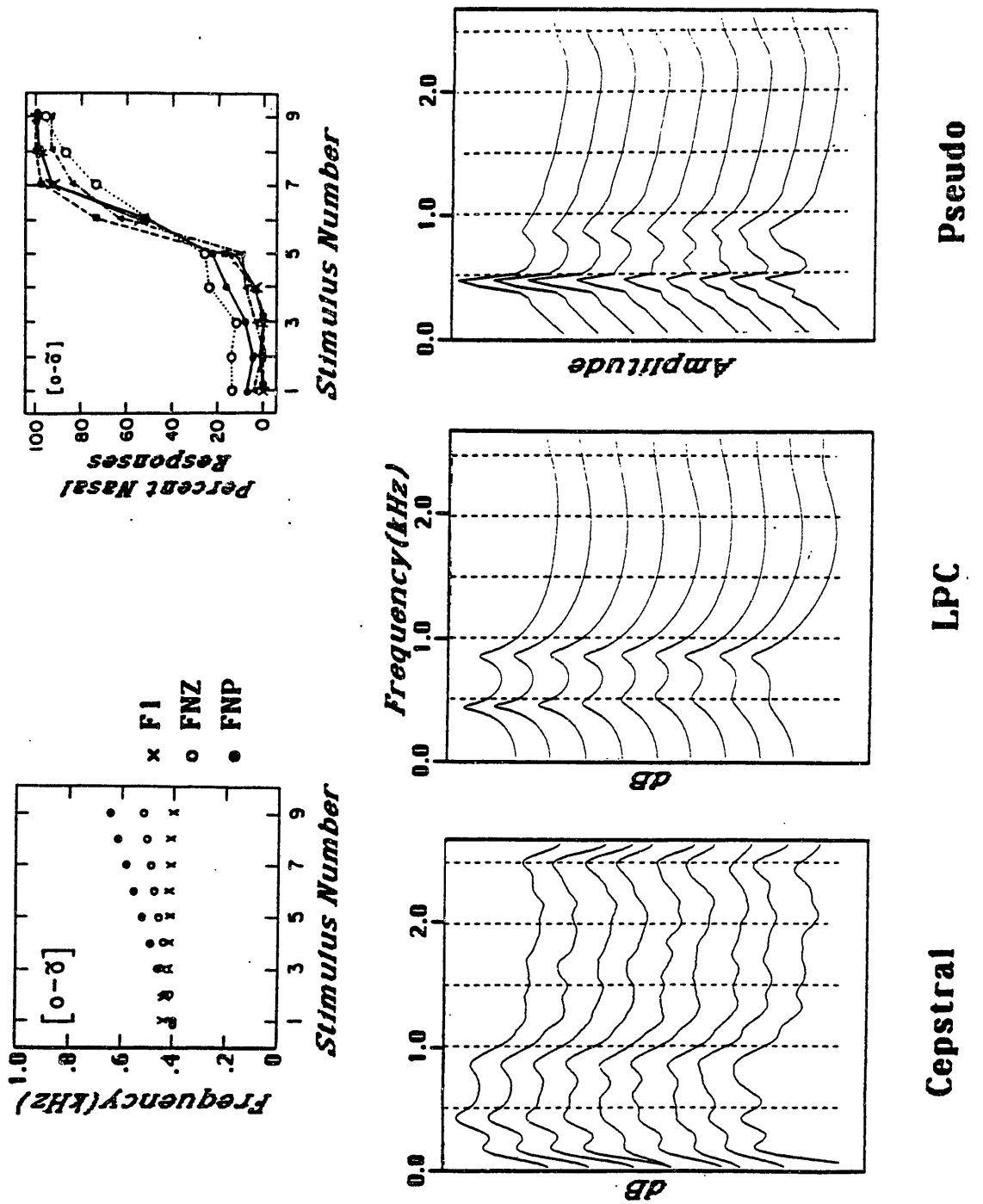

Figure 10.13: Results of cepstral analysis, LPC analysis, and pseudo spectral analysis of the synthetic "to" stimulus series. Stimulus number increases in each case from top to bottom. Analysis was performed at a time slice near the end of the vowel in each case. The acoustic and perceptual data for /o/ are included at the top left and top right corners of the Figure for reference.

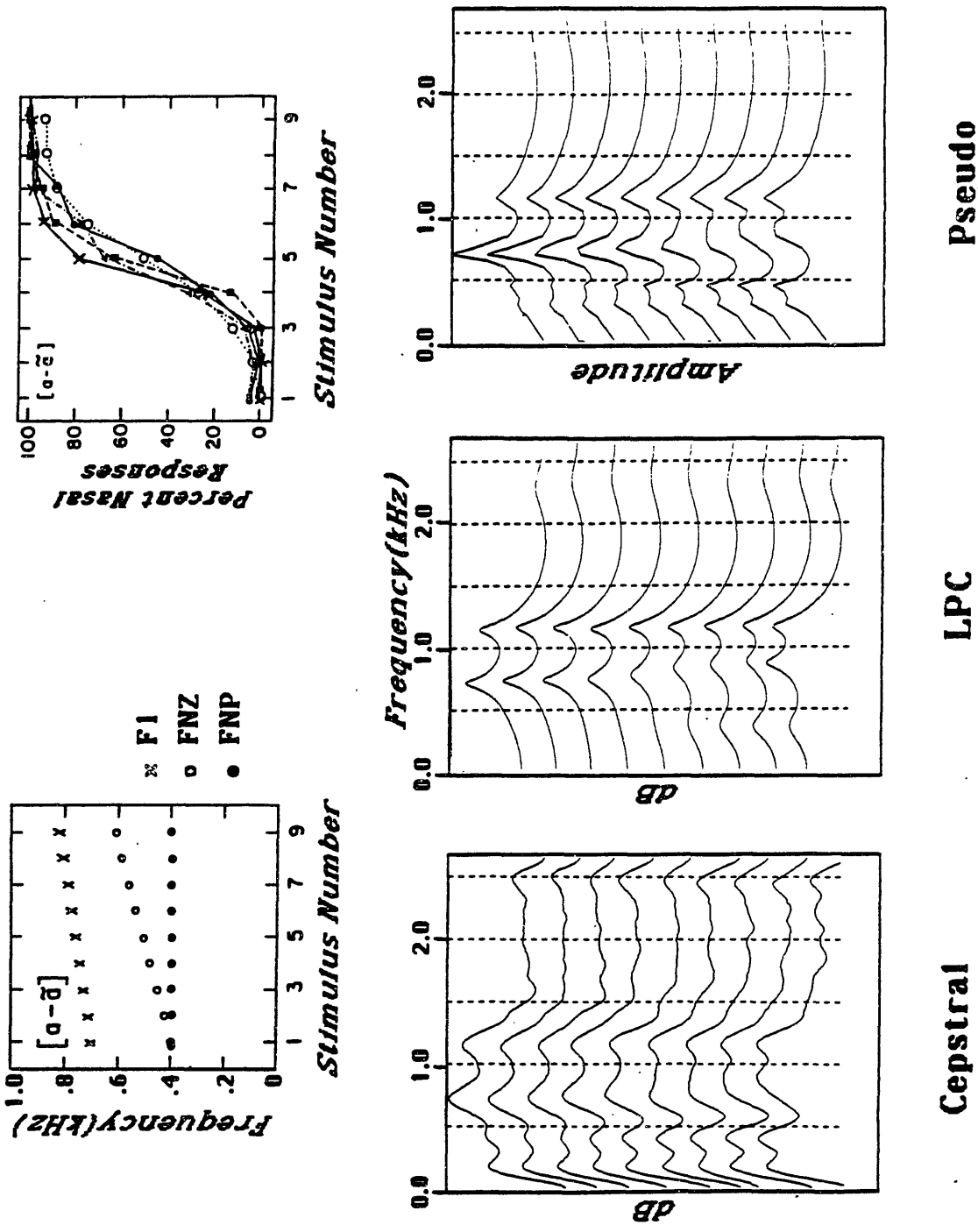

Figure 10.14: Results of cepstral analysis, LPC analysis, and pseudo spectral analysis of the synthetic "t*a*" stimulus series. Stimulus number increases in each case from top to bottom. Analysis was performed at a time slice near the end of the vowel in each case. The acoustic and perceptual data for /*a*/ are included at the top left and top right corners of the Figure for reference.

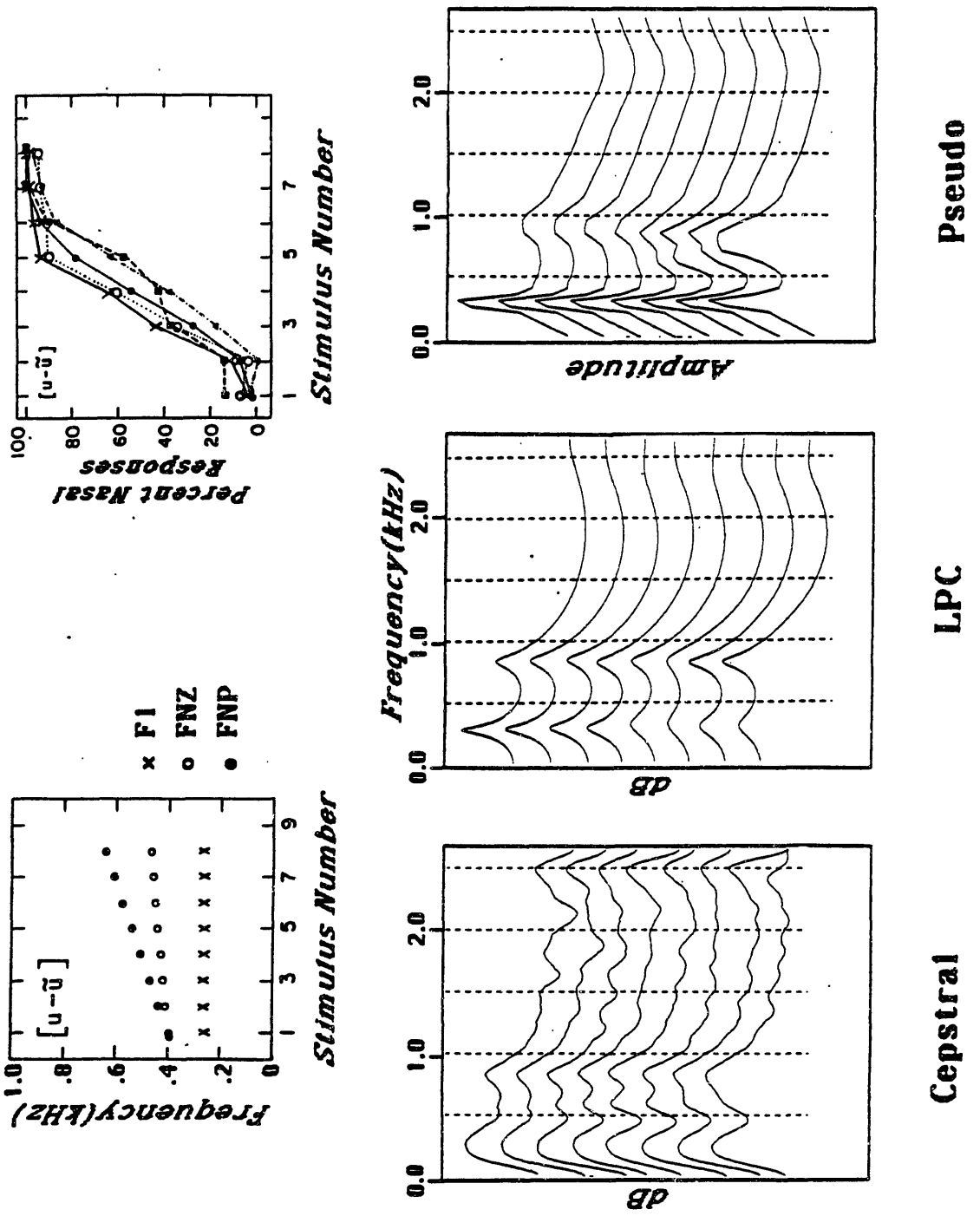

Figure 10.15: Results of cepstral analysis, LPC analysis, and pseudo spectral analysis of the synthetic "tu" stimulus series. Stimulus number increases in each case from top to bottom. Analysis was performed at a time slice near the end of the vowel in each case. The acoustic and perceptual data for /u/ are included at the top left and top right corners of the Figure for reference.

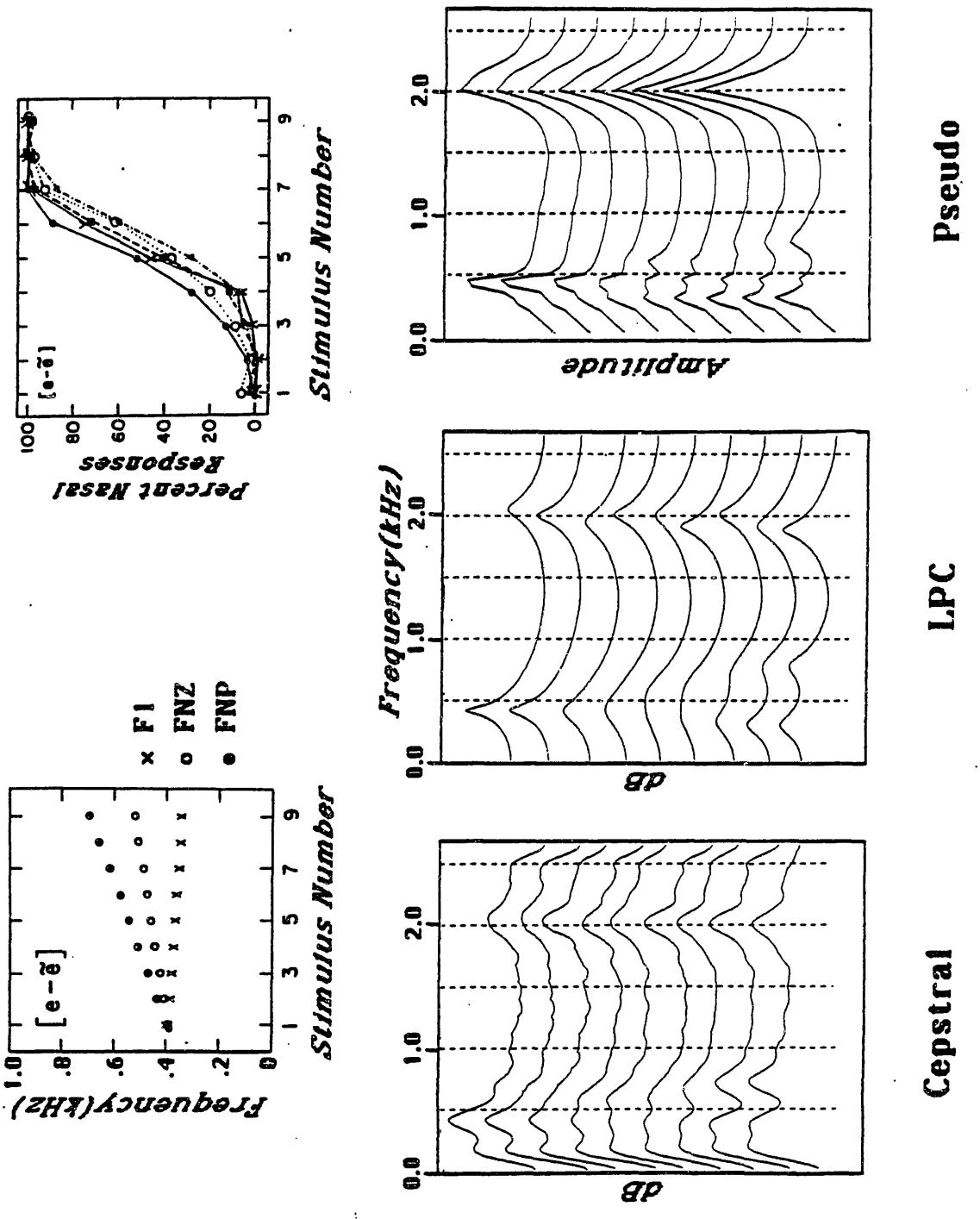

Figure 10.16: Results of cepstral analysis, LPC analysis, and pseudo spectral analysis of the synthetic "te" stimulus series. Stimulus number increases in each case from top to bottom. Analysis was performed at a time slice near the end of the vowel in each case. The acoustic and perceptual data for /e/ are included at the top left and top right corners of the Figure for reference.

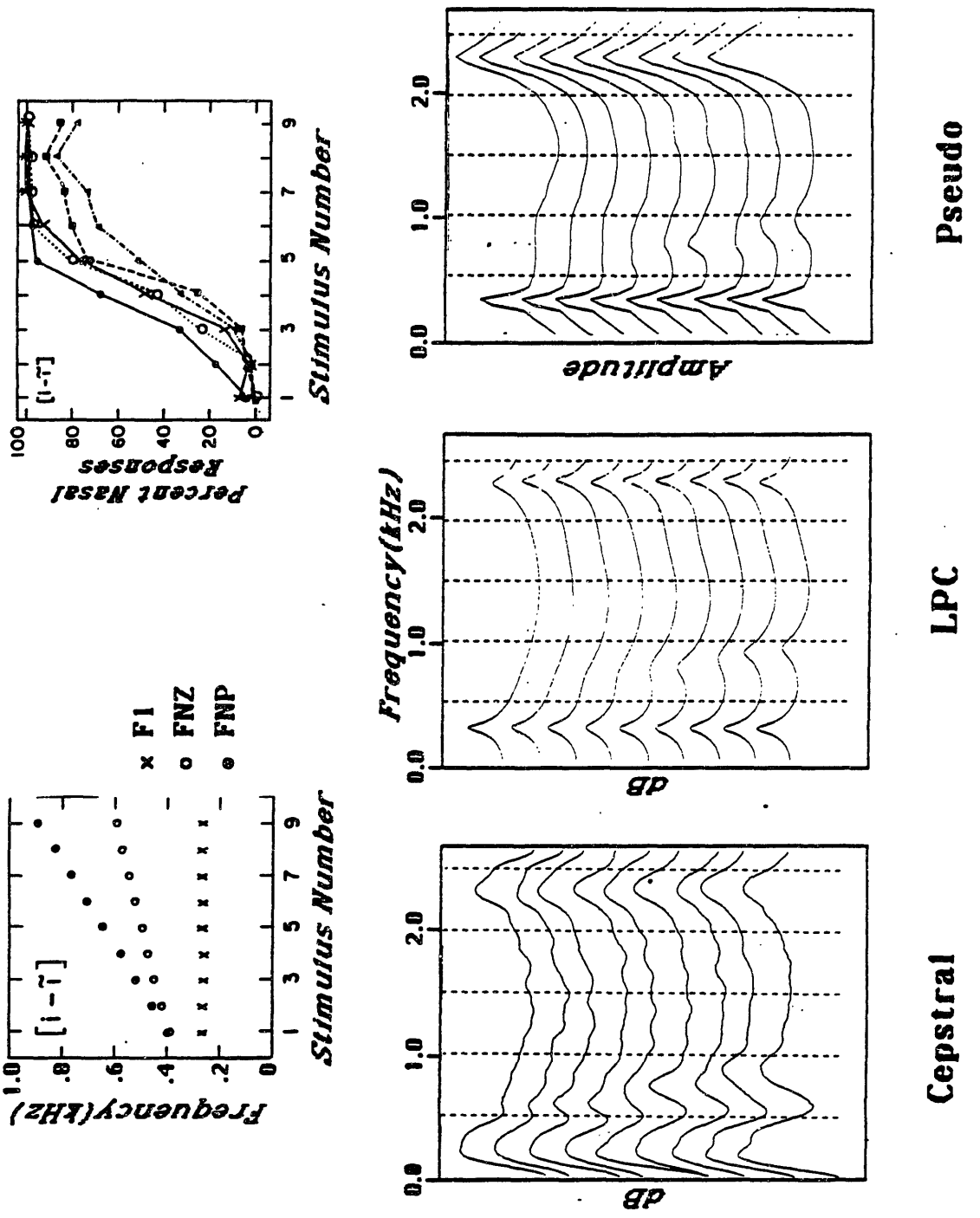

Figure 10.17: Results of cepstral analysis, LPC analysis, and pseudo spectral analysis of the synthetic "ti" stimulus series. Stimulus number increases in each case from top to bottom. Analysis was performed at a time slice near the end of the vowel in each case. The acoustic and perceptual data for /i/ are included at the top left and top right corners of the Figure for reference.

around 400 Hz for the first three tokens to around 300 Hz for the last three. This is in accord with the transition downward of the first formant frequency in the stimulus conditions. For the LPC analysis, the first formant region contains two peaks for the last two tokens, and a clear upper shoulder on F_1 for the preceding one. A prominent peak is only present for the first two tokens, with the middle four showing a single broad peak. Hence the natural groupings do not accord as well with the psychophysics. The cepstral analysis shows a prominent peak for the first two tokens, and a prominent valley for the last two. Between these is a gradual transition, with no clear prominences or boundaries.

Figure 10.17 shows the data for the /ti/ series. In this case both the LPC and pseudo spectral analysis obtain a steady peak at the first formant frequency, with very little change in amplitude or shape. For the LPC analysis, a second peak at the nasal pole shows up first as a shoulder on the first formant in the third token, and as a separate peak starting with the fifth token. The pseudo spectral analysis detects the valley for the zero near 500 Hz at token number 4, and this valley is maintained, with a growth upward in frequency, for the subsequent tokens. The pseudo spectral analysis detects only a broad, weak peak at the nasal pole, beginning at about token number 5. The cepstral analysis produces a very broad peak at the first formant frequency, for all tokens, but the valley for the nasal zero is pronounced, almost from the beginning. There is also a more prominent peak for the nasal pole than in either of the other two analysis methods.

Perceptual results for /ti/ were not as clear as for /to/, and seem to depend significantly on subjects' background. However, at least for the naive English and Gujarati subjects, the first three tokens are nonnasalized, and the last four are nasalized. Tokens 4 and 5 are ambiguous. This grouping fits quite well with all of the analysis methods, although the changes between groupings are not dramatic. Thus, for example, LPC shows little or no shoulder on F_1 for the first three tokens, and has a clear extra peak at the nasal pole for the last four. A similar story holds for the other methods. However, the results here indicate that the contrast between nasalized and nonnasalized would have to be realized in the case of a high vowel such as /i/ very differently from the realization for a low vowel such as /o/. With /o/, it was hypothesized that the prominence of the first formant could be a nasal indicator, whereas for /i/, the first formant varies little across all stimuli and for all three analysis methods. Instead, nasalization would be conveyed by the presence of a second peak in the spectrum somewhat above the first formant.

On the other hand, this particular stimulus series for /i/ may not be representative of typical nasalized high front vowels. In our experience with natural speech, it is usually the case that the first formant of a nasalized /i/ is substantially reduced in amplitude relative to a nonnasalized /i/, when subjected to pseudo spectral analysis [compare the /i/ in "He" with the /i/ in "cream" of the sentence in Figure 9.15]. The fact that some listeners had trouble identifying the nasal quality of these synthetic stimuli lends validity to this hypothesis.

In summary, the pseudo spectral analysis is able to extract the underlying pole-zero-pole complex in the F_1 region, and tends to show a distinct contrast between nasal and nonnasal vowels. This contrast is usually manifested by a lack of prominence of the peak at the first formant frequency, expressed either as an amplitude reduction relative to the amplitude of F_2, or as a dispersion of the peak energy over a diffuse region as contrasted with the presence of a sharp prominence for nonnasalized vowels.

Chap er 11

Alternative Forms for Spectral Representation

11.1 Introduction

In this chapter we will investigate a number of alternatives to the basic spectral analysis system as described in chapters 7 and 8, and illustrated in chapters 9 and 10. The first section examines the effect of inserting a filter to remove DC prior to the GSD computation for spectral analysis. This step is important for pitch processing, and, in the interest of keeping the two branches as similar as possible, we should determine the effect of a similar step on spectral processing. The second section shows the effects on the GSD analysis of modifying or leaving out certain stages of the peripheral model. The final section compares a number of alternatives to the GSD processing for synchrony measurement.

In the first section, the peripheral model is left unchanged from the original system, except that the final output of each channel is processed through a notch filter at DC. The removal of DC has a major effect on the numerator of the synchrony measure, but is almost irrelevant to the denominator. We will show that the addition of a half-wave rectifier at the output of the GSD results in an adequately constrained spectrum in this case.

The second section focuses on the peripheral aspects of the system, and examines which components of the peripheral system are essential to system performance. A computationally less expensive version which eliminates some aspects of the peripheral model may be adequate for certain purposes. In particular, an engineer who is only interested in producing a usable spectral representation for computer speech recognition may be able to make use of a simplified version of the peripheral model.

The final section discusses alternative forms for the synchrony measure, such as using autocorrelation as a synchrony measure, omitting synchrony detection altogether [i.e., using amplitude information only], and detecting synchrony to the adjacent filter output [i.e., cross-correlations rather than autocorrelations]. The main purpose of this section is to compare the synchrony measure that was used for the standard system with other methods, and to illustrate that, for the most part, its performance is superior to the performance of the other methods. Some of the other methods show some promise, however, and should be pursued in more depth before a definitive statement can be made.

Evaluation of the performance of a given system is always made in the context of certain biases. Our assumptions of "quality" are that the spectrogram representation should show clear peaks at the formant frequencies, preferably with prominent valleys between them. The spectrogram

should also be "smooth" in both frequency and time, and, in particular, it should be conceptually straightforward to extract the formant information from the given spectral representation.

11.2 Effects of Removing DC

The basic system for spectral analysis begins with a peripheral model consisting of a linear filtering step followed by a sequence of two AGC's, one with a fast τ followed by one with a slow τ, and a raised hyperbolic tangent as a half wave rectification scheme. In this section we compare a number of variants on the synchrony measurement process, keeping the model for the peripheral processing intact. Each of these variants includes the basic premise of a ratio of the integrated sum waveform over the integrated difference waveform. In the standard system, a threshold close to spontaneous rate is subtracted from the numerator, and the final output is the arc tangent of the ratio, as described in Chapter 8.

All of the variants that will be described begin with a filter to remove the DC level from the half-wave rectified waveform. This step is motivated by issues related to pitch detection. As was described in Chapter 7, the pitch waveform was filtered to remove the DC component prior to the application of the GSD algorithm. Such a step has major consequences on the definition of the GSD process. Now that the waveform is no longer guaranteed to be positive, the sum waveform must be interpreted as a filtering process, much like the difference waveform. In fact, adding the waveform to itself delayed by N samples is equivalent to processing the waveform through a linear filter consisting of a set of N zeros equally spaced around the unit circle. However, the zeros are located at $z = exp(j2\pi i/N + j\pi/N)$, i.e, precisely half-way between the zeros of the denominator. In the case of pitch extraction, the numerator becomes more useful when the DC component is removed, because now it can act as a filter to filter out selectively information **between** the proposed harmonics, accentuating the energy remaining at the proposed harmonic frequencies.

This effect might also be important for spectral analysis, particularly in the first formant region. For example, if the filter were centered at the second harmonic of the pitch, then the numerator would filter out information at the first and third harmonics, leaving the second harmonic component intact. The denominator, on the other hand, would filter out the second harmonic, leaving the first and third harmonic energy. Thus the second harmonic response would only be prominent in the final output if the second harmonic amplitude were sufficiently high, relative to the amplitudes of the first and third harmonics.

Given the premise that pitch and spectral analysis procedures should be as similar as possible, and given the above discussion, which suggests that removing DC could be useful for spectrum as well as pitch, it was felt to be appropriate to try removing the DC components from the individual filter outputs prior to processing them through the GSD's for the spectral analysis. The results were quite encouraging: it seemed generally that harmonics below the first formant frequency in females were somewhat reduced in amplitude relative to the original system. The process did not however completely remove additional harmonics below F_1. For the most part, there was not a substantial difference between spectrograms obtained from the filtered peripheral outputs and spectrograms

obtained through the standard system, as shown in Figure 11.1, for the word "intelligent", spoken by a female speaker.

A major problem with the removal of DC is that one loses control over the notion of a spontaneous rate to use as the threshold to be subtracted from the sum wave integration in the numerator. A big advantage of the raised hyperbolic tangent half-wave process is that weak sounds are essentially raised on a pedestal whose value equals the spontaneous rate level. An integration of the amplitude of these weak sounds will produce an output very close to spontaneous level. As the level increases, the shape of the original sine wave becomes distorted such that the negative half of the cycle remains small [it is constrained to be between zero and spontaneous level], while the level of the positive half of the cycle grows large [up to the limit of twenty times spontaneous in the standard system]. Thus the threshold to subtract from the numerator can be set to be slightly greater than spontaneous, assuring that weak inputs will result in a slightly negative level for the output.

In the case of a weak signal, the DC filter filters out the spontaneous level, and thus the amplitude of the sum waveform, in the case of perfect synchrony, is exactly twice the amplitude of the input signal. A fixed constant to subtract from the integrated sum waveform is no longer tractable, because synchronous signals that are very weak will produce a relatively large negative output, rather than an output that is constrained to be only slightly less than zero. Particularly in the case of high front vowels, the spectral region between the first two formants will often show deep negative valleys.

A simple solution to this problem is to pass the final pseudo spectral output through another raised hyperbolic tangent, remapping large negative values to zero, increasing the gain in intermediate regions, and mapping large positive values to a saturation level. This step results in a tightly constrained spectral representation. Typically there are low flat valleys between formants, with very rapid slopes at formant edges, leading to peaks that can tend to become flattened if the saturation level is set too low. If the parameters of the half-wave are selected properly, it is possible to obtain a spectral representation that shows prominent features and maintains a limited range of values.

The above concepts are illustrated in Figure 11.2, for the word "variety", spoken by a female speaker. Two cross sections are examined, in the /*a*/ [left] and in the /i/. Figure 11.2a shows narrow-band spectra for reference. Figure 11.2b shows the results of processing through the standard system. In Figure 11.2c, the DC removal has been included, but not the final hyperbolic tangent half-wave. Figure 11.2d shows the results of including the final hyperbolic tangent half-wave rectifier, replacing the arc tangent of the original system. In the region between the formants in the /i/ and in the low frequency region of the /*a*/, the amplitude of the pseudo spectrum dips substantially below zero in (b). This problem is solved by replacing the final arctan with a hyperbolic tangent. It is clear, at least for this utterance, that the complete system, including the DC filter and the final hyperbolic tangent half-wave rectifier, can produce prominent peaks with deep intermediate valleys.

Another possible solution to the problem of large negative numerators is to add a small constant

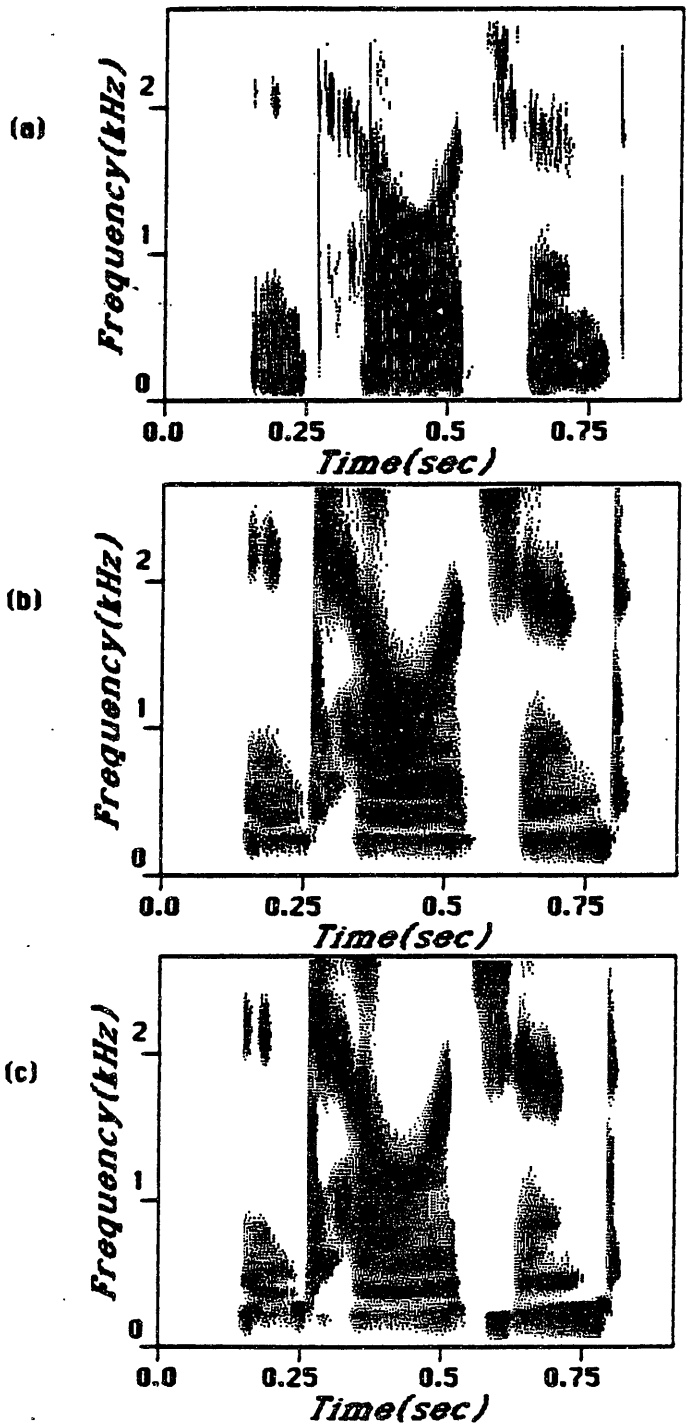

Figure 11.1: Comparison of pseudo spectral analysis with and without removal of DC component of peripheral level outputs.
a) Wide-band spectrogram of word "intelligent", spoken by a female speaker.
b) Pseudo spectrogram of same word as in (a).
c) Pseudo spectrogram obtained by first passing peripheral stage outputs through a filter to remove the DC component.

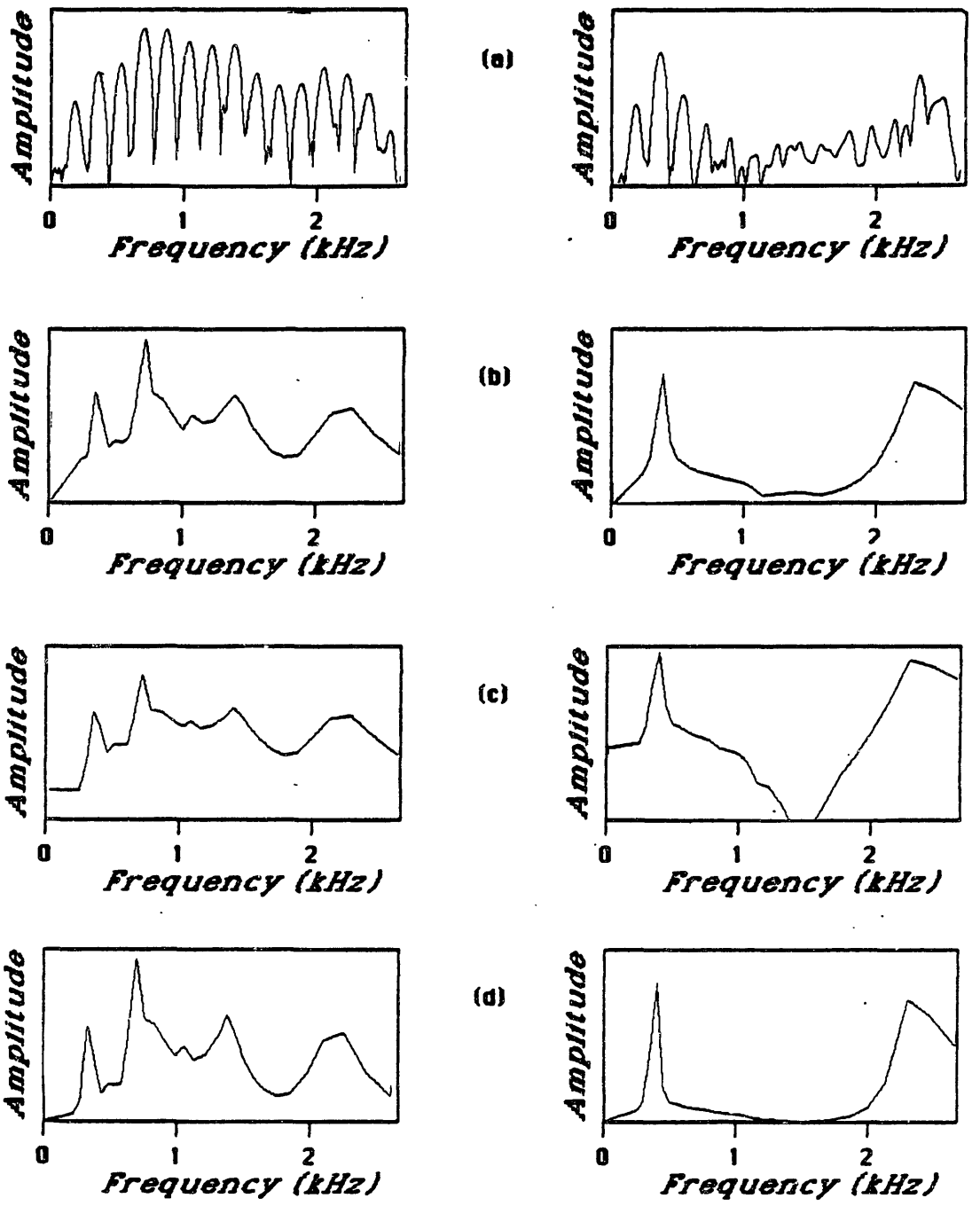

Figure 11.2: Example illustrating effect of removing the DC component from the peripheral stage outputs on the final system. Cross sections are taken in the /*a*/ [left] and the /i/ of "variety", spoken by a female speaker. a) Narrow-band spectral analysis [78 Hz Hamming window]. b) Standard system. c) Results obtained when DC component is filtered out of peripheral level outputs. d) Same as (c), except final arctan soft-limiter has been replaced with a hyperbolic tangent, the same one that was used for the half-wave rectification in the peripheral model.

to the denominator rather than subtracting a small constant from the numerator. With this approach, the spectrum is guaranteed never to be negative, and no hyperbolic tangent half-wave is necessary. Unfortunately, it is not possible to select a single constant for this strategy which will successfully suppress responses at weak inputs without also grossly reducing the prominences at the formant resonances. The problem is that the denominator can become very small when the input is synchronous, such that the constant that is added is substantially larger than the difference waveform level, even when the input signal is strong. Thus the constant plays a major role in reducing the amplitude at precisely those times when it is most important for the amplitude to be strong.

In summary, the pitch and spectral processing methods can be made to be more similar by removing the DC component from the input to the synchrony measure in both cases. The two structures are still not completely identical, because it is necessary to suppress a response to weak signals in the case of spectral analysis, but not in the case of pitch analysis. This suppression cannot be achieved by adding a constant to the denominator, because such a strategy has a severe effect in attenuating the response when synchrony is excellent. Subtracting a constant from the numerator is effective, except that very weak signals get mapped into large negative outputs. The inclusion of another hyperbolic tangent half-wave rectifier at the final output stage not only solves this problem but also enhances the contrast between peaks and valleys in general, producing spectral outputs with sharp contrasts.

11.3 Effects of Changes in Peripheral Model

In this section, we will examine the effects on the synchrony outputs of altering or eliminating altogether certain aspects of the peripheral model. The effects will be demonstrated by way of examples of spectrograms and spectral cross sections in selected speech tokens.

Perhaps surprisingly, the GSD algorithm is remarkably insensitive to distortions on the wave-shape of the input waveforms, introduced by the parts of the peripheral model that follow the linear filters. The **filter** characteristics, on the other hand, can affect the spectral prominences of the formant peaks in a major way. Furthermore, the delay τ_c must be an accurate reflection of the center frequency of the filter. The delay is necessarily quantized to integer sample units, and for some of the higher frequency filters, it was necessary to **upsample** the peripheral outputs to 32000 Hz in order to generate the appropriate delay τ_c.

Given the discussion in the preceding section about removing the DC component from the peripheral level outputs, it is perhaps not unexpected that the GSD algorithm can deal adequately with waveforms that have not been subjected to half-wave rectification at all. In fact, if the peripheral model is reduced to a bank of linear filters, with no AGC's or half-wave-rectification, the final output is not altered by much. The major effect is that weak sounds fall below threshold, with onsets being affected the most.

Some of the effects described above are illustrated in Figures 11.3 through 11.9. Figure 11.3 shows what happens if the high frequency filters are not upsampled to 32000, but instead the

nearest integer delay at a 16000 Hz sampling rate is used in the computation of the GSD. The sentence is "The young kid jumped the rusty gate", spoken by a male speaker. Part (a) shows the wide-band spectrogram, Part (b) shows the result of processing through the standard synchrony analysis, and Part (c) shows the results when the delay for the synchrony measure is quantized to the nearest τ in units of 1/16 ms [16000 Hz]. In the rapid motion of F_2 of the word "young" and the word "jumped", it should be evident that the quantized version moves in a staircase fashion.

The effects of modifying the filter characteristics and of omitting either the half-wave or the AGC's are illustrated for the word "intelligent" in Figure 11.4. Two modifications of the filter characteristics are illustrated: broader tails on the low frequency side of the high frequency filters, and the addition of high frequency tails on the low frequency filters [below 500 Hz]. The filter characteristics for the original system and the two modified systems are shown in Figures 11.5a, b, and c.

Six different cases are investigated in Figure 11.4 as follows:

(i) Standard Fourier Analysis,

(ii) Original System,

(iii) System with Filters as in Figure 11.5b,

(iv) System with Filters as in Figure 11.5c,

(v) System with Original Filters but no half-wave rectification, and

(vi) System with AGC's omitted, and thus linear filter outputs feed directly into the hyperbolic tangent half-wave rectifier.

Figure 11.4a shows pseudo spectrograms for the entire word for the six cases, and Figures 11.4b and 11.4c show pseudo spectra at the time slices indicated by the vertical bars in Figure 11.4a, in the /ɛ/ and the /ɪ/.

In the case of the broader filter characteristics, (parts (iii) of the Figure) the definition of critical bandwidth in terms of the 3dB points is preserved, but the slopes are reduced on both the upper and lower sides. It is clear that the second formant is very weak, particularly in the stressed syllable, where it encroaches upon F_1. Furthermore, the third formant is almost nonexistent. This effect is even more evident in the two spectral cross sections than in the pseudo spectrogram. Such a loss of prominence of formant peaks is undoubtedly undesirable.

Parts (iv) of the Figure show the results of adding broad tails on the high frequency side of the low frequency filters. The effects are limited to filters below 500 Hz, but it is clear, particularly in the cross-section in the /ɛ/, that the glottal formant peak below F_1 has been substantially reduced.

Parts (v) of Figure 11.4 show the effects of removing altogether the AGC's, and parts (vi) show the effects of removing the hyperbolic tangent half-wave rectifier. When one of these two is removed, the other then takes on the entire load of compressing the dynamic range, but both are effective in achieving this compression. Removing the half-wave rectifier is not conceptually significantly different from first performing the half-wave rectification and then removing the DC component. The major difference is that in the former case the original waveshape is preserved, and negative and positive halves of the sine wave cycle are thus more similar to each other. Without the AGC's, the half-wave rectifier can turn the waveform into essentially square waves, at least when

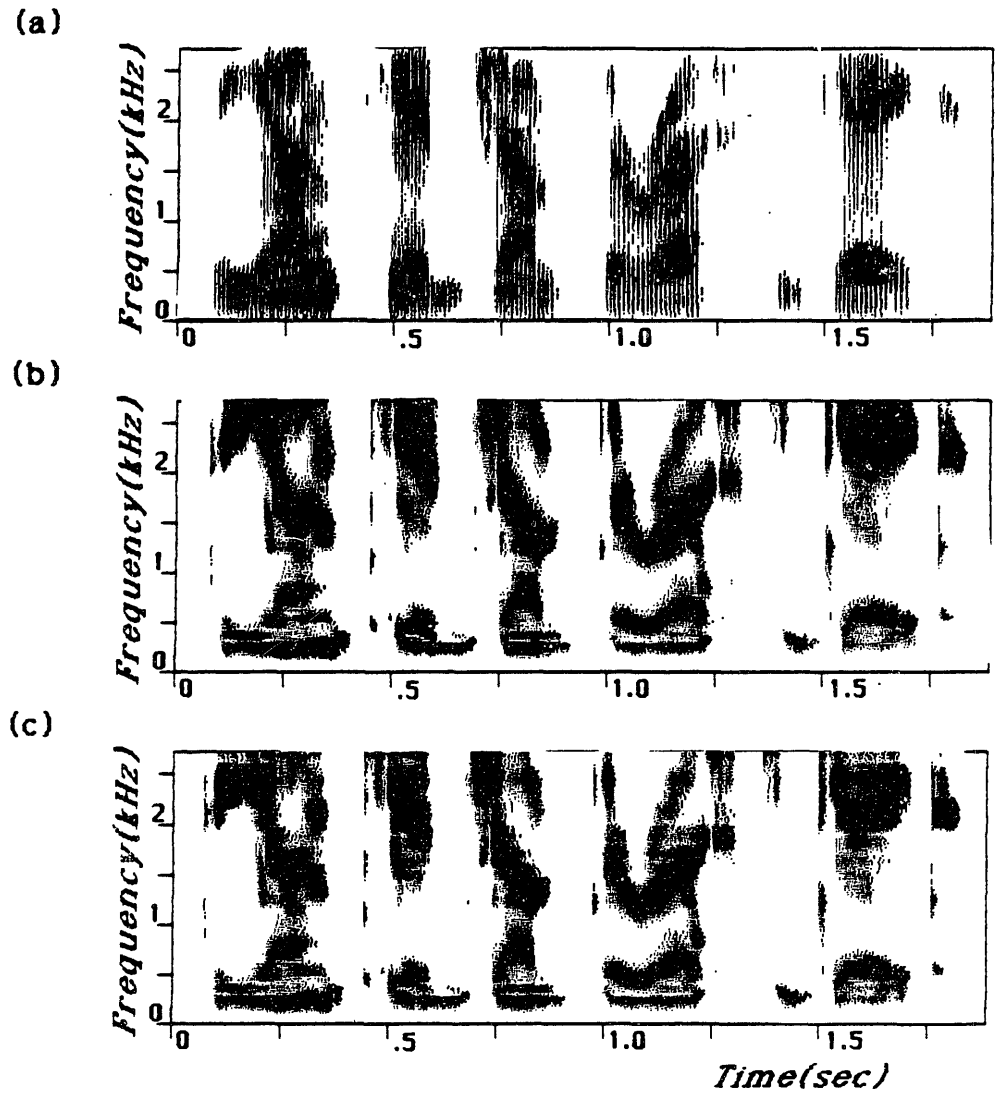

Figure 11.3: Example illustrating the effect of quantizing high-frequency filter taus to integer multiples of 1/16 ms [16000 Hz sampling rate]. a) Wide-band spectrogram of the sentence, "The young kid jumped the rusty gate", spoken by a male speaker. b) Standard pseudo spectrogram for same utterance, where the filter outputs of the high frequency filters are upsampled to 32000 Hz to obtain more accurate delay for GSD algorithm. c) Pseudo spectrogram obtained by leaving all high frequency filter outputs at the original 16000 Hz sampling rate, thus forcing inaccurate delays for some of the filters.

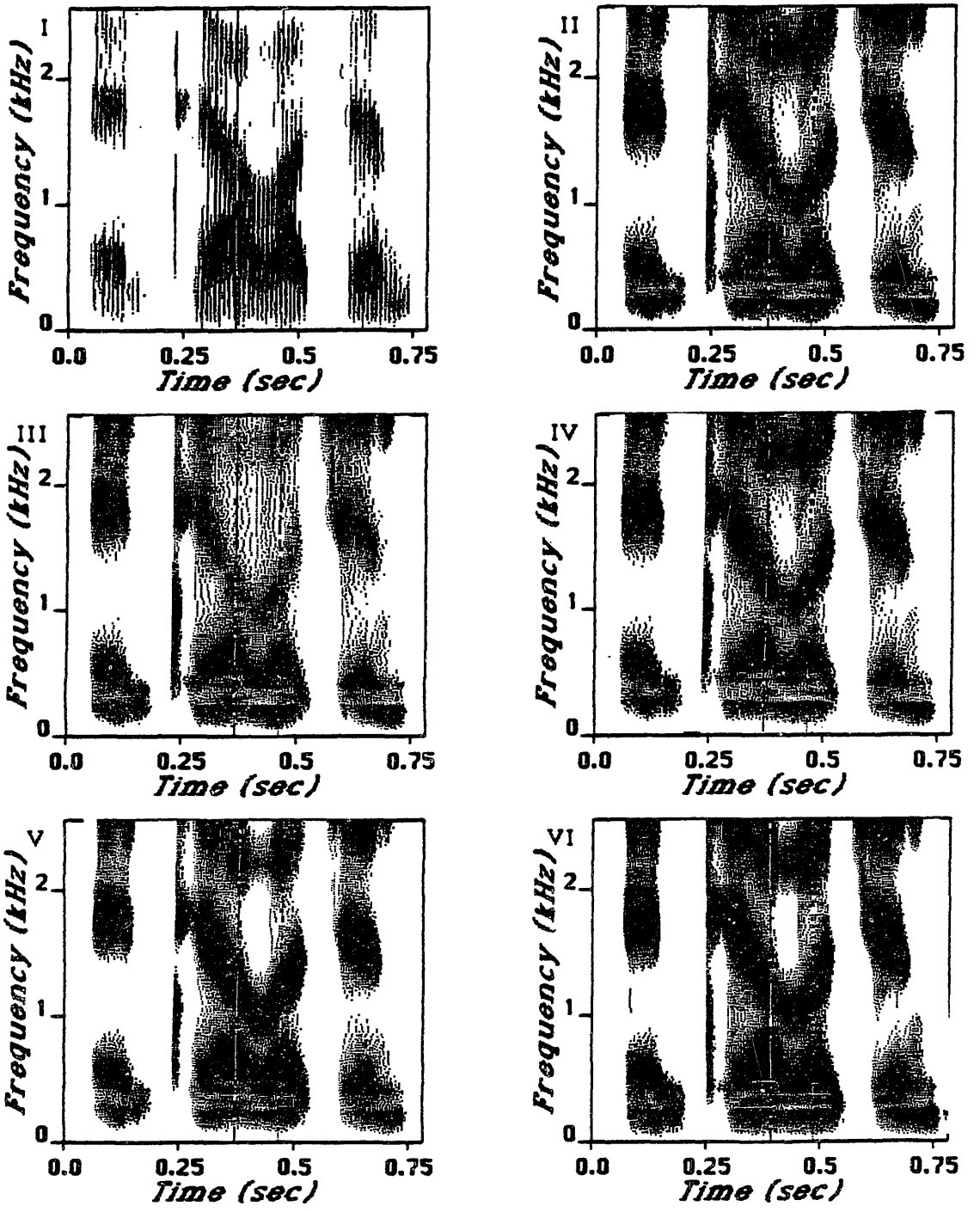

Figure 11.4: Example illustrating the effects of removing certain aspects of the peripheral model. For each part, the following conditions hold:
i) Standard Fourier analysis, ii) Original peripheral model, with filters as in Figure 11.5a, iii) System with filters modified as in Figure 11.5b, iv) System with filters as in Figure 11.5c, v) System with original filters but no half-wave rectification, and vi) System with AGC's omitted, and thus linear filter outputs feed directly into half-wave rectifier.
a) Wide-band spectrogram and pseudo spectrograms for a series of different cases for the word "intelligent", spoken by a male speaker.

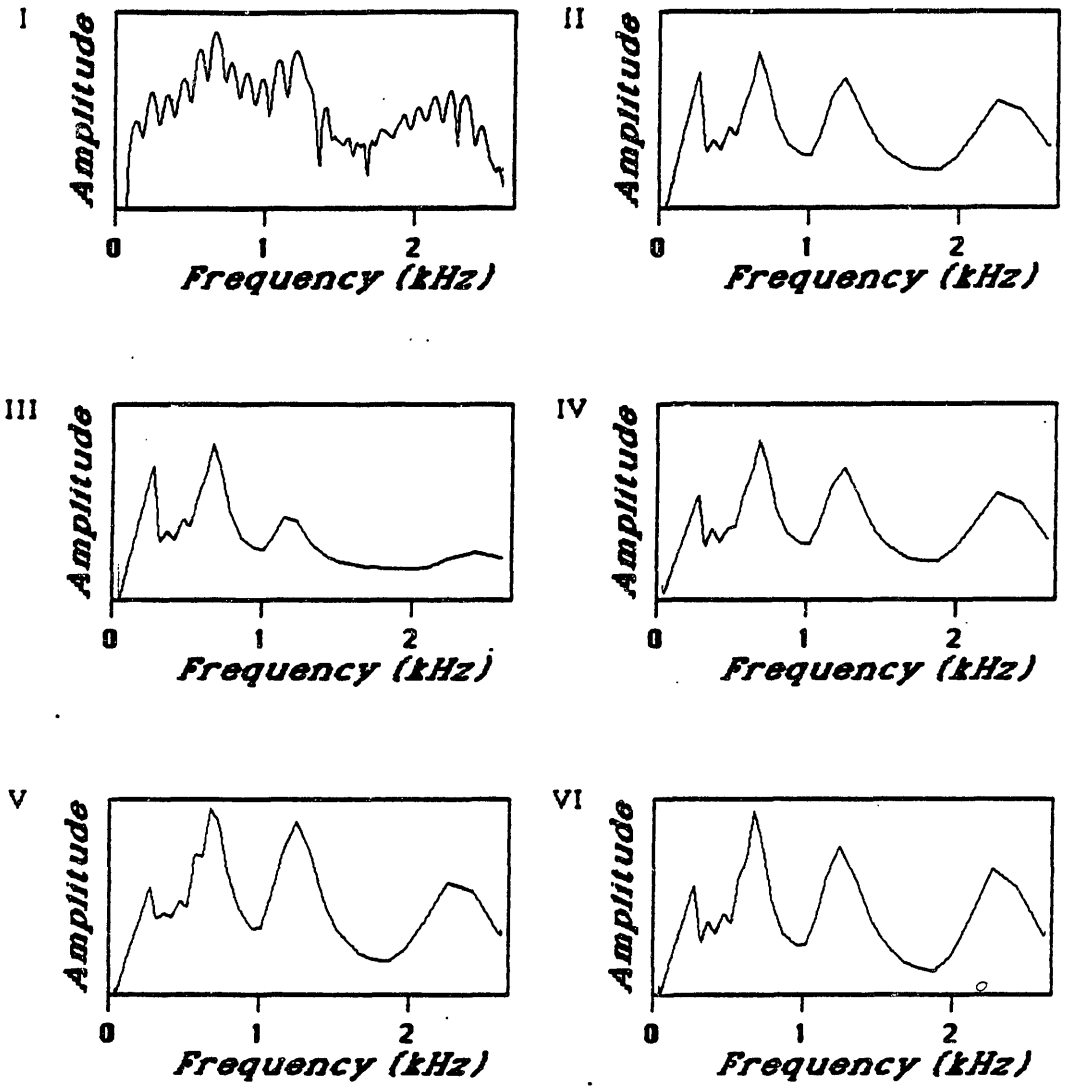

Figure 11.4:
b) Cross sections taken at point marked by vertical bar in Part (a) during /ɛ/ in "intelligent", spoken by a male speaker.
Parts (i) through (vi) as defined in Part (a).

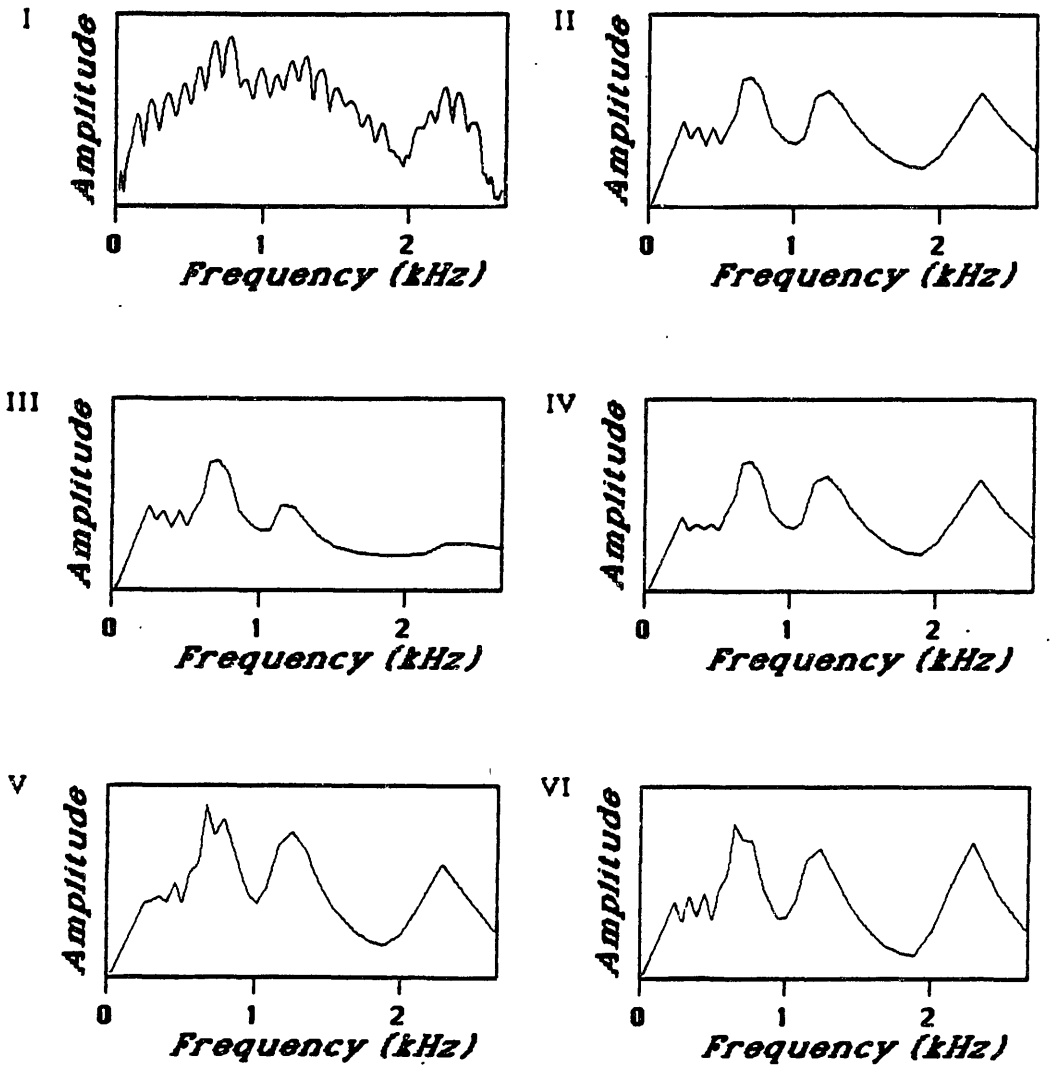

Figure 11.4:
c) Cross sections taken at point marked by vertical bar in Part (a) during /ɪ/ of "intelligent". Parts (i) through (vi) as defined in Part (a).

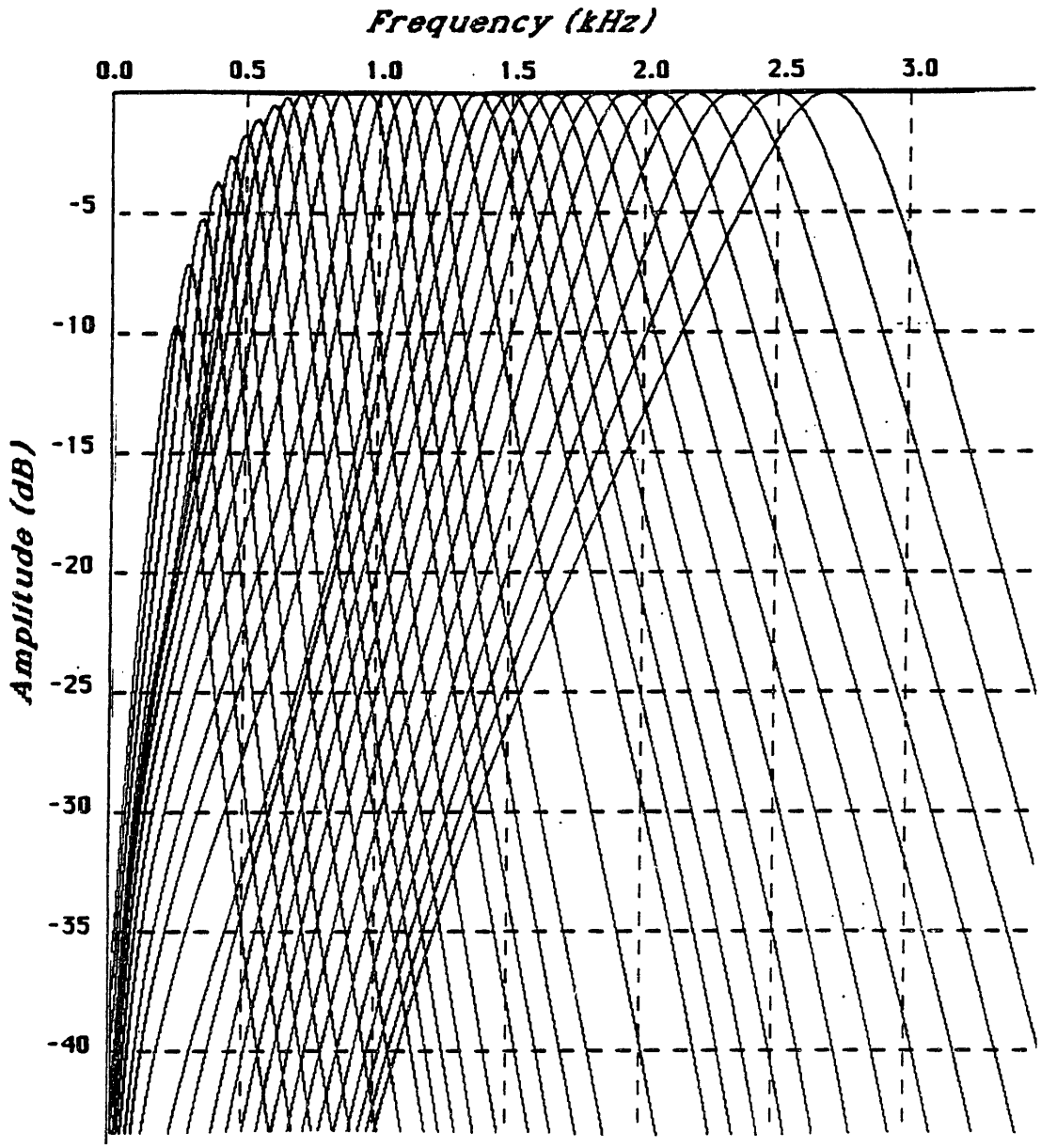

Figure 11.5: Filter characteristics for systems illustrated in Figure 11.4, Parts (ii), (iii), and (iv), to show effect of broader filters on synchrony measure.
a) Standard filter characteristics, displayed on linear frequency scale.

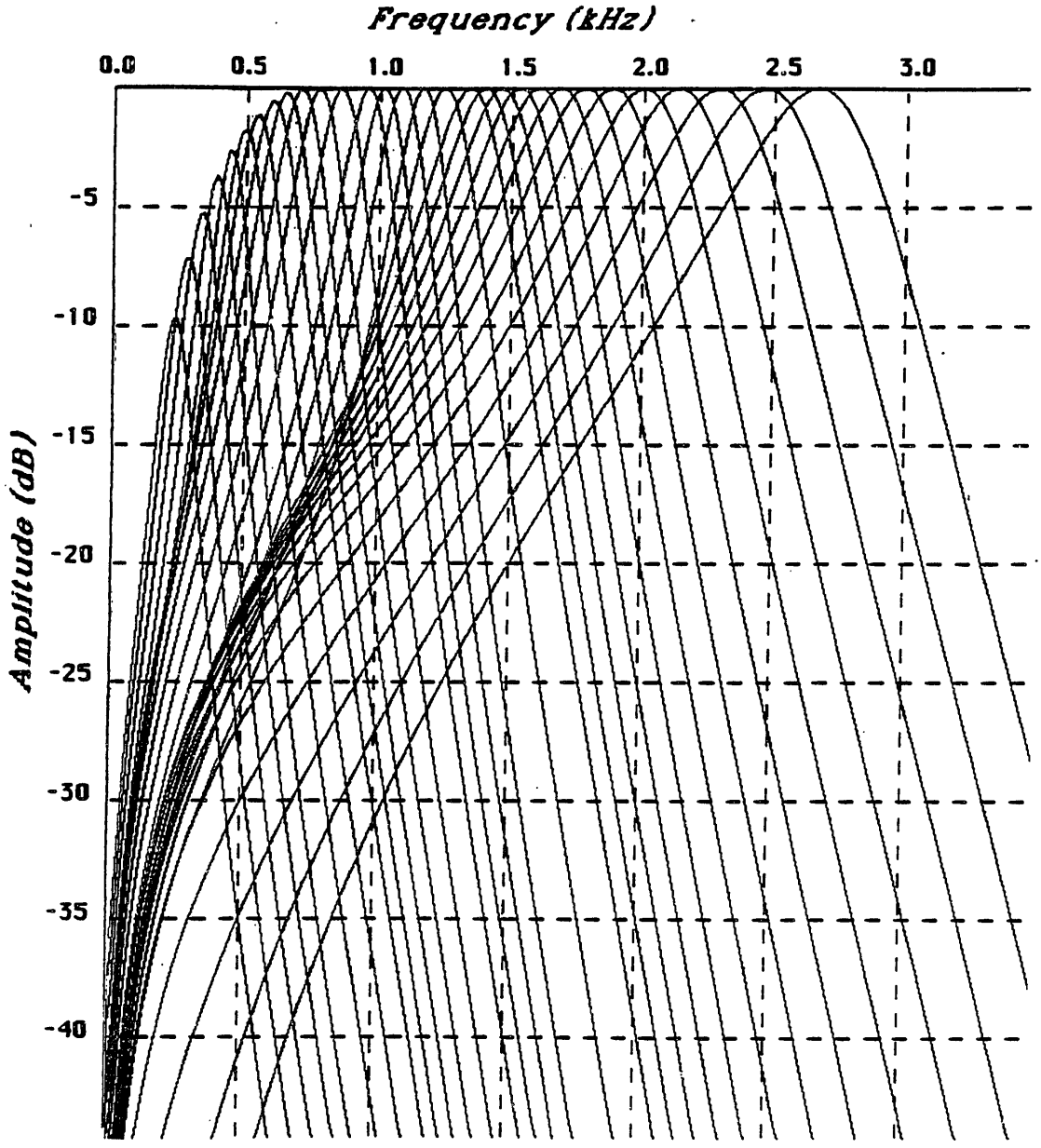

Figure 11.5:
b) Filter characteristics used for Figures 11.4, Part (iii). The filters centered above about 1000 Hz have broader low frequency tails than in Part (a).

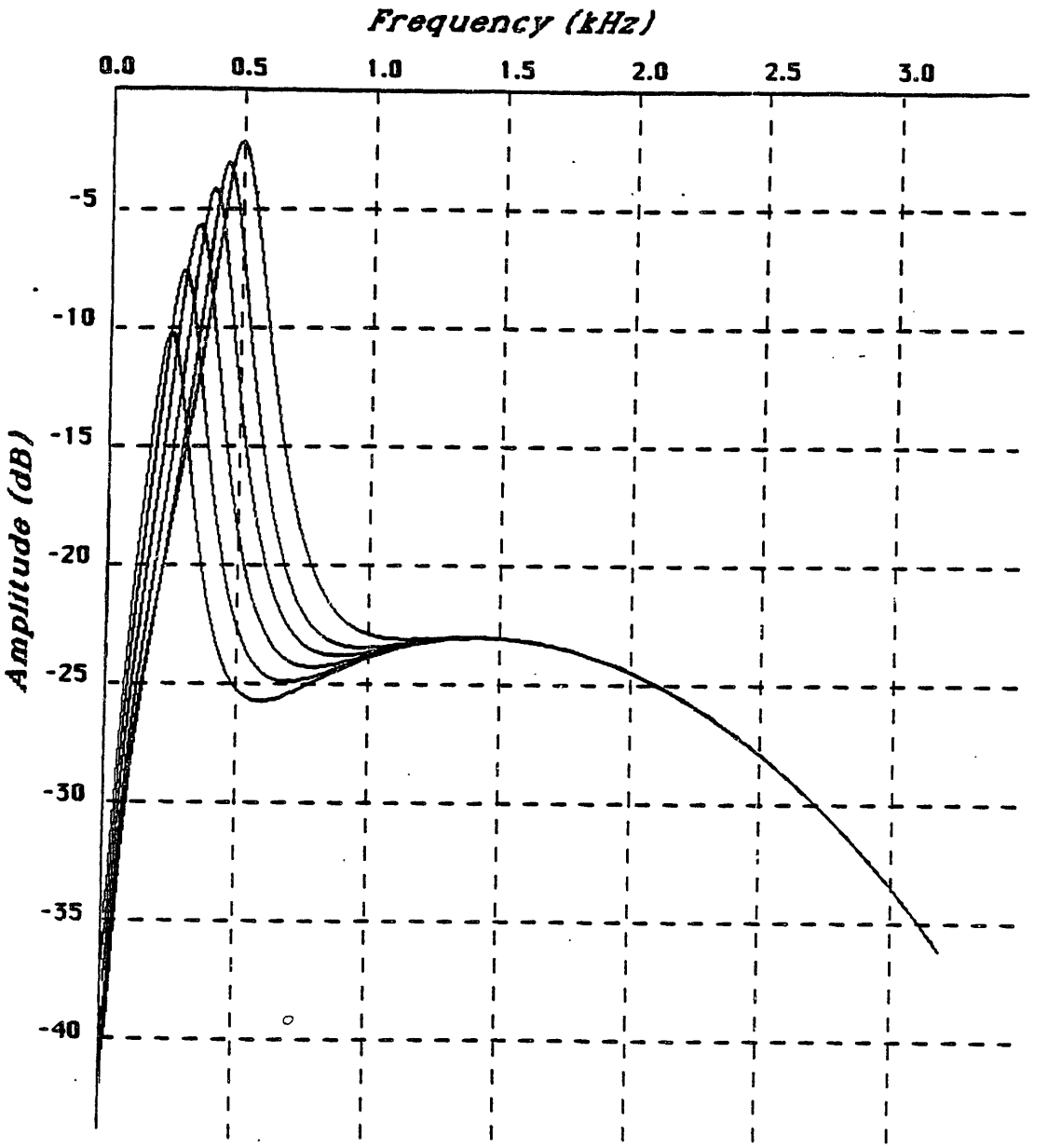

Figure 11.5:
c) Filter characteristics used for Figures 11.4, Part (iv). Only the filters centered below 500 Hz were modified, by adding a broad tail on the high frequency side.

the output of the peripheral filter is strong. But in general these square waves are still strongly synchronous with the underlying sine wave period.

Such a square wave effect is illustrated in Figure 11.6, for a spectral cross section in the syllabic /*a*/ of the word "hospital", spoken by a female speaker. The narrow-band spectrum is shown in Figure 11.6a, and is to be compared with the pseudo spectrum below it in Figure 11.6b. This pseudo spectrum was obtained from a system with the AGC's omitted. The output of the filter centered at 700 Hz, is shown at the top right in Figure 11.6c. After the half-wave rectification [Figure 11.6d], the wave shape becomes square, because there were no preceding AGC's to compress the waveform while preserving its shape. Figure 11.6e shows the rectified output after the waveform in Figure 11.6d has been subtracted from the waveform delayed by 1/700 sec, the center period. Although there are sharp spikes at the edges where adjacent square waves are not exactly the same size, there is still a substantial reduction in energy in the waveform of Figure 11.6e relative to the waveform of Figure 11.6d. Hence the pseudo spectrum displays a strong peak at the 700 Hz formant frequency.

Figure 11.7 shows a wide-band spectrogram (a) compared with a standard pseudo spectrogram (b) and a pseudo-spectrogram computed when the AGC's are omitted in the peripheral stage (c). The utterance, "He ordered peach pie with ice-cream", was illustrated in detail in Chapter 9. It appears that little has been sacrificed in terms of overall performance in this case, and in fact such is generally the case. The amplitude of the peaks is usually somewhat higher, which is reflected in a broader black region around formant frequencies. The contrast between weak sounds and strong sounds is greater; as a consequence, for example, the first formant region of the nasalized /i/ in "cream" is **weaker** in part (c) than in part (b).

The AGC's are occasionally useful, however, as shown in Figure 11.8, which compares spectral cross sections in the /*a*/ of "variety" spoken by a female speaker. At the top is the narrow-band spectrum, and in the middle is the pseudo spectrum when the AGC's are included. At the bottom is the result of omitting the AGC's. In this case, the narrow region between the first and second formants has been badly disrupted as a consequence of distortions in the waveshape introduced by bypassing the AGC stage.

Figure 11.9 illustrates the result of omitting both the AGC's and the half-wave rectifier, and therefore feeding the outputs of the peripheral filters directly into the GSD algorithms. The phrase is "paste can cleanse", spoken by a male speaker. Figure 11.9a shows the wide-band spectrogram, and Figure 11.9b shows the standard pseudo spectrogram. Figure 11.9c shows the pseudo spectrogram applied directly to the linear filter outputs. The differences between parts (b) and (c) are subtle; mostly they concern the reduction of amplitude at onsets. Thus, for example, the /k/ of "cleanse", and, to a lesser extent, the /p/ of "paste" and the /k/ of "can", do not show up in part (c) as a sharp line onset. Although this difference may not seem great, it is probably extremely important for a successful identification of the stop consonants. The sharp onsets provide an important mechanism for locating the burst in time, and the features that would lead to an identification of the burst are probably better expressed if the nonlinearities are included in the peripheral model. Without the enhancement of energy at onset, there is a danger that all filter outputs will be too weak, even those that are at a local spectral peak.

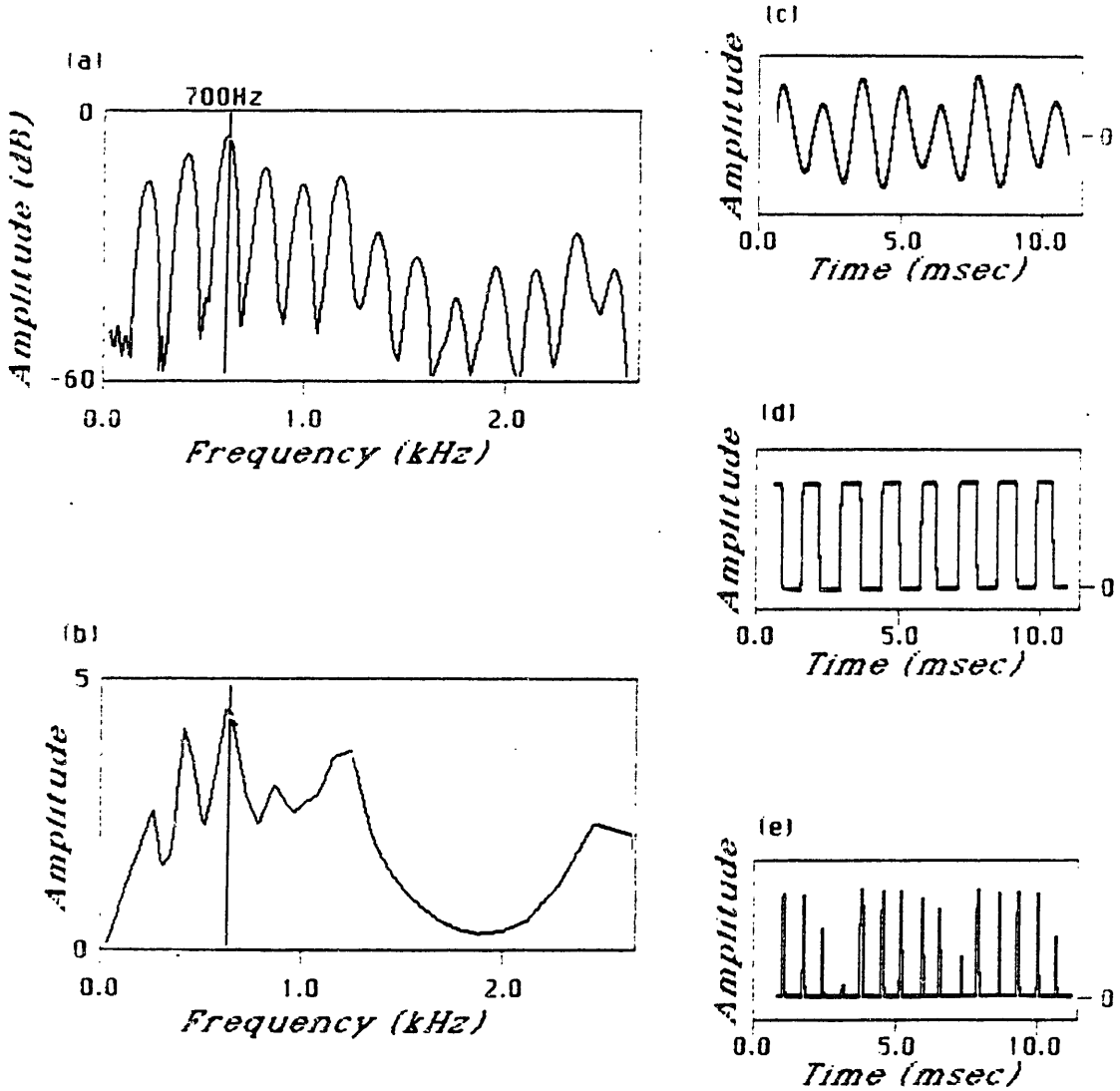

Figure 11.6: Illustration of why omitting AGC's does not necessarily have a major effect on the synchrony outputs.
a) Narrow-band spectral cross section taken in /ɑ/ of "hospital", spoken by a female speaker.
b) Pseudo spectral cross section taken at same place as (a), when AGC's of peripheral model have been omitted. The vertical bar is at 700 Hz, the frequency of the first formant.
c) Output of linear filter centered at 700 Hz, for example in (a).
d) Output after hyperbolic tangent half-wave rectification. Because there is no AGC, the wave shape has been converted to square.
e) Difference waveform when waveform in part (d) is subtracted from itself delayed by 1/700 sec.

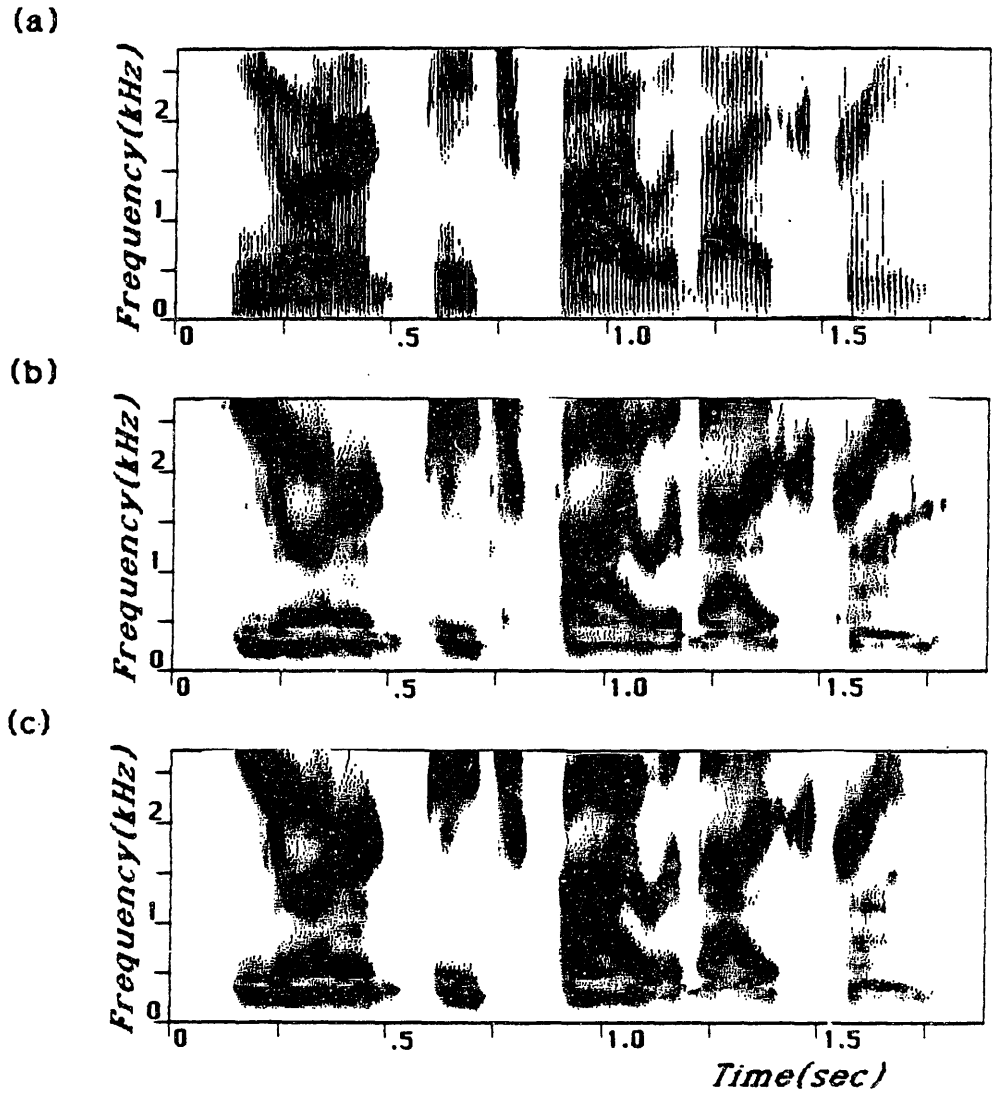

Figure 11.7: Example showing that synchrony system is robust against distortions on the waveshapes of the inputs from the peripheral model. The sentence is "He ordered peach pie with ice-cream", spoken by a male speaker.
a) Wide-band spectrogram
b) Pseudo spectrogram with standard system
c) Pseudo spectrogram, with AGC's omitted from peripheral model.

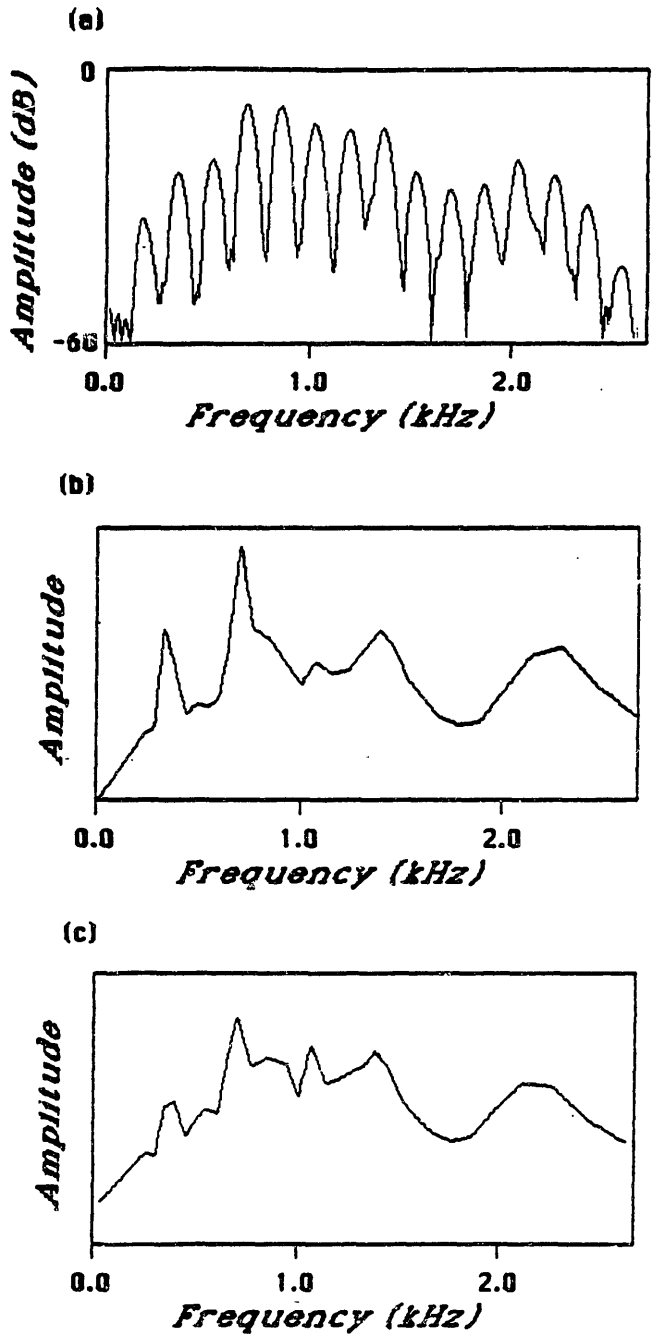

Figure 11.8: Example where lack of AGC's in peripheral model makes a difference.
a) Narrow-band spectrum in /æ/ of "variety", spoken by a female speaker,
b) Pseudo spectrum at same place as (a), with standard peripheral model, and
c) Pseudo spectrum obtained when AGC's are omitted from peripheral model.

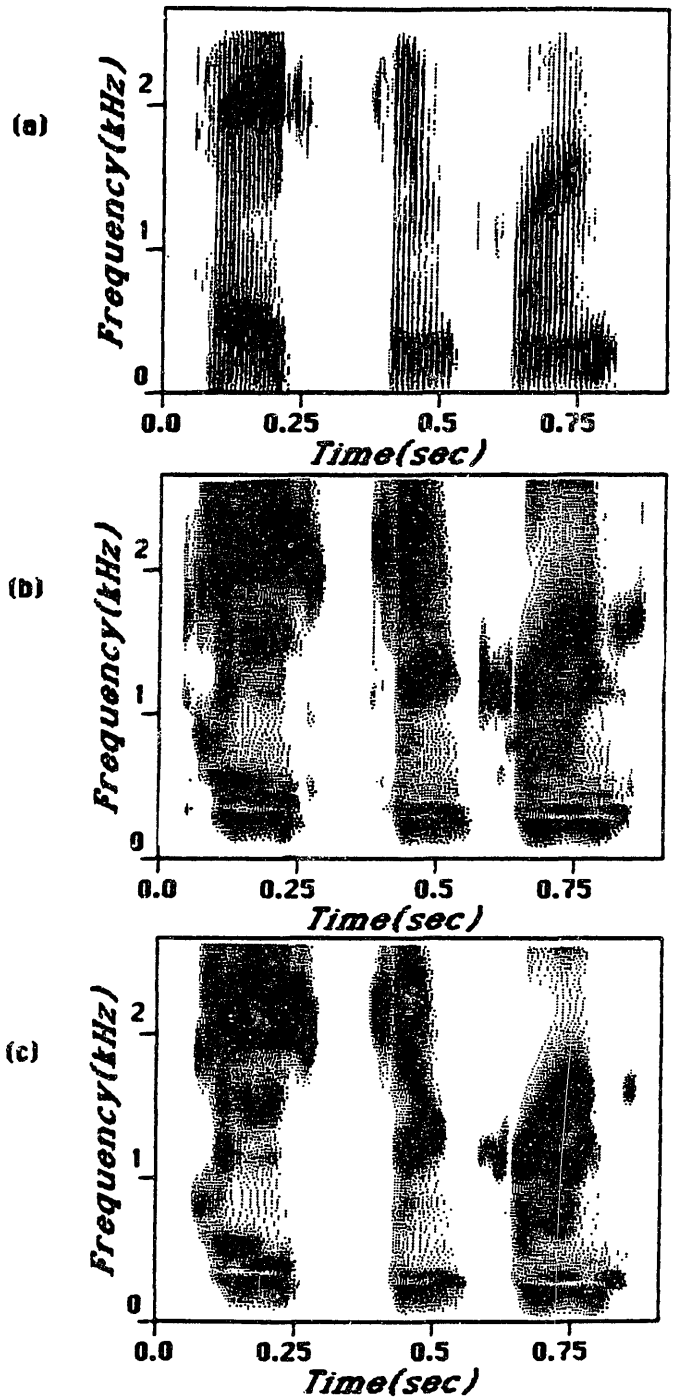

Figure 11.9: Example illustrating effect of leaving out everything in peripheral model except linear filters. a) Wide-band spectrogram of the phrase "paste can cleanse", spoken by a male speaker. b) Pseudo spectrogram of same phrase as in (a). c) Pseudo spectrogram obtained when peripheral model is simplified to a bank of linear filters, with no AGC's and no half-wave rectification. The major difference between (b) and (c) is that onsets are less pronounced in (c) [compare the /k/ of "cleanse"].

11.4 Alternative Forms of Synchrony Detection

In this section we will examine briefly the consequences of replacing the standard synchrony measure with other methods, while leaving the peripheral model intact. We will show that methods based on the energy domain are in general less effective than methods operating in the magnitude domain.

An itemized list of the alternative methods is given below. In each case, the appropriate function is passed through an envelope detection process, symbolized by $< >_r$, which is identical to the one used in the standard system, implemented by means of a double pole on the x-axis. Thus, the integration window is infinitely long, with a peak at r [r was always fixed at 4 ms]. $y[n]$ is the output of the peripheral model for a particular channel, and $z[n]$ is the output of the "synchrony" measure, which may not in fact be a measure of synchrony at all in some cases. When specified, n_0 is a delay in samples equal to the center period of the filter.

1. Log Energy in y[n]

$$z[n] = 10 \log < y^2[n] >_r \tag{11.1}$$

2. Log Autocorrelation Function at period τ_c

$$z[n] = 10 \log < y[n]\, y[n - n_0] >_r \tag{11.2}$$

3. Log Magnitude of y[n]

$$z[n] = 20 \log < |y[n]| >_r \tag{11.3}$$

4. Log Autocorrelation Square Root Function at period τ_c

$$z[n] = 20 \log < \sqrt{(y[n]\, y[n - n_0])} >_r \tag{11.4}$$

5. Square Root Normalized Autocorrelation Coefficient

$$\begin{cases} z_1[n] &= y[n]\, y[n - n_0] \\ z_2[n] &= y[n]\, y[n] \\ \\ z[n] &= \sqrt{\dfrac{< z_1[n] >_r}{< z_2[n] >_r + \delta}} \end{cases} \tag{11.5}$$

6. Normalized Autocorrelation Square Root Coefficient

$$\begin{cases} z_1[n] &= \sqrt{y[n]\, y[n - n_0]} \\ z_2[n] &= |y[n]| \\ \\ z[n] &= \dfrac{< z_1[n] >_r}{< z_2[n] >_r + \delta} \end{cases} \tag{11.6}$$

7. Comparing Adjacent Filters

$$z_k[n] =< |y_k[n] - y_{k-1}[n]| >_r \tag{11.7}$$

where $y_k[n]$ is the output of the peripheral channel centered at frequency ω_k.

It is important to note that method [3] differs from method [1] by much more than the simple factor of 1/2 that relates the log of a number to the log of its square root, and likewise a simple square root relationship does not hold between methods [2] and [4]. The square root is computed on a sample by sample basis before integrating over time, and the two results are therefore quite different.

The seventh definition involves a comparison of adjacent filter outputs, as suggested by Prof. Merzenich of the University of California [personal communication]. It represents a major departure from the normal synchrony definitions, and depends heavily upon the phase characteristics of adjacent filters, and also upon how far apart "adjacent" is defined to be. Prof. Merzenich suggested 0.4 Bark spacing, which was used for the experiment reported here. Of the several possibilities that were tried for defining the comparison, this simple magnitude of a difference turned out to produce the most promising spectral representation of this sort.

The above seven methods were compared with each other and with the standard synchrony model of the thesis. Methods based on square roots generally produce much smoother spectra than methods operating in the squared domain. We will show later that the hyperbolic tangent half wave function in the peripheral model is the major contributor to enhancing the contrast between these two types of methods. The two methods that include amplitude normalization, methods [5] and [6], produce more promising results than methods [1] through [4], that are based on absolute amplitudes. Method [7] is a novel departure from the other methods, in that it compares adjacent filter outputs rather than delayed versions of the same output. Attempts were made to define an amplitude-normalized version of this method, but with little success.

Figure 11.10 compares the various methods pair-wise using a spectrogram-like representation, in conjunction with two spectral cross sections taken at the two time slices indicated in the spectrogram, during the /a/ and during the off-glide. Figure 11.10a shows the standard wide-band spectrogram and narrow-band spectral cross sections of the word "desire", as a reference for evaluating the rest of the methods. This word was chosen for illustration because it contains a reduced syllable, very weak in energy, and a strong syllable, containing relatively rapid formant motion. A processing method that can enhance the peaks at the formants in the /i/, and produce clear tracks of the formants in the second syllable, would be considered good.

Figure 11.10b compares and contrasts methods [1] and [2], which both integrate in the energy domain. Method [2] involves R_i instead of R_0, and is therefore a synchrony method. Formant peaks are indeed somewhat enhanced relative to method [1], but in both cases there are major irregularities in the spectrum, related to the fundamental frequency, that appear to have been introduced by the hyperbolic tangent half wave.

Methods [3] and [4] are illustrated in Figure 11.10c, which differ from methods [1] and [2] respectively in that a square root is inserted prior to the integration step. It is remarkable that both of these methods show very few horizontal bars in the spectrogram, such as are characteristic of the first two methods. Method [4], which involves synchrony detection, shows improved spectral resolution over method [3], which does not use synchrony, although additional irregularities are introduced into the spectrum with the synchrony measure.

Methods [5] and [6], which are both amplitude-normalized schemes, are illustrated in Figure 11.10d. Method [6] is basically an amplitude-normalized version of method [4], and in fact the spectral cross sections look very similar to the ones for method [4]. Because of the normalization, the amplitude of the unstressed first syllable is enhanced relative to that in method [4].

Analogously, method [5] relates to its counterpart method [2]. However, in this case taking the ratio of R_i over R_0 has a larger effect in terms of reducing the prominence of the horizontal bands of energy. The upper cross section, taken at the /a/, preserves the salient formant information, and has lost most of the roughness that was evident in method [2].

Figure 11.10e compares method [7] with the standard GSD method. Method [7] is a measure of synchrony of adjacent filter outputs, which is quite a bit different from other methods examined here. The steep edges on the high side of the filters should have an interesting effect. Thus the detailed wave shape of the output of a filter just below a formant should be quite different from the wave shape of the output of an adjacent filter which includes the formant frequency. The filter just above the formant should, however, give a response that is quite similar to the response at the formant frequency. A strong difference in waveshape is reflected in a strong response in the synchrony output, because the two waveforms are subtracted sample by sample to produce the difference measure. Thus peaks are introduced at formant frequencies, as shown in the Figure.

The GSD algorithm produces prominent peaks at the frequencies of the first two formants, and the inherent amplitude normalization results in a balance of the amplitudes of F_1 and F_2 [compare the two bottom plots of cross sections in the off-glide]. The valley between the two formants in the /a/ is considerably deeper than in the case of method [7].

The horizontal bands that are evident in the spectrograms for methods [1] and [2] were unexpected. The effect is apparently being introduced at the level of the hyperbolic tangent half wave rectifier. To show that this is the case, two repeats of method [1] [which is just an integral of the energy in the outputs of the peripheral stage] are shown in Figure 11.10f, with simplified versions of the peripheral model. On the left, the energy is computed from the outputs of the peripheral filters, before either the AGC's or the half wave rectification have taken place. The resulting spectrogram and cross sections are very smooth, with no evidence of pitch ripple either in frequency or in time. On the right, the energy in the outputs of the AGC's is computed, thus sampling the output just before the half-wave rectification. It is clear that the AGC's have indeed smeared the formant peaks, as anticipated, but the jagged edges seen in method [1] are not yet introduced.

Although it is not reasonable to omit the half-wave rectifier for processing autocorrelations at the center period, it is feasible to replace the hyperbolic tangent with a simple piece-wise linear half-wave rectifier. The results for method [2] in this case are shown in Figure 11.10g. Interestingly,

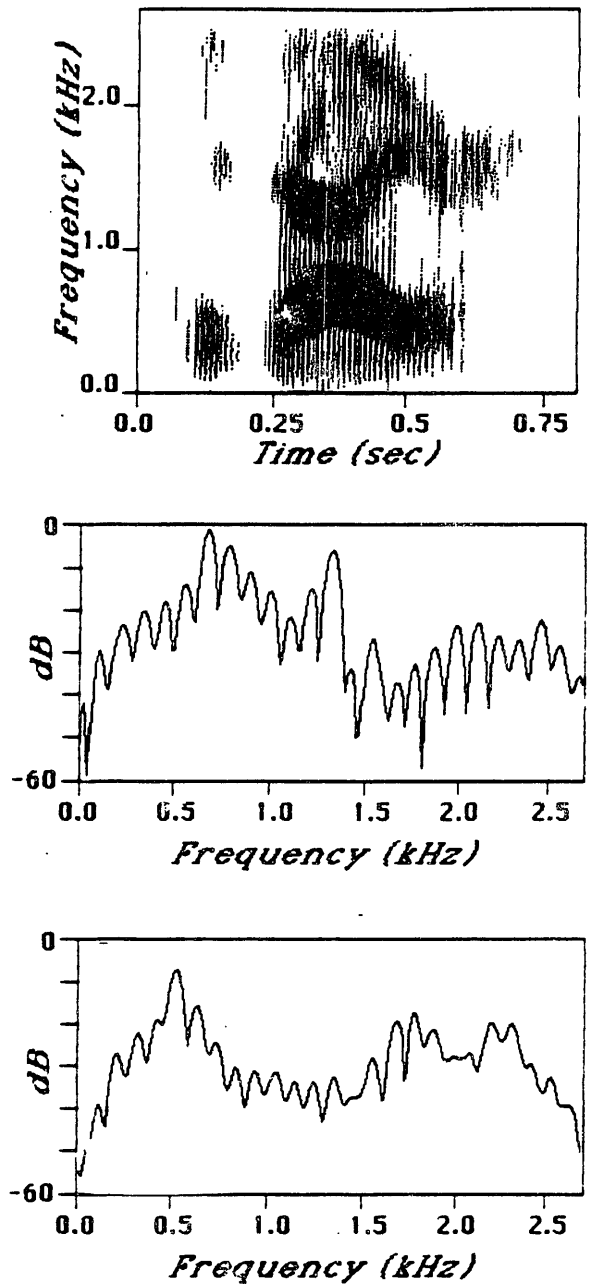

Figure 11.10: Illustration of processing of the word "desire", spoken by a male speaker, through a number of different synchrony measures, as discussed in the text. In all cases, the standard model is used for the peripheral system. Cross sections are shown at the two places indicated by the vertical bars in the spectrograms. Each part of the Figure, except the first part, illustrates two, usually related, versions of synchrony measurement.
a) Wide-band spectrogram [top], and narrow band spectra taken at time slices indicated by the vertical bars in the spectrogram, in the /*a*/ [middle] and /I/ [bottom].

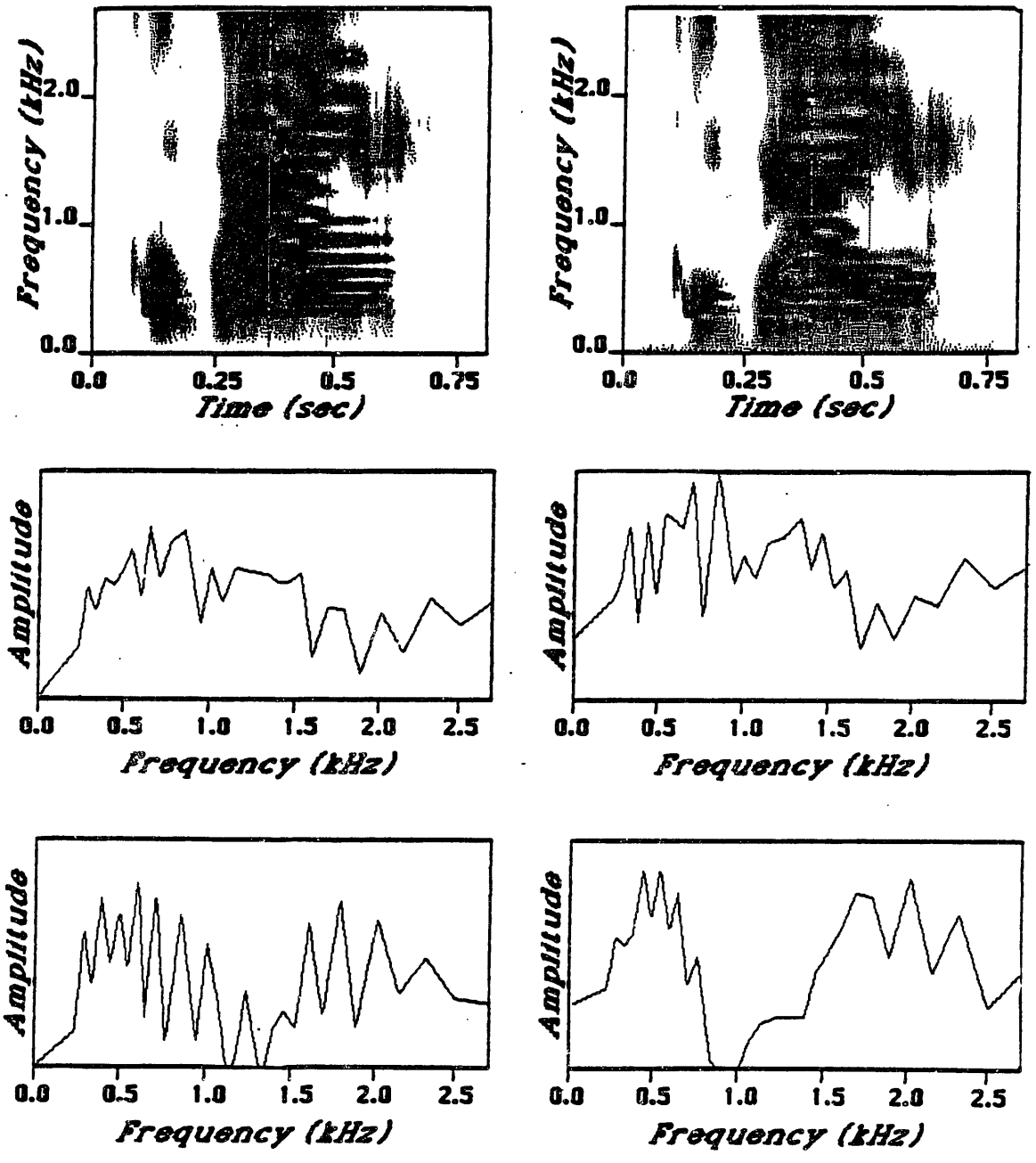

Figure 11.10:
b) Spectrograms and spectra of same data as in (a), obtained from the log energy in peripheral model outputs [left], and from the log of the autocorrelation coefficient [non-normalized] at center period [right].

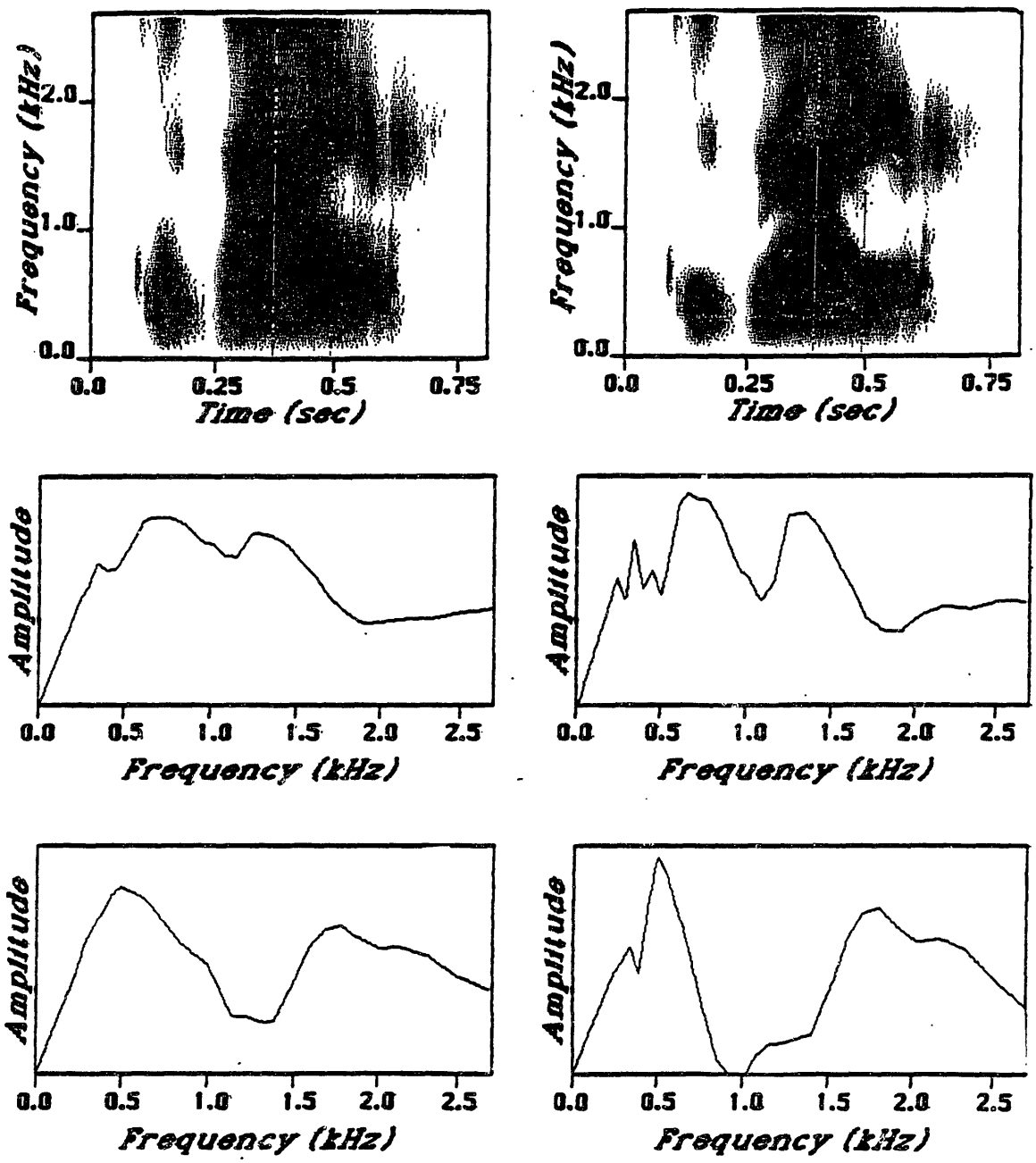

Figure 11.10:
c) Same as in (b), except the square root of each sample output is computed prior to integration over time. Thus a log magnitude spectrogram is obtained on the left, and a log square root autocorrelation spectrogram on the right.

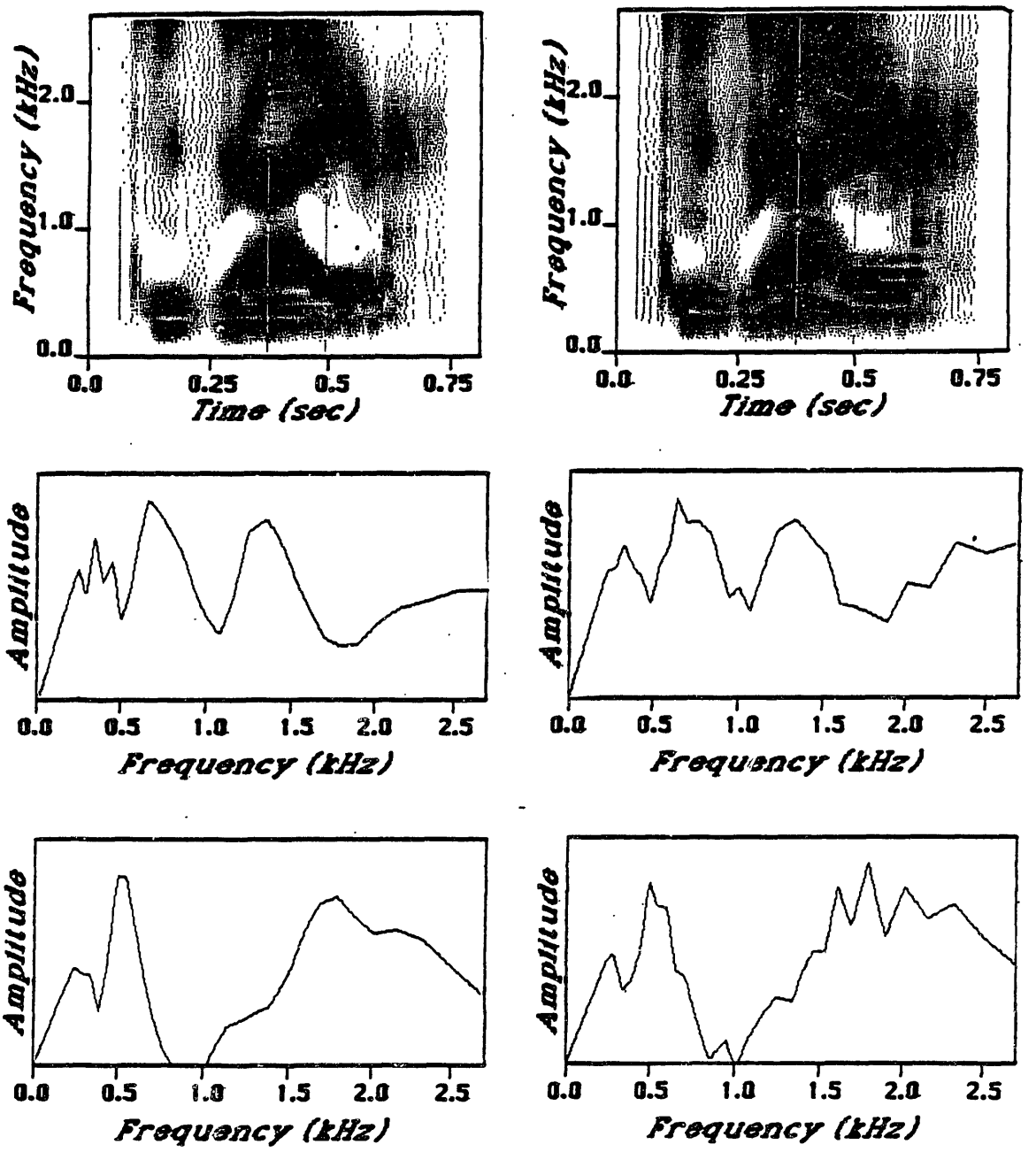

Figure 11.10:
d) Spectrograms and spectra obtained from normalized square root autocorrelation coefficient [left], and square root normalized autocorrelation coefficient [right], where normalization is with respect to average magnitude [left], and R_0 [right]. [See text for precise definitions].

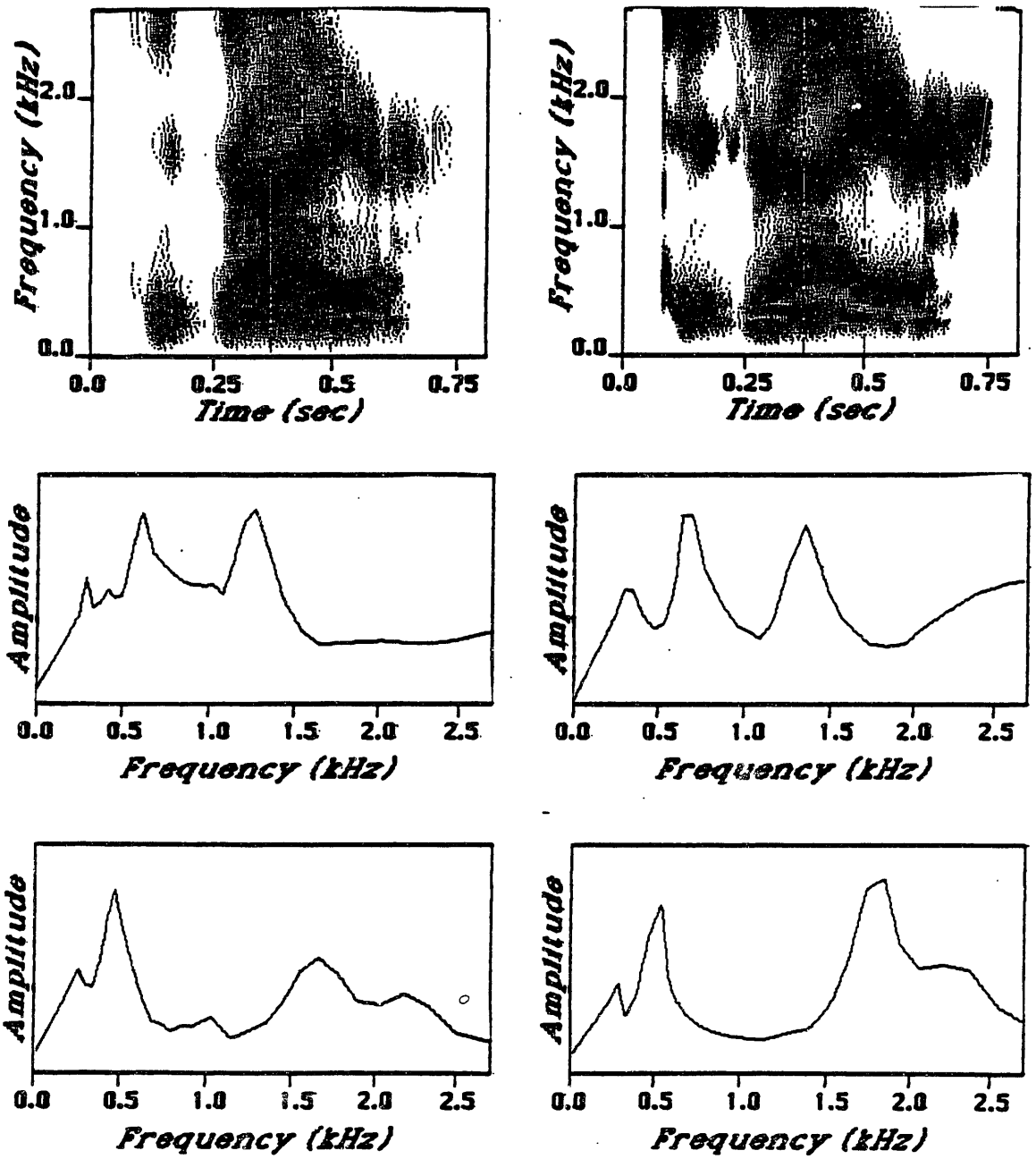

Figure 11.10:
e) Spectrograms and spectra obtained by comparing adjacent filter outputs [left] and by standard synchrony measure [right]. Adjacent filters are spaced by .4 bark, and comparison is made by measuring the magnitude of a difference waveform obtained by subtracting the adjacent filter output from the filter output.

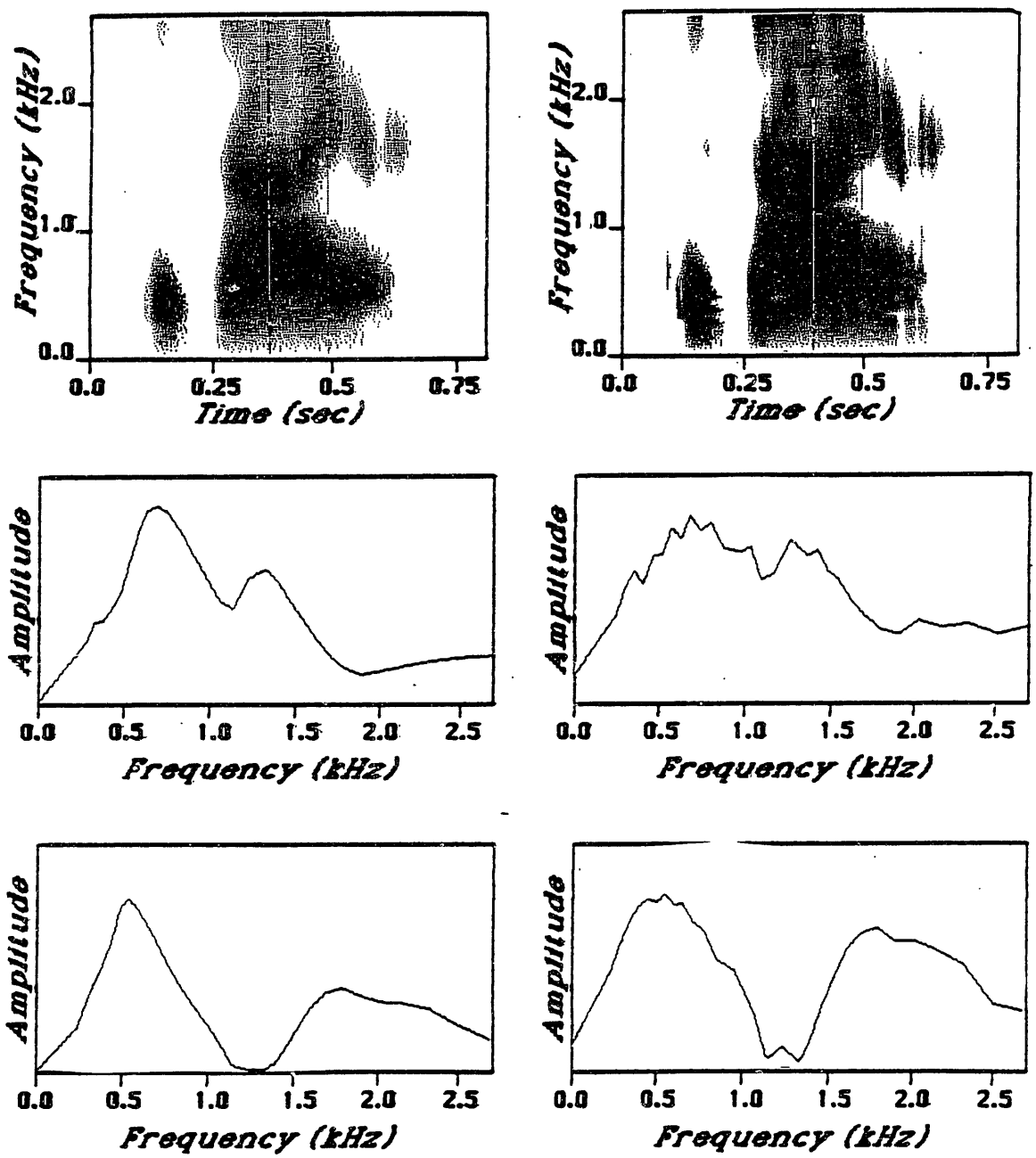

Figure 11.10:
f) Spectrograms and spectra obtained from the log energy in the outputs at intermediate stages of the peripheral model. The data is tapped just after the peripheral filters on the left, and just after the AGC's on the right. This data shows that irregularities in methods based on squared terms are introduced mainly by the half-wave rectification step.

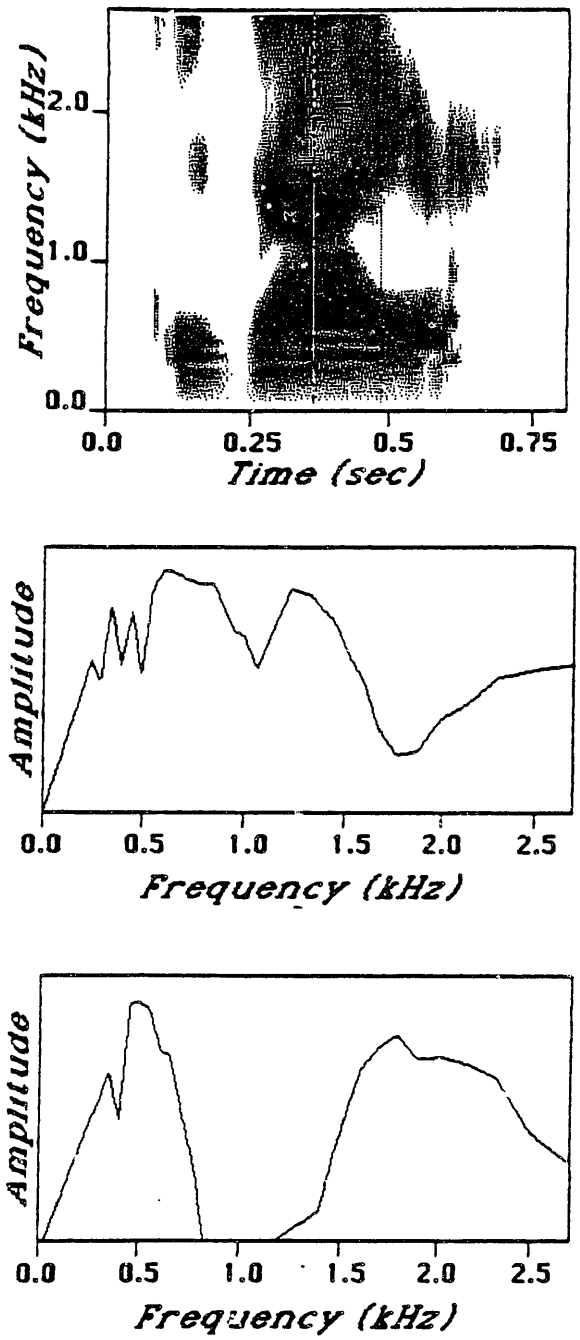

Figure 11.10:
g) Spectrograms and spectra obtained from log autocorrelation coefficient, when hyperbolic tangent half-wave is replaced with a piecewise-linear half-wave rectifier.

there is very little of the jaggedness that is apparent in the autocorrelation method when the hyperbolic tangent is used for half-wave rectification. The spectral peaks have been somewhat sharpened relative to the log energy, which is equivalent to a mean rate response.

The hyperbolic tangent half-wave rectifier has a much greater effect on the energy domain than on the magnitude domain. This is probably because the half-wave rectifier has the effect of enhancing the contrast between peaks and valleys in the waveform. The greater this contrast, the greater the difference between integrating magnitudes and integrating magnitudes squared.

The most promising of the alternate methods described here is the one that compares adjacent filter outputs. It is surprising that this method is able to preserve the salient features of the spectrogram, since it is so different from a standard spectral analysis procedure. There is no obvious way to incorporate energy normalization into this method, however. Nonetheless, the method should be pursued further in the future.

11.5 Conclusions

In this chapter we have attempted to illustrate certain properties of the specific system of the thesis as well as of several other methods for detecting synchrony. One major result is the observation that the GSD algorithm is relatively insensitive to major distortions applied to the input waveform. As long as essentially the same distortions are applied to each period of the waveform, the periodicity will be maintained, and thus the denominator will converge on zero. The method is more sensitive however to the frequency characteristics of the peripheral filter. In addition, the delay, τ_c, in the GSD algorithm, must be accurate, at least for the high frequency filters.

It is interesting that the DC component can be filtered out of the final output of the peripheral stage without a major effect on the overall system. This result is a little surprising, since the absence of DC results in a very different interpretation for the numerator of the GSD algorithm. The major problem is a loss of control over the value to use for the constant to subtract from the numerator to eliminate large responses to weak signals. Due to the possibility of very large negative outputs when the input is weak, it is necessary to process the GSD output through another half-wave rectifier. Such a step may not be unreasonable, given the general characteristics of nerve fibers.

Several other synchrony methods were investigated in the final experiment described in this chapter. Methods that included some form of amplitude normalization, such as R_i/R_0, in general produced smoother, less noisy, spectrograms than other methods. The autocorrelation methods were surprisingly sensitive to distortions on the waveshape of the input, imposed, for example, by the hyperbolic tangent half-wave rectifier. We consider such sensitivity to reflect a certain lack of robustness that is an unattractive property in a synchrony measure. The method that involved a comparison between adjacent filters represented a major departure from the basic premise of comparing a single channel output with a delayed version of itself. It was surprising that this method worked as well as it did, especially considering the differing phase characteristics of the adjacent filters, which should have caused a complex interaction.

Chapter 12

Pitch Detection: System Details and Examples

12.1 Introduction

In Chapter 8, a general description of the pitch estimation procedure was provided, but details of the process to actually extract an estimate of the fundamental frequency of voicing were omitted. This chapter is concerned with a detailed description of the pitch extraction process, including heuristics to decide which peak in the pseudo autocorrelation function represents the fundamental periodicity, as well as a voicing decision [a given frame is considered unvoiced if no prominent periodicities are present].

Pitch detection remains an unsolved engineering problem. Part of the difficulty is in the simplistic model that is typically assumed: either a given segment of speech is voiced, in which case a value of the fundamental period is given, or the frame is unvoiced, in which case the excitation is assumed to be white noise. In fact, a mixed source function is not uncommon in certain speech sounds: both a periodic glottal source and a noise source component at a constriction in the vocal tract are present simultaneously. Good examples of such a situation are voiced fricatives such as /v/ and /z/. Another common situation is vocal fry, which tends to occur frequently during vowels at the end of an utterance. In this case the source is at the glottis, but the glottal pulses occur with an apparently random spacing, or at least are highly irregular. A periodicity at a fundamental frequency is not evident; nor is a white noise source an appropriate model. It is inevitable that a pitch detector which forces a binary decision will "fail" in such circumstances.

In spite of the above objections, it was felt appropriate to attempt such a binary decision, derived from an examination of the outputs of the synchrony model. The exercise is a convenient mechanism for deducing how well periodicities, when present in the original signal, are preserved by the processing. We will begin with a discussion of the generation of the pitch waveform, which has been described briefly in Chapters 7 and 8. We will compare the pitch waveform with the original waveform by examining the detailed waveshapes, as well as spectra and pseudo autocorrelations applied to both signals. The next section will describe in detail the pitch detection process. Included are a description of heuristics used to select the peak in the pseudo autocorrelation, the strategy used to make a voiced/unvoiced decision, and the details of a post-processing stage to smooth the pitch track. The final section shows some examples of pitch detection applied to a set of 19 short words, such as dog, spoken by children under three years of age, which are an especially difficult data set.

12.2 Generation of Pitch Waveform

As discussed previously, the pitch waveform is generated by summing the outputs of the individual filters across the frequency or place dimension. The resulting waveform usually resembles the original waveform, although often, particularly for high pitched voices, the periodicities at the formant frequencies have been reduced; i.e., a form of spectral flattening has been achieved. The filters that are summed are spaced by half a Bark, and thus a significant frequency overlap exists between the individual components of the sum. The half-Bark spacing was determined somewhat empirically. It was found that full Bark spacing did not qualitatively produce as good results.

It is not clear how important the phase characteristics of the filters are to the overall performance of the summing process, although it is obvious that phase relationships among the individual filter outputs play a major role in the overall shape of the pitch waveform. The phase characteristic of each filter includes a strong linear phase component on which is superimposed a rapid 2π shift at resonance. Zeros on the x-axis also introduce a nonlinear phase component, which is more pronounced for the low frequency filters than for the higher frequency ones [see Figure 8.2b]. Each filter output was delayed appropriately before summing, so as to remove the linear phase component, but aside from this partial alignment of the components, phase characteristics should interfere in an unpredictable fashion.

The half-wave rectification process introduces a series of higher harmonics superimposed on the fundamental in each independent channel output. There is no obvious mechanism for controlling the phase characteristics of these higher harmonics relative to those of the outputs of filters tuned to the higher-harmonic frequencies. Thus the pitch waveform represents a complex sum of fundamentals and higher harmonics that may reinforce or cancel other fundamentals and/or higher harmonics present in distal channels.

An analytic treatment of the pitch waveform is thus impossible to obtain, but it is feasible to compare a narrow-band spectral analysis of the pitch waveform with the same analysis applied to the source waveform. Part (a) of Figure 12.1 shows a narrow band spectral cross section for the original waveform, lowpass filtered to 2667 Hz, and for the pitch waveform, during the /e/ of "major", spoken by a female speaker. Any information in the spectrum of the pitch waveform above 2700 Hz was introduced by the distortion processing in the peripheral model. The harmonics between the first and second formants are much higher in amplitude in the case of the pitch waveform than in the original waveform. In part (b) of the Figure, three other examples are shown, where frequencies above 2700 Hz are omitted, for the /w/ in "worship", the /ɛ/ in "yellow", and the /ɔ/ in "cost". In each case, the pitch waveform has a flatter overall spectrum than the original waveform.

Figure 12.2 shows narrow-band spectrograms of the original waveform and of the pitch waveform for the word "museum", spoken by a female speaker. The amplitude in the region between the first and second formants has been enhanced in the pitch waveform, for example, during the /i/. There is also a strong component at twice F_2, due to nonlinearities in the peripheral model. In fact, it was generally the case that the pitch waveform's spectrum was flatter than the spectrum of the original waveform. To the extent that spectral flattening is useful for pitch detection, the pitch

waveform should be a better source for determining pitch than the original waveform.

The pseudo autocorrelation does not depend upon a flat spectral envelope in order to produce a prominent peak at the fundamental period. In fact, most of the time, the pseudo autocorrelation function applied to the original waveform is not substantially inferior to the pseudo autocorrelation obtained from the pitch waveform. Figure 12.3 shows pseudo autocorrelations computed from the original waveform [left] compared with pseudo autocorrelations computed from the pitch waveform, for the word "hotel", spoken by a male speaker. The sequence is obtained at 80 ms increments during the /ɛl/. Although in all cases the peaks in the pseudo autocorrelation not at multiples of the pitch period are reduced in amplitude for the pitch waveform relative to the original waveform, in no case would an error be made if the largest peak in the pseudo autocorrelation of the original waveform were selected as the pitch period.

However, in the case of the 19 tokens of young children's speech, a major improvement was often obtained for the pitch waveform as contrasted with the original waveform. Figure 12.4 shows a plot as in Figure 12.3, where processing was done at 20 ms increments in time throughout the vowel, during the word "cat", the first word of the series. The original waveform contains a substantial amount of information at the formant frequencies, which shows up as additional peaks that clutter up the pseudo autocorrelation function. The pitch period is in fact not evident in the sixth token, for the original waveform, but is prominent in the pitch waveform's pseudo autocorrelation. The loss of information at the formant frequencies is in part a consequence of the spectral flattening process that takes place at the level of the peripheral processing.

12.3 Pitch Period Estimation

12.3.1 Heuristics to Estimate Pitch Period from Pseudo Autocorrelation

Usually, the correct pitch period is the time delay associated with the largest peak in the pseudo autocorrelation function. However, in the case of a perfectly periodic signal with a high fundamental frequency, there will be peaks of equal heights at multiples of the fundamental period. Given the minor variations that cause deviations from perfect periodicity, it is not uncommon for a peak at a higher harmonic of the pitch to be the largest peak in the pseudo autocorrelation. Hence, the major task of the heuristics is to look for a peak at a submultiple of the period of the largest peak, whose amplitude is close to the amplitude of the largest peak, and to accept such a peak as representing the true pitch period.

A relatively simple strategy that was found to work quite well most of the time is the following:

1. Find the absolute peak in the pseudo autocorrelation function, and mark its period, τ_{pk} as the candidate pitch period.

2. If the period of the candidate pitch is close to the period of the final pitch in the previous frame, accept it as the current pitch period.

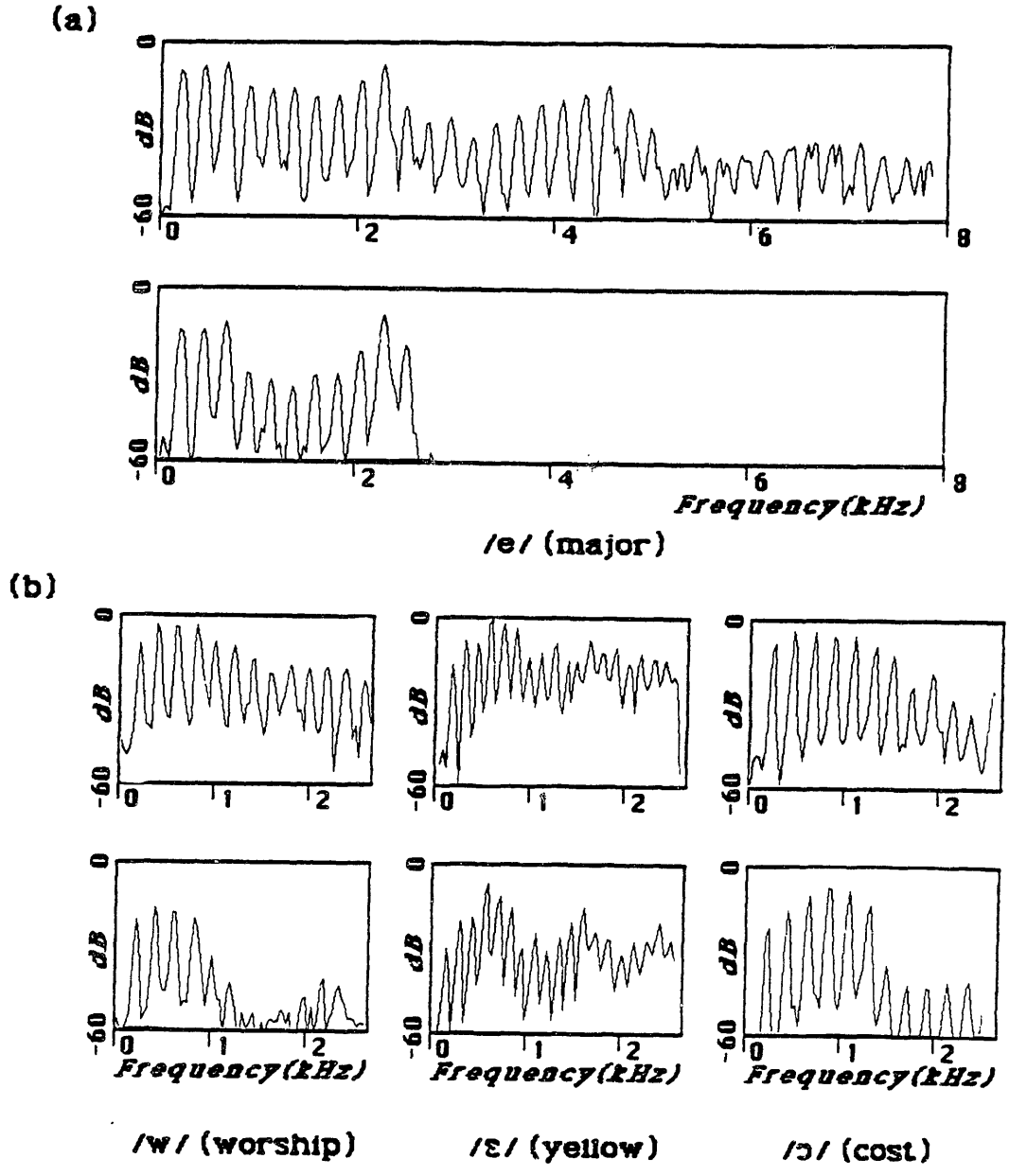

Figure 12.1: Illustration that pitch waveform is spectrally flattened relative to original waveform. a) Narrow-band spectra from 0 to 8kHz of pitch waveform [top] and of original waveform, lowpass filtered to 2.67 kHz, during /e/ of "major", spoken by a female speaker. The harmonics above 2.67 kHz in the pitch waveform were generated as a consequence of nonlinearities in the peripheral model. b) Three additional examples, showing only spectrum below 2.7kHz, the /w/ of "worship", spoken by a female speaker, the /ɛ/ of "yellow", spoken by a male speaker, and the /ɔ/ of "cost", spoken by a female speaker. In each case, the pitch waveform spectrum is at the top.

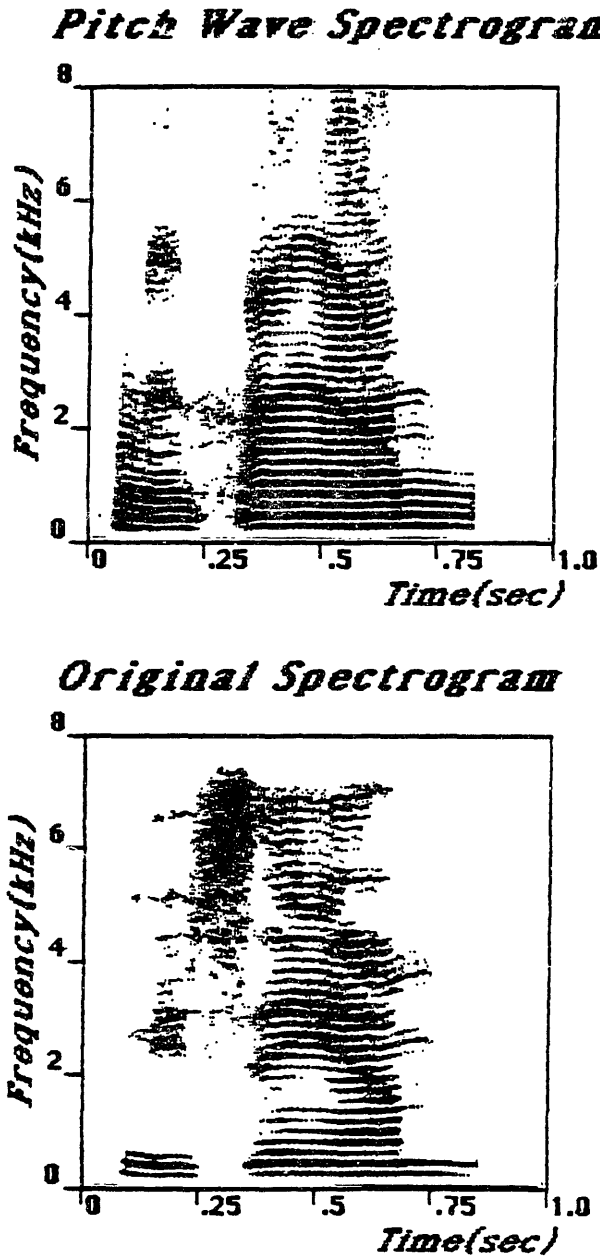

Figure 12.2: Comparison of narrow-band spectrograms [25 ms Hamming window] for pitch waveform versus original waveform for the word "museum", spoken by a female speaker.

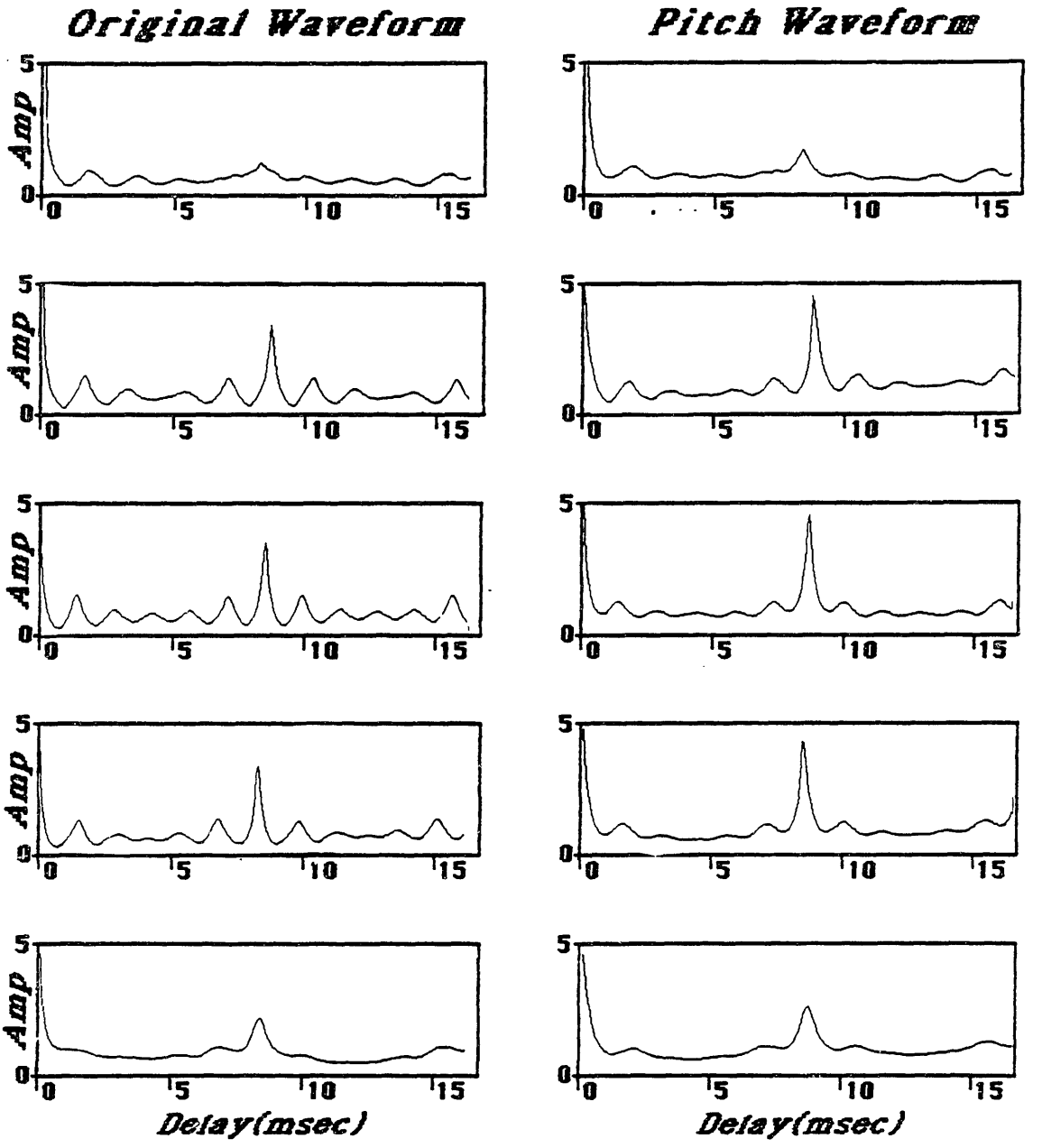

Figure 12.3: Series of pseudo autocorrelations of original waveform [left] versus pitch waveform [right] at 40 ms increments [time increasing from top to bottom] during the /ɛl/ of "hotel", spoken by a male speaker. Periodicities at formant frequencies are reduced somewhat in the pitch waveform relative to the original waveform, but the differences are not dramatic.

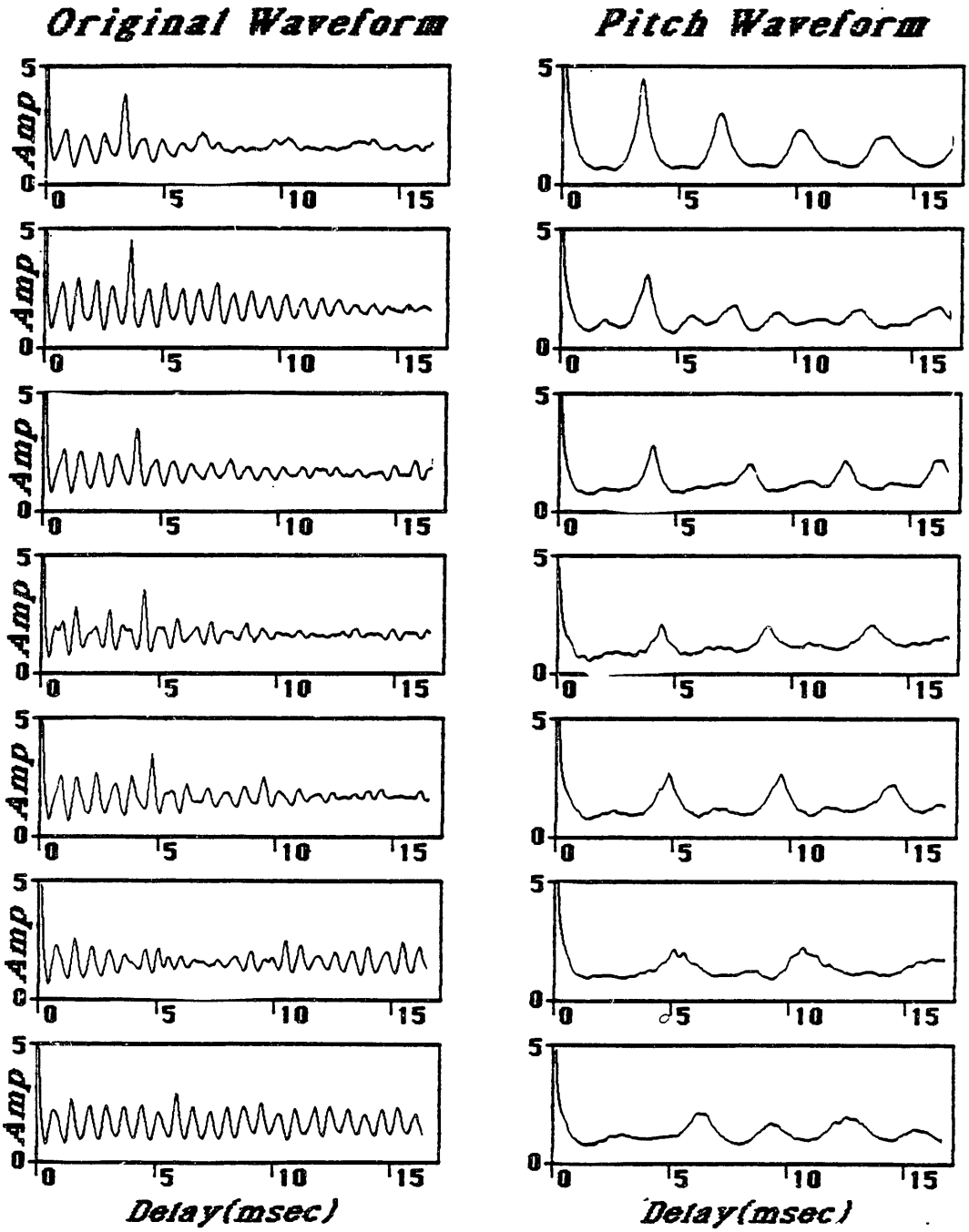

Figure 12.4: Series of pseudo autocorrelations of original waveform [left] versus pitch waveform [right] at 20 ms increments [time increasing from top to bottom], during the vowel portion of the word "cat", spoken by a preschooler. The periodicities at the formant frequencies have been substantially reduced in the pitch waveform relative to the original waveform.

3. Otherwise, find all the peaks in the pseudo autocorrelation from τ = 1.5 ms to $\tau = \tau_{pk}$, and consider each one by comparing its associated period, τ_i with the period of the closest submultiple of τ_{pk}. If a submultiple of τ_{pk} can be found that is very close to τ_i, and the amplitude of the peak at τ_i is within a threshold of the amplitude of the largest peak, accept τ_i as the pitch period. Even if the period of the peak fails the submultiple test, it can be accepted as the correct peak under a more stringent amplitude test. The **shortest** period that meets one of the above criteria is the final pitch estimate.

The thresholds that were used for amplitudes for the two different conditions (submultiple, not submultiple) are .85 and .95 respectfully. The threshold for considering a peak as a submultiple is defined as follows:

$$|\tau_i - \tau_{sub}| < .25\text{ms} + .04\tau_i$$

where τ_{sub} is the delay of the submultiple that is closest to the candidate peak.

12.3.2 Voiced-Unvoiced Decision

A voiced/unvoiced decision is usually an integral component of a pitch detection algorithm. Voiced sounds are produced by vibrations of the vocal folds at the glottis, whereas a noise source is a pressure source at a constriction placed somewhere in the vocal tract. Although in some cases, as mentioned previously, a noise source and periodic source may be present simultaneously, it is usually adequate to assume that the source function is either periodic or random. The task of the pitch detector is then to decide whether there is sufficient periodicity in the waveform such that the source function could be represented by a periodic pulse train, or whether a random noise source would be more appropriate, for a given segment of speech. If the source is periodic, then it is almost certain that it is a consequence of vibrations of the vocal folds at the glottis. However, the reverse is not necessarily true. In conditions such as vocal fry or a glottal stop, the source function is in fact a sequence of pulses, and therefore "voiced", but the pulses are not spaced by equal intervals, a necessary condition for periodicity. Thus a nonperiodic voiced source is also possible. In this case, the source function can not be described by a number representing the period, but rather the times of occurrence of each isolated glottal pulse would need to be specified. An example of such a situation is given in this section. Because the assumed model is too simple, the pitch detector can not hope to get the "right" result in such circumstances.

To complete the pitch estimation process in this system, a voiced/unvoiced decision was made, based only on information derived from the pitch waveform. Although in practice multiple cues are available from the original speech waveform for making such a decision, it was decided to restrict the system to work only from the pitch waveform, in conjunction with its derivatives, such as the pseudo autocorrelation function, in order to see how well the task could be accomplished from such a restricted source.

Several factors enter into the voiced/unvoiced decision. One factor is a simple amplitude measure, applied to the pitch waveform, after the DC component has been removed. A threshold on amplitude throws out silence and sounds whose energy is restricted to frequencies above 2700 Hz. Another important parameter is a measure of the prominence of the peak in the pseudo autocorrelation function at the period corresponding to the pitch estimate. This measure, "peakiness", is defined as the difference between the amplitude of the pseudo autocorrelation at the peak and the average amplitude over a region 2 ms wide, centered on the peak. A final measure useful for the voiced/unvoiced decision is the overall range of values of the pseudo autocorrelation; i.e., the difference between the maximum level and the minimum level. If the segment of speech is voiced, it will tend to have prominent peaks between deep valleys, whereas if it is unvoiced, there will be no periods for which the denominator will be small enough to generate a large peak.

Although the voiced/unvoiced decision is based on a set of hard thresholds, we make no claim that such an approach is representative of what might be happening in the natural system. It is likely that a voiced/unvoiced "decision" is never really made, but instead that certain features will excite a voicing detector, and certain other features will inhibit it. It is not the intent of this thesis to attempt to model anything as complex as a voicing decision. Instead, the simple set of heuristics developed here are only intended to turn the pitch detector into a practical system that could be used by a researcher interested in studying pitch or using a pitch track in speech analysis/synthesis systems.

The thresholds on the three parameters, pitch waveform amplitude, peakiness, and range, are set differently for regions in which the pitch value is not continuous, than for regions where it is. That is, if the current pitch estimate differs from the previous pitch estimate by a sufficient amount, a more relaxed set of thresholds on the parameters is allowed to disqualify the region as voiced. Such a continuity constraint is justified because in general pitch changes slowly over time. If the pitch changes substantially within a short time interval, it is likely that the peak is not really representative of a sustained periodicity. Pitch is updated every 5 ms in the system, and thus in general would be expected to change little from frame to frame.

An example of a heuristic decision tree for voicing that was found to work quite well is given in Table 12.1. In the table, an **AND** of all of the conditions in any one row will rule out voicing for the given frame. Thus there is an independent threshold on each of the three parameters, amplitude, peakiness, and range, that will rule out voicing. In addition, if all three parameters are below the thresholds indicated in row 4, the frame is considered unvoiced. The next two rows give thresholds for peakiness and amplitude that will rule out voicing if the pitch is discontinuous at the frame [if the change in pitch period is less than 0.5 ms plus eight percent of the pitch period].

Special treatment was necessary on rare occasions for the detection of a false pitch estimate at the first formant frequency. Such a circumstance can occur at the onset of a high vowel, when the periodicity at the formant frequency, which requires less integration time than the periodicity at the fundamental, triggers a false peak in the response. A proper solution would be to detect the anomalous nature of the peak, and to set the frame to be unvoiced [because two glottal pulses have not yet accumulated in the integration window]. It was found that a simple form of artificial

Peakiness	Amplitude	Range	Pitch Period
< 0.3	----	----	----
----	< 0.5	----	----
----	----	< 1.0	----
< 1.0	< 2.0	< 1.5	----
----	< 2.0	----	Discontinuous
< 0.8	----	----	Discontinuous
< 0.8	< 2.0	----	$P < 2.0$ms

Table 12.1: Heuristics used to make voiced/unvoiced decision based on data derived exclusively from pitch waveform. An **AND** of all conditions in any given row determines an **UNVOICED** decision. Pitch track is discontinuous if $\Delta P < .5\text{ms} + .08P$

intelligence could eliminate most such occurrences. If the peak for the pitch estimate is below two milliseconds, the peakiness factor is measured on the peak in the pseudo autocorrelation function at **twice** the fundamental period, rather than the peak at the fundamental period. The logic is that, if it is in fact a pitch period, then there will be a prominent component at the double period, whereas if it is a formant, then it would be fortuitous for a strong component to exist at half the formant frequency. Because the overall energy tends to be weak in such anomalous circumstances, an additional pair of thresholds on overall energy and peakiness considered jointly is included to rule out voiced, in the event that the fundamental frequency is high [last row in the table].

An example where the above logic failed to work in determining that a periodicity was in fact at the formant frequency rather than the pitch frequency is shown in Figure 12.5. The example is the word "museum", spoken by a different female speaker from the example in Figure 12.2. Part (a) shows a comparison of the pitch waveform with the original waveform at the onset of the /i/. Part (b) shows the narrow-band spectrum for a 25 ms segment for both waveforms. The original waveform has a very strong component at the formant frequency, which is quite close to the second harmonic of the pitch frequency. The half-wave rectification in the peripheral model has introduced an additional component at twice the formant frequency. The pseudo autocorrelation of the pitch waveform is shown in part (c). In this case, the periodicity with the formant period is too strong to be ignored by a decision algorithm. Hence, a small segment of the pitch track [part (d)] is an erroneous track of the formant frequency, at vowel onset. Such an error could only be corrected by a second pass based on high-level knowledge of reasonable pitch tracks.

The most difficult problem to deal with is the glottalization that frequently occurs at the end of an utterance, particularly during an unstressed or reduced syllable. Often the spacing between

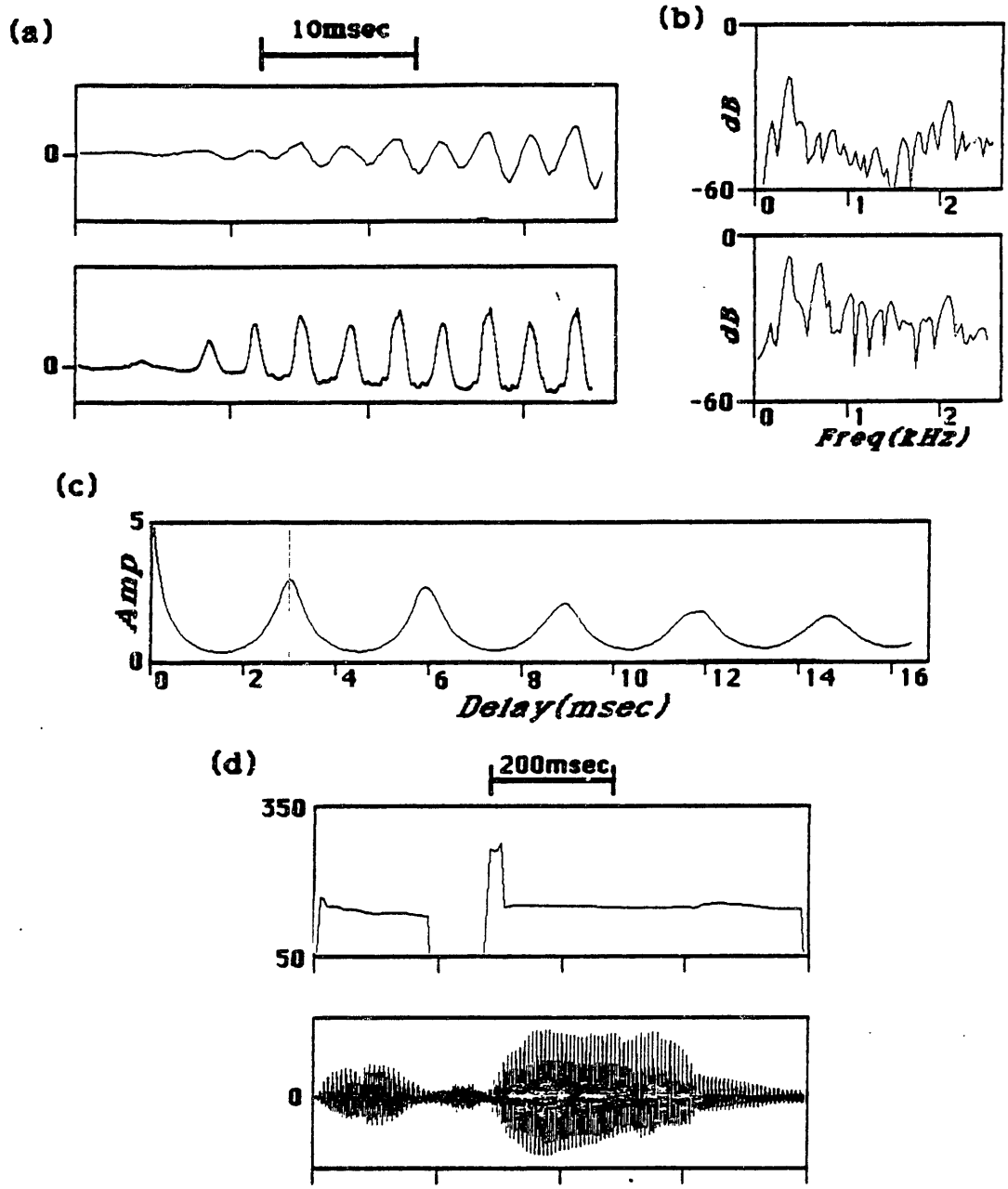

Figure 12.5: Example where the pitch detector made an error. The formant frequency was erroneously labelled as the pitch frequency at the beginning of the /i/ in "museum", spoken by a female speaker.

a) Comparison of original waveform [top] with pitch waveform at onset of /i/.

b) Comparison of narrow-band spectrum of original waveform [top] with that of pitch waveform.

c) Pseudo autocorrelation of pitch waveform at vowel onset, showing strong periodicity at formant frequency [300 Hz].

d) Pitch track aligned with original waveform on reduced time scale, showing the error at the onset of the /i/.

individual glottal pulses is highly irregular, making it impossible to define a periodicity surviving beyond two glottal pulses. Even if the algorithm detects the corect spacing at a given instance in time, the spacing will change substantially from frame to frame, thus causing a relaxation on the thresholds for the voiced/unvoiced decision.

An example of such a situation is given in Figure 12.6, for the last syllable of the word "hesitate", spoken by a male speaker. The original waveform, pitch waveform, and pitch track are shown time-aligned in part (a) of the Figure. Two conflicting periodicities that were detected by the pseudo autocorrelation towards the end of the vowel, are at about 9 and 14 ms, as shown in part (b) of the Figure. This segment of speech sounds rough, and the pitch track is correspondingly irregular, including a break into a short "unvoiced" region. As mentioned previously, neither a periodic model nor a noise model for the source function is sufficient in such circumstances.

12.3.3 Post-Hoc Editing

Most pitch detectors include a final stage of cleaning up the pitch track using some simple smoothing strategies such as median smoothing. The final editing that is used for the pitch detector described here includes a check for single frames out of line, followed by the elimination of any voiced regions less than three frames long [15 ms].

The algorithm to eliminate a single bad frame is as follows. Let P_{i-1}, P_i, and P_{i+1} be the periods in ms of the three sequential frames under consideration, and let $\theta = .8 + .1P_i$. If P_i deviates from both P_{i-1} and P_{i+1} by more than θ, and $|P_{i+1} - P_{i-1}|$ is less than θ, reset P_i to the average of P_{i-1} and P_{i+1}. If the two surrounding pitch values deviate by more than θ, set the middle value to unvoiced.

12.4 Study of Nineteen Words Spoken by Preschoolers

As noted previously [see Figure 12.4], the spectral flattening inherent in the procedure for generating the pitch waveform from the original waveform is particularly effective for voices with a very high pitch frequency. Consequently, this pitch detector is capable of extracting the fundamental period of the speech of very young children. To demonstrate this capability, we have chosen to illustrate pitch tracks derived from a series of 19 short words ["dog", "cat", etc.] spoken by preschoolers, which were provided by Dr. Philip Lieberman. The study is a test of the voiced/unvoiced decision as well, as long segments of burst and aspiration are apparent in many cases.

Results are shown in Figure 12.7. For many of these words, it is difficult at times to assess the pitch visually, and hence it was not expected that the pitch detector would always produce a continuous pitch track. For regions where there is an obvious gap in the pitch track, an insert of the corresponding original waveform is shown in the Figure. Such an insert was included for word3, word8, word18, and word19. Four glottal pulses are apparent in the insert for word3, but they are not equally spaced, and the waveshape changes substantially from pulse to pulse. Hence

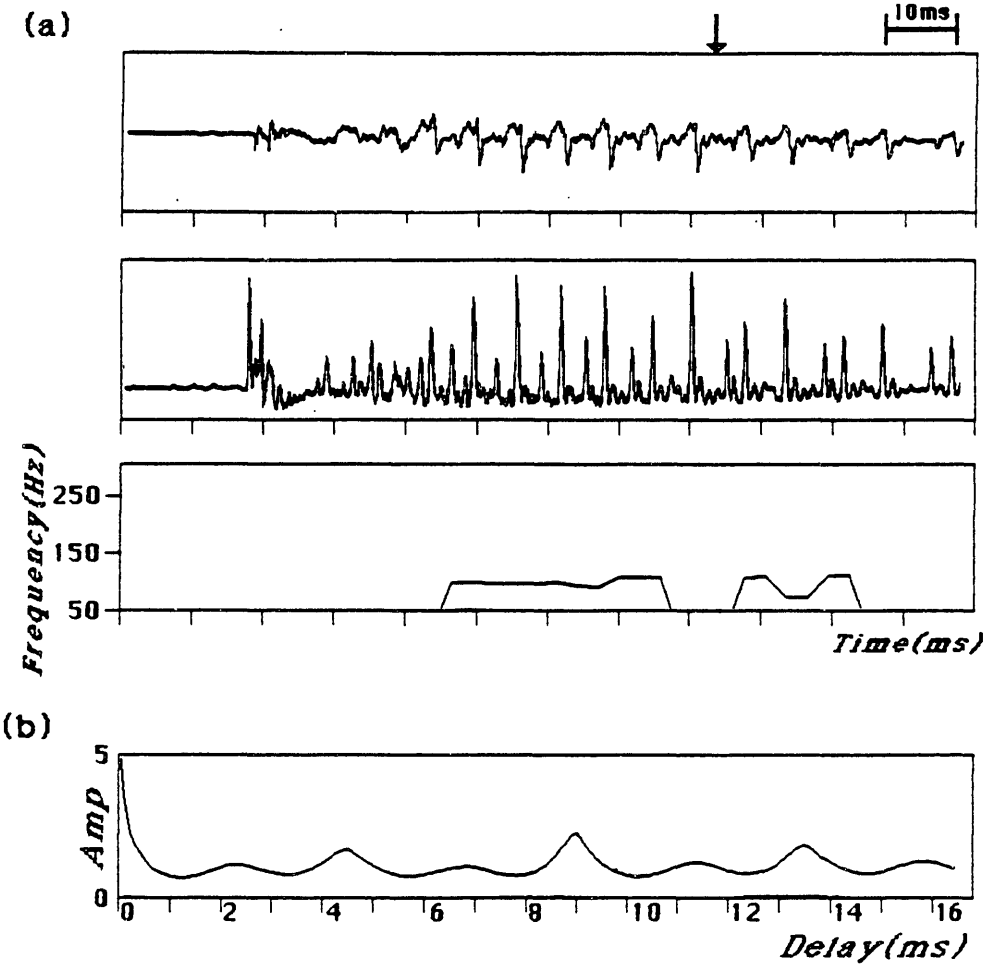

Figure 12.6: Example of glottalization at end of the word "hesitate", spoken by a male speaker.
a) Original waveform aligned with pitch waveform and pitch track, during /e/ of "hesitate". Pseudo autocorrelation is detecting multiple periodicities; as a consequence, voicing drops out in the pitch track during a portion of the vowel.
b) Pseudo autocorrelation of pitch waveform in the vicinity of arrow in part (a), showing ambiguous periodicities

the waveform was not sufficiently periodic for a VOICED decision. This was an inflection point in the pitch, showing rapid change.

The pitch track is quite broken up at the end of word8. Likewise, it is not clear at all from the waveform where the glottal pulses are occurring. At the onset of word18, a formant frequency was detected as a pitch frequency. In this case, as in the word "museum" examined above, a sophisticated knowledge-based approach could correct this error. A pseudo-periodicity at the fundamental is detectable by eye, and therefore this is an error on the part of the pitch detector. The gap near the end of word19, on the other hand, is probably legitimate, to the extent that the waveform is very irregular, and no periodicities can be detected visually.

The pitch detector is able to track the very rapid rising pitch of word10, word11, and word16. In some cases, such as the latter third of word14 and most of word17, the pitch track is somewhat rough. It is not easy to assess whether such irregularities are inherent in the waveform or introduced as a consequence of noise in the measurement process, since in these cases the pseudo-periodicities are not clear visually. Except for a few cases as discussed above, the pitch detector was able to extract the correct pitch and to make a correct voiced/unvoiced decision for these 19 words.

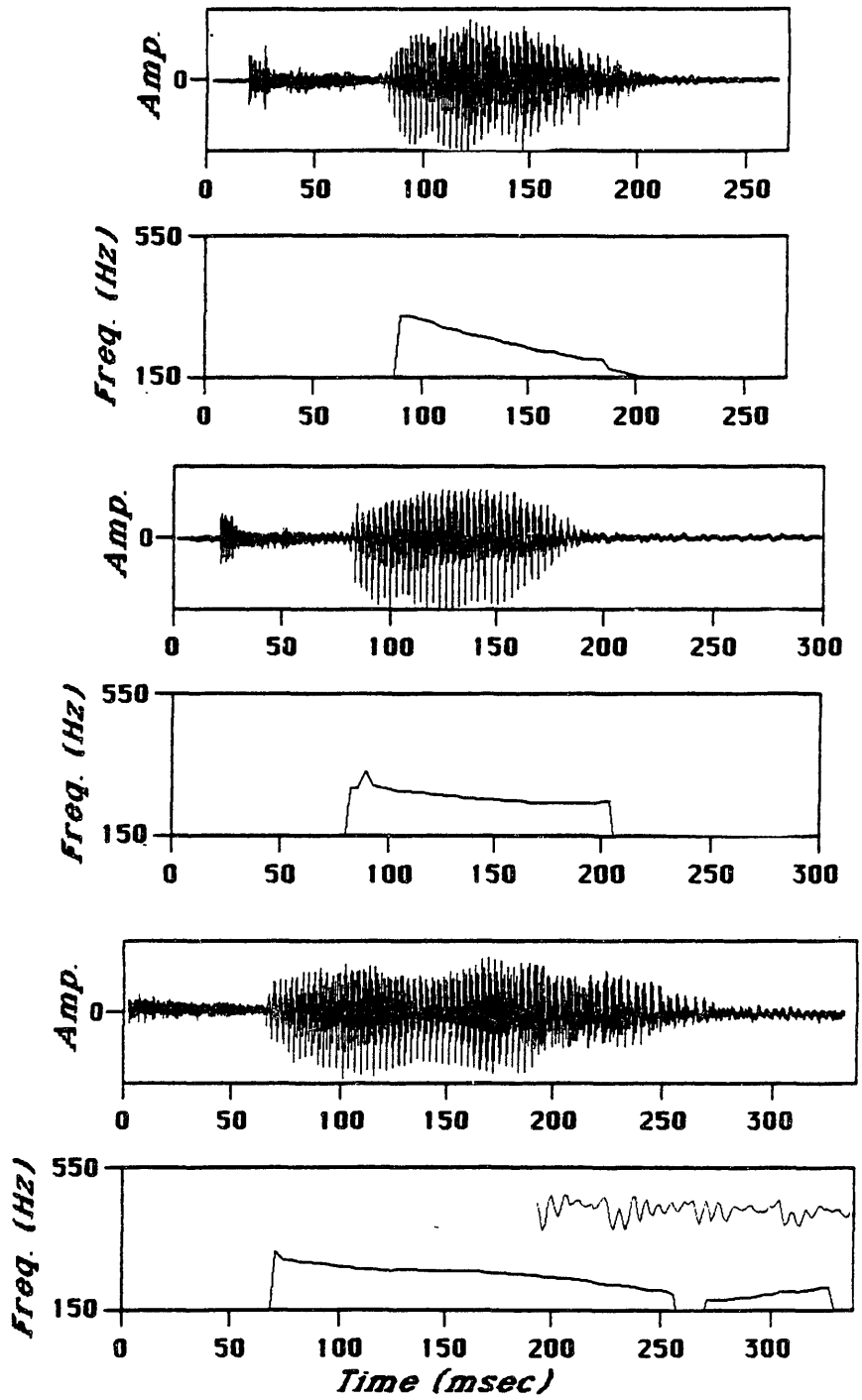

Figure 12.7: [Continued on subsequent pages] Pitch tracks aligned with original waveform for series of 19 short words spoken by preschoolers. In certain cases where the pitch track is discontinuous, an insert of the original waveform on an expanded time scale is included to provide additional information.

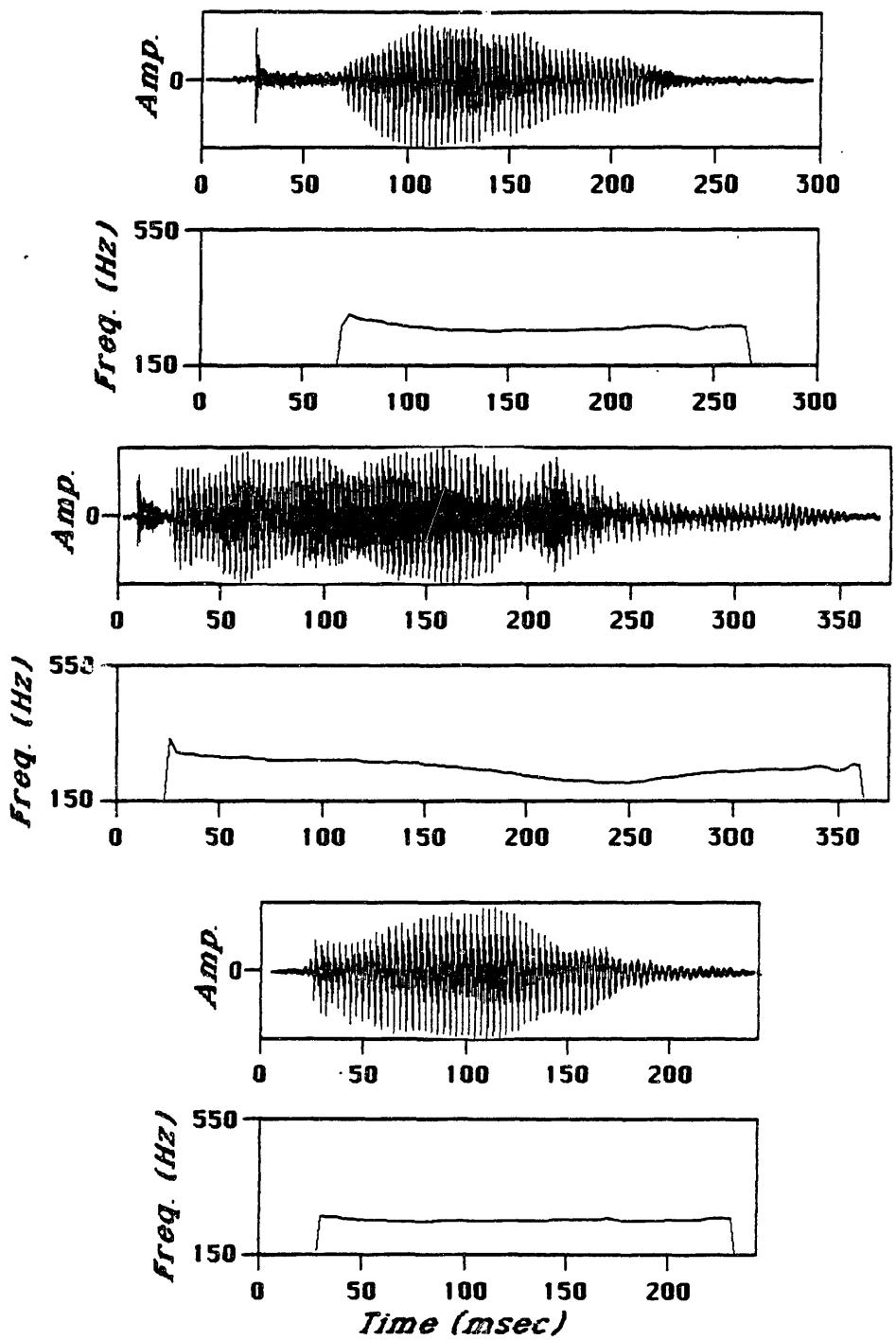

Figure 12.7: [Continued]

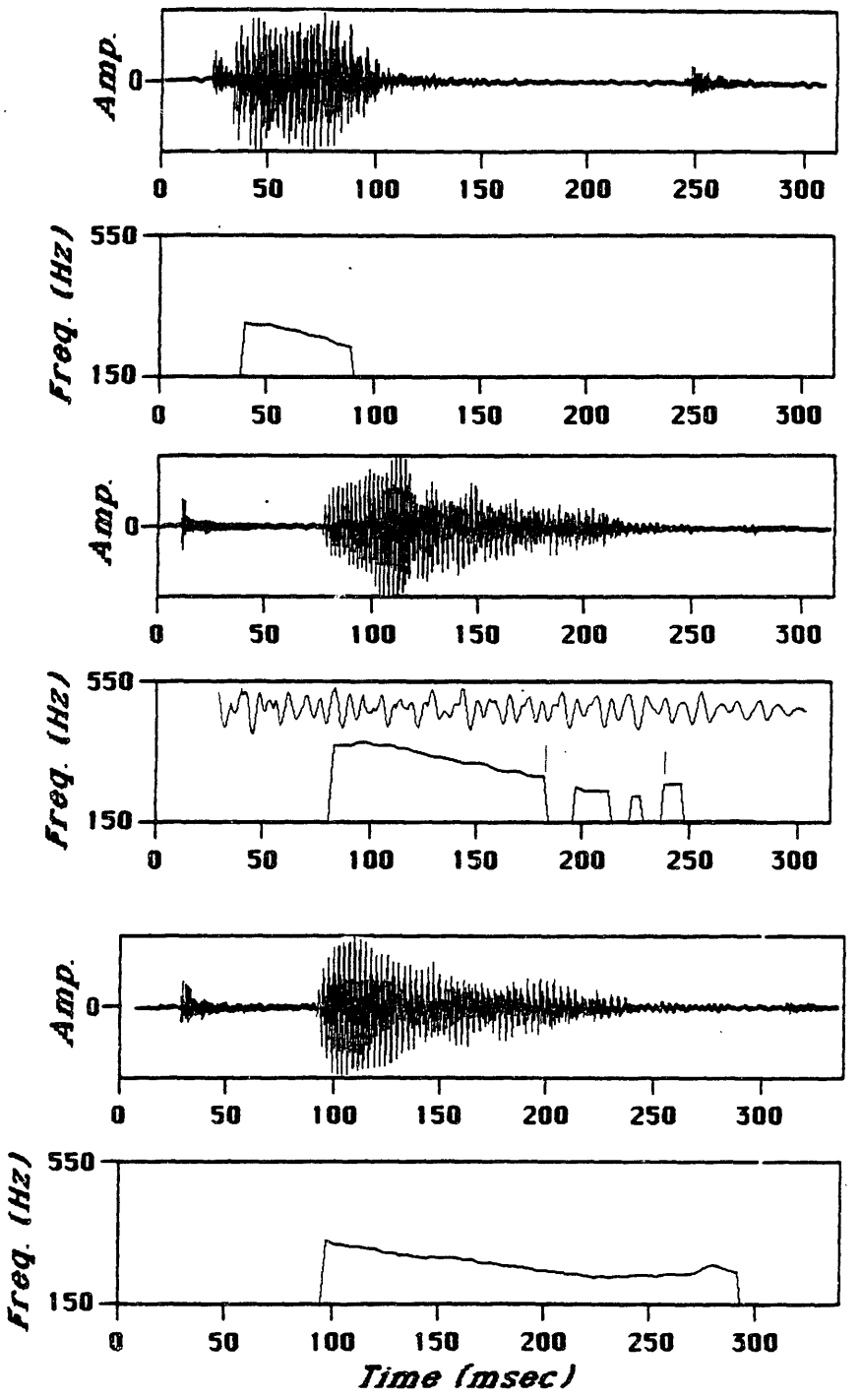

Figure 12.7: [Continued]

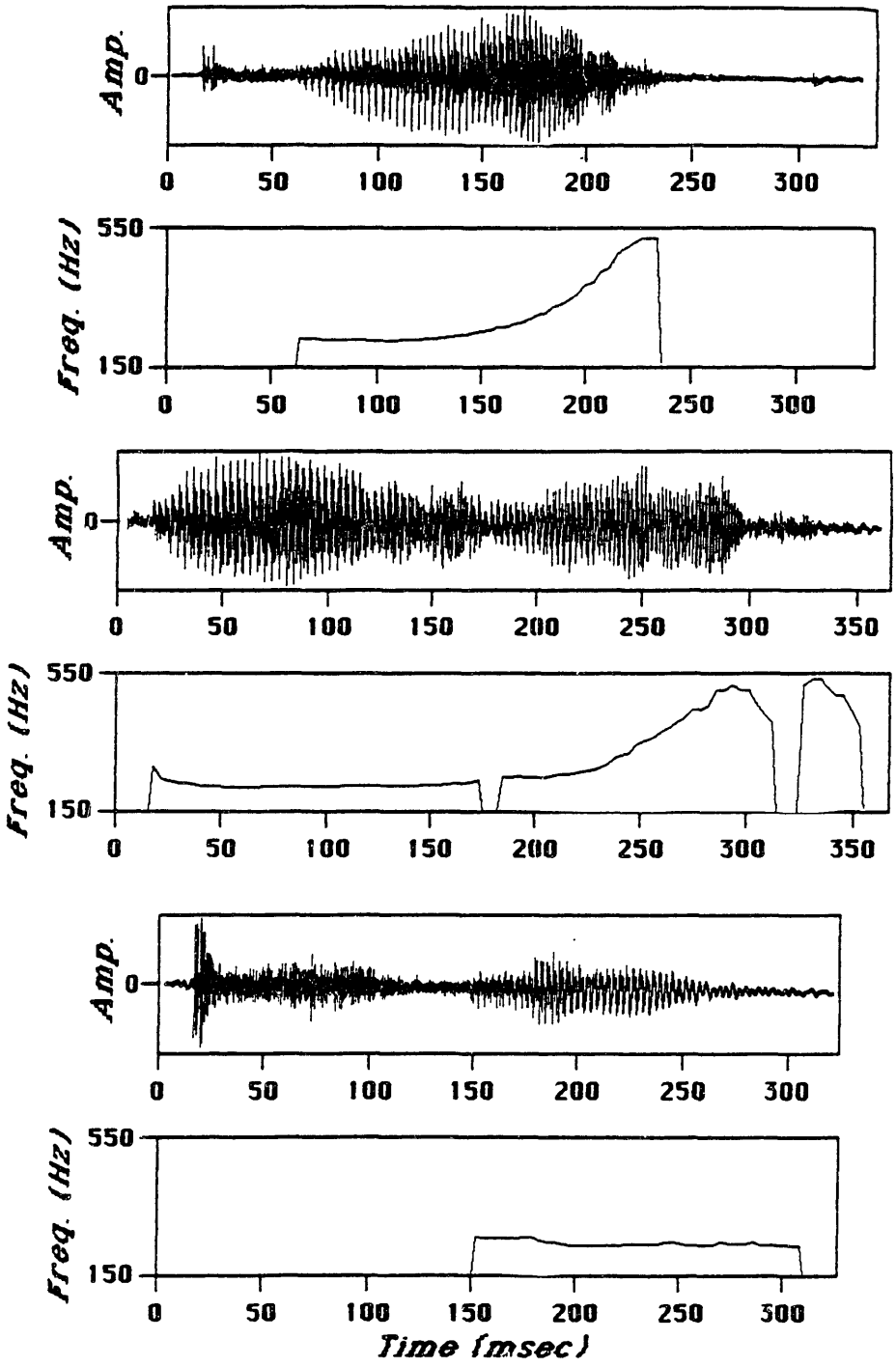

Figure 12.7: [Continued]

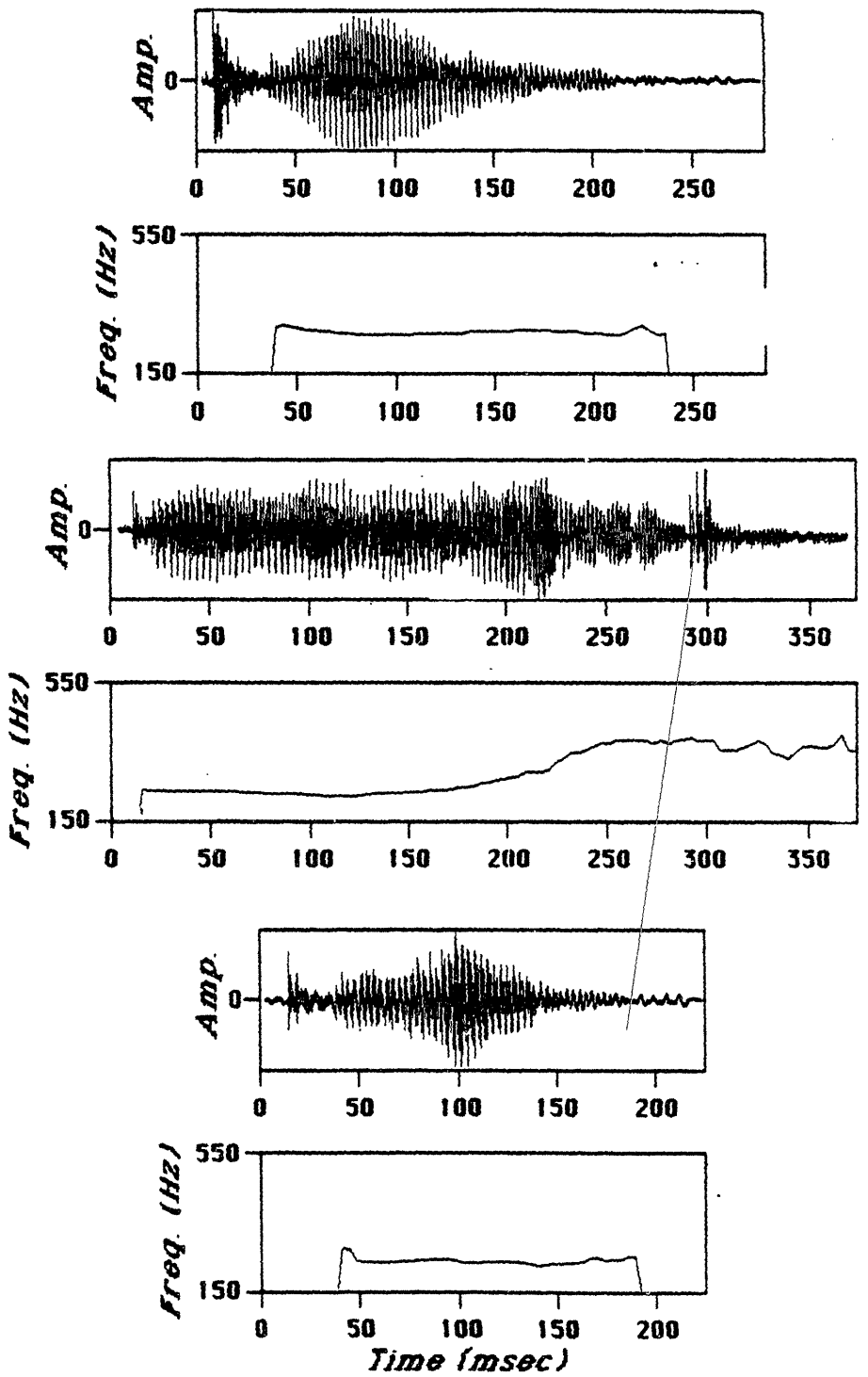

Figure 12.7: [Continued]

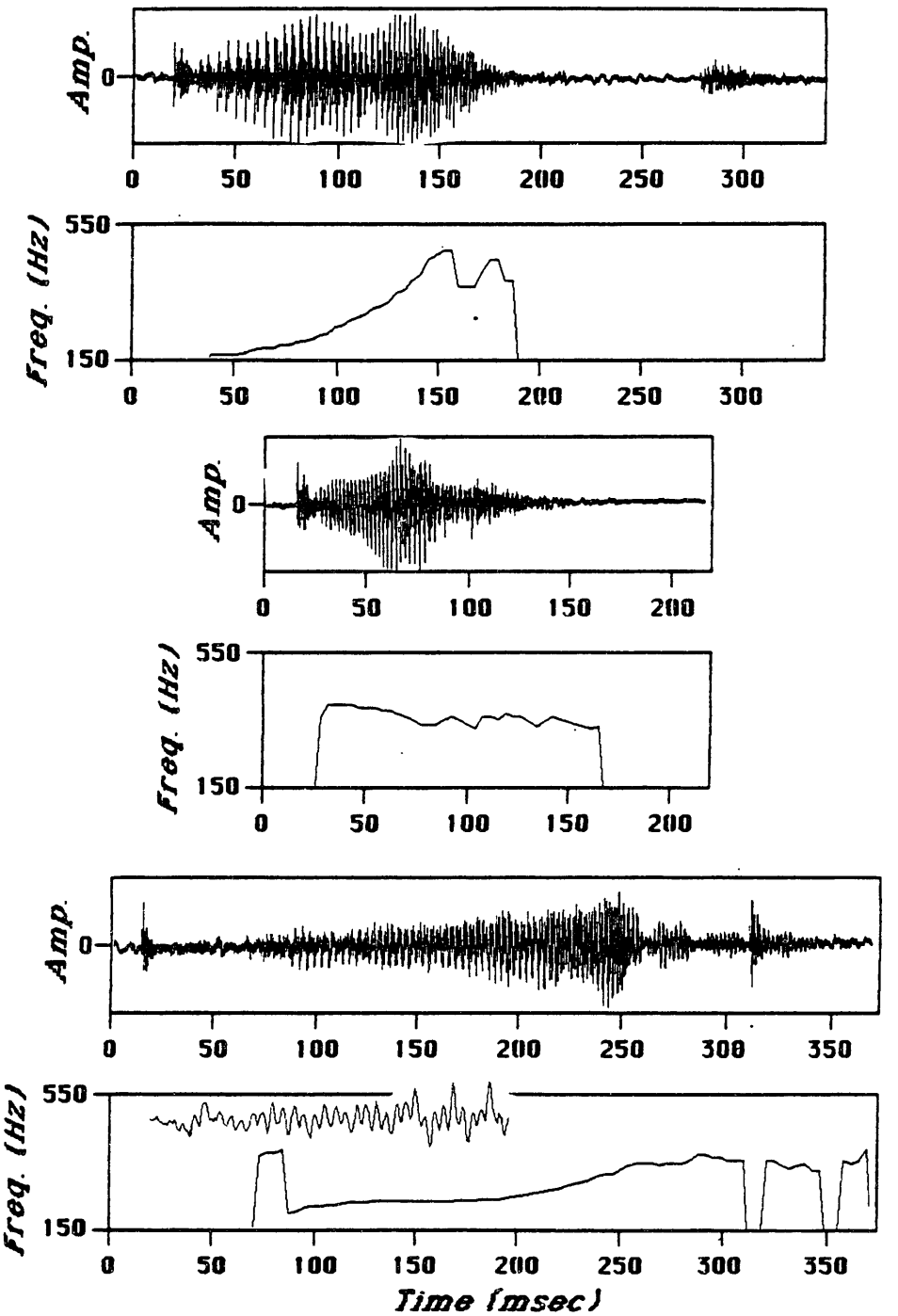

Figure 12.7: [Continued]

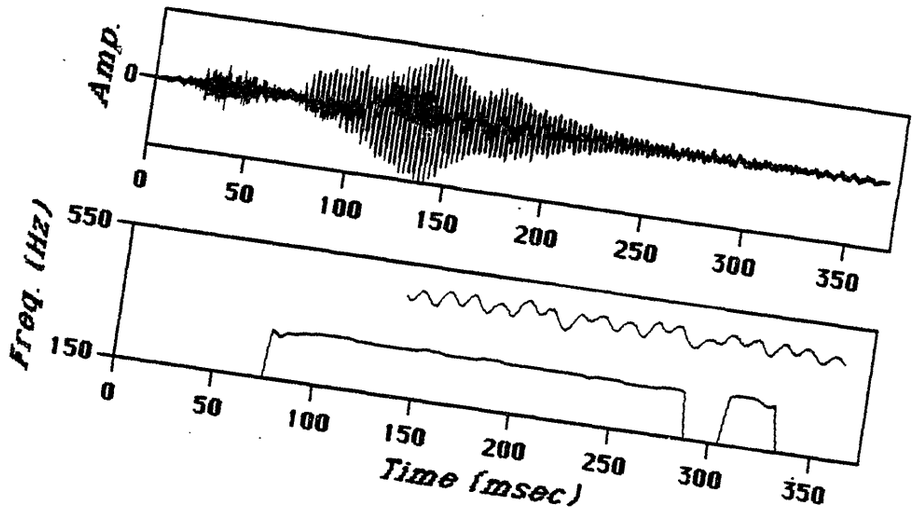

Figure 12.7: [Continued]

Chapter 13

Summary and Discussion

13.1 Summary

In this thesis, we have described a system for processing the speech waveform, which borrows from what is known about human auditory processing at the peripheral level, and then invents strategies for taking advantage of the synchronous information in the firing patterns. It obtains a representation for the speech spectrum in the low frequency region [from about 200 to about 2700 Hz], as well as an estimate of the fundamental period of voicing. A model for the peripheral level processing is developed, which is based on histogram data obtained from the spike patterns in the response of isolated nerve fibers to known signals. The model is represented as a waveform describing the probability of firing; i.e., the data are never reduced to a spike sequence.

The outputs of the peripheral level model are delivered to a series of "Generalized Synchrony Detectors" [GSD's], which further process the data so as to emphasize spectral prominences. The GSD algorithm is the most significant feature of the thesis system. It is computed as the ratio of an integrated sum waveform over an integrated difference waveform, where the sum and difference are computed with a specified delay period. It obtains a very strong response to signals that are periodic with the delay period. In the case of spectral analysis, the delay for each channel is set to be equal to the inverse of the center frequency of the corresponding peripheral level filter. The outputs of the GSD's can be plotted as a function of frequency to yield a "pseudo spectrum", which is similar to a standard spectrum in that the frequency locations of peaks and valleys are preserved. The peaks in the pseudo spectrum are in general considerably sharper than what would be obtained by simply integrating the response measured at the peripheral level, to produce a "mean rate spectrum".

We also hypothesize a method for determining the fundamental frequency of voicing of the speech signal, which begins with a combining of all the peripheral level outputs to produce a single waveform, the "pitch waveform", from which to extract fundamental periodicities. This waveform is then fanned out to a series of GSD's at periods covering the full human range of fundamental frequency. The result can be plotted as a "pseudo autocorrelation", which shows peaks at multiples of the fundamental period. A pitch extraction algorithm was also developed, mainly as a mechanism to aid in the evaluation of the quality of the pseudo autocorrelation for capturing relevant pitch information.

One main purpose of the thesis is to demonstrate the properties of the pseudo spectrum and the pseudo autocorrelation, by means of a set of examples of outputs for a variety of different natural and synthetic speech tokens. We have shown that the pseudo spectrum usually produces a single prominent peak at the second formant frequency, which may in certain cases, such as /ɝ/,

merge with a third formant peak. On the other hand, the situation in the first formant region [below about 800 Hz] is much more complicated. For low vowels, which have a high first formant frequency, there is usually at least one extra peak below the first formant, which may correspond to a glottal formant. In speech with a high fundamental frequency of voicing, the individual harmonic components are usually resolved below the first formant. However, the harmonics just above the formant are usually suppressed by the synchrony extraction process, because the low frequency tail on the peripheral filter picks up enough information at the formant frequency to suppress synchrony to the filter center frequency. When vowels are nasalized, the situation is complicated by additional poles and zeros in the F_1 region. These complexities usually show up in the pseudo spectrum as a reduction in amplitude and, in many cases, peak-splitting of the first formant peak.

Stop bursts are usually much more evident in the pseudo spectrogram than in standard wide-band spectrogram displays. The main reason is that the peripheral level model takes into account the high-level onset response characteristics of nerve fibers. The large increase in energy level at onset carries through to the synchrony analysis to produce, in many cases, a sharp line to mark the location in time of the burst. This feature should be particularly significant for segmentation algorithms, where it is important to measure the frequency characteristics of the burst at a precise point in time.

In addition to demonstrating the properties of the pseudo spectrum we also attempted to characterize the sensitivity of the GSD algorithm, as applied to spectral analysis, to diverse factors. We showed that the frequency characteristics of the peripheral filters are an important factor in shaping the pseudo spectrum. It is significant that peripheral filters have a sharp cutoff on the high frequency side. If a filter picks up information at twice its center frequency, such information will appear synchronous to its center frequency. The low frequency tail also cannot be too broad, or else the peak at the second formant will be significantly weakened by the presence of energy in the signal at the first formant frequency. We also showed that it is important for the delay in the GSD algorithm to be precise, at least for the high frequency filters [above about 1500 Hz]. This necessitated an upsampling of some of the high frequency filter signals to a 32,000 Hz sampling rate.

On the other hand, the GSD algorithm is remarkably insensitive to distortions on the waveshape, as long as distortions from cycle to cycle of the signal are consistent. This is an attractive feature for a synchrony-detection device, because it implies a robustness against distortions resulting from nonlinearities in the hair cell mechanism and/or in the transformation from voltage to spike sequence in the auditory nerve. Although distortions in the steady-state waveshape are unimportant, the sharp onset response characteristic of the peripheral level model plays an important role in enhancing the prominence of stop bursts. If the dynamic range compression and half-wave rectification are omitted, stop bursts are much weaker in the final pseudo spectrum, and edges are not nearly as sharp.

We compared the GSD algorithm with a number of different alternative mechanisms for detecting synchrony, most of which were based on autocorrelations of peripheral level outputs computed at the center period of the peripheral filter. We showed that the GSD algorithm produces in general

a smoother spectral representation, with prominent peaks at the formants. Autocorrelation tends to be more sensitive to distortions on the wave shape due, for example, to the nonlinear half-wave rectifier. One unique method is the "Merzenich" method that involves a comparison of adjacent channel outputs, rather than measuring synchrony to a specific center frequency in each channel independently. This method holds some promise and is an area for further research.

For pitch extraction, all of the peripheral level outputs are added together to produce the pitch waveform, which maintains the fundamental periodicities of the original waveform, but tends to show a reduction in the periodicities at the formant frequencies. This single waveform is then analyzed through a series of GSD's to produce the pseudo autocorrelation, from which an estimate of the pitch period is determined. We showed that this method can be used for determining the fundamental periodicity of a sine wave or a harmonic sequence in any region within the range of the system. We also showed that, at least in a general sense, the results are consistent with data on the perception of the pitch of inharmonic sequences.

It is a feature that the pitch extraction problem is reduced to an analysis of a single waveform for periodicities. It has been hypothesized that long delays can not be maintained accurately in a nerve network. If a pitch estimator had to combine periodicity estimates obtained from multiple waveform sources, there would be concern about inaccuracies in the individual delays, which would tend to blur the combined estimate. If there is only a single waveform to be analyzed, it could be processed through a tapped delay line, with a GSD detector stationed at each tap. Each GSD has no explicit knowledge of its absolute delay, but it is certain that the delay is longer than that of the preceding GSD and shorter than that of the following one. Thus a peak in the pseudo autocorrelation is anchored in a relative but not an absolute sense. This aspect seems to correspond to human pitch processing, since absolute pitch is usually elusive in humans. The pitch waveform may also be useful for the extraction of other complex attributes of the signal, such as roughness and loudness.

13.2 Interpretation in Terms of Auditory System

At the present time, very little is known about the auditory system beyond the peripheral level. For example, it is not at all clear at what level of the brain stem a neural mechanism for processing analogous to the pseudo spectral extraction process should be sought. Nonetheless, it is instructive to consider how a mechanism such as the GSD algorithm might be realized using simple units that are at least feasible neurologically. If the input to the GSD algorithm were a sequence of pulses instead of a waveform, then the difference waveform in the denominator would reduce to an XOR gate, with a suitable narrow time window over which the delayed input and the undelayed input "coincide". The threshold that is subtracted from the numerator, and the soft-limit that is superimposed on the division, resemble features that are often properties of neural cells.

Figure 13.1 shows a schematic "neural" model for the GSD. This Figure should be compared with Figure 8.7, where the GSD is described mathematically. The sum waveform in the numerator is realized as an adder, and the difference waveform reduces to an XOR gate. Both sum and

difference waveforms are passed through leaky integrators, and then the division is schematized by an excitatory/inhibitory unit. This unit or "cell" would have a minimal response threshold, related to the δ in the numerator, and a saturation level, as a mapping of the arctan function. This diagram is of course a grossly simplified model; for example, we have left out a mechanism for combining the outputs of a large number of fibers with similar response patterns to obtain a sufficient sampling of the statistical response. It is only meant to be suggestive of the sorts of neurological devices that would be appropriate for a GSD type of implementation.

The generation of the sum of all channel outputs for pitch extraction is not visualized as happening in a single massive summation of all fiber responses on the 8th nerve, but rather as a cascade of local sums followed by sums of sums, etc. The first-stage local sums would certainly also have utility in representing a better statistical sampling of the corresponding localized frequency region than is available from any single fiber. Thus these localized sums could be fed to the spectral analyzer as well as to the next stage summer, that would eventually lead to the pitch waveform.

The GSD algorithm is well-suited for other tasks such as lateralization and echolocation as well as pitch and spectral estimation from speech. In some cases, the two waveforms to be compared would be distinct from one another; for example, in lateralization the important parameter is the delay between the arrival at one ear and the arrival at the other. It seems attractive in an evolutionary sense to make use of a previously available set of tools for a new task, by modifying and adjusting certain parameters such as delays or node connections, rather than inventing entirely new computations.

13.3 Potential Improvements to the Thesis System

Some obvious, relatively straightforward changes in the system described in this thesis are to improve the peripheral level model and to extend the frequency range for spectral analysis beyond the arbitrary 2700 Hz cutoff frequency. In order for the frequency extension to make sense, a model for partial synchrony would have to be included at the peripheral level. The model for the auditory periphery used in the thesis did not account for partial loss of synchrony as the f_c of the peripheral level filter becomes high.

There has been a growing recent interest in developing models for peripheral auditory processing. Many of these destroy the synchronous information in the transition from the output of the linear filter to the nerve-fiber rate response. However, the model recently developed by Jont Allen [1984] retains the synchronous response, and models the nonlinearities such as half-wave rectification and partial loss of synchrony in a physiologically meaningful way. It would be interesting to see how the overall system changes if this physiologically well-motivated model were substituted for the model used in the thesis. Of course, the model would still be incomplete; our understanding of such important effects as two-tone inhibition [Sachs and Kiang, 1968] and the influence of the efferents is not yet thorough enough to suggest an accurate model for these effects.

Another interesting possibility is to replace the peripheral auditory model with some actual auditory data from cats' ear experiments. Histogram data taken from the measured responses

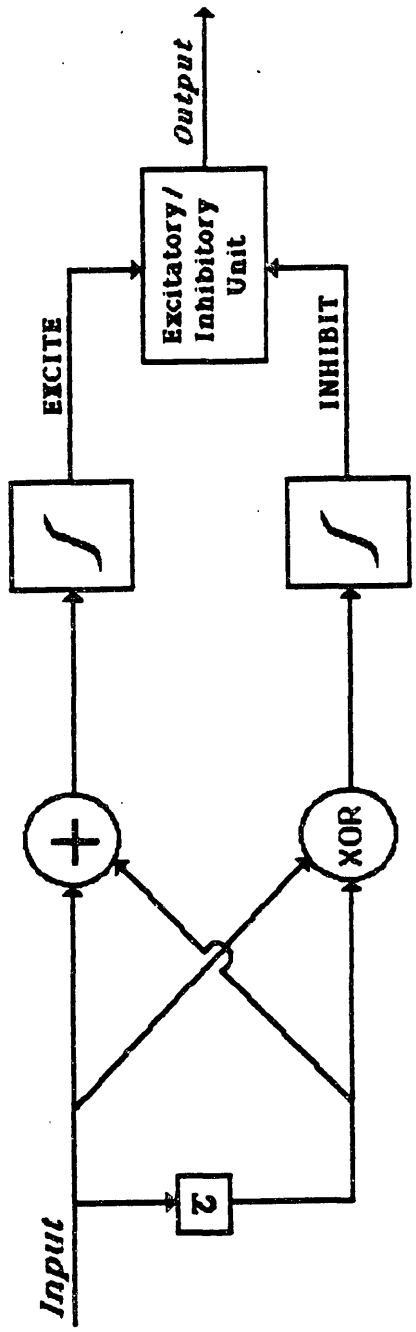

Figure 12.1 Block diagram of model for GSD algorithm based on a schematized "neural" interpretation. This is to be compared with Figure 8.7.

could be used as the inputs to the GSD's for spectral analysis. Of course, such long-term averaging of the output of a single nerve fiber would not be possible in the actual auditory system. However, a period histogram taken over a long time from a single fiber is not unlike a signal generated by adding together the outputs of a collection of adjacent fibers with similar response characteristics. The latter is a clear possibility for the auditory system. Adjacent fiber responses have been shown to be statistically independent [Johnson and Kiang, 1976], and therefore combining the responses of two fibers over the same cycle of the signal would be very similar to combining the responses of a single fiber over two cycles in sequence.

If enough data were available for a major portion of the frequency space, it might also be feasible to add together the measured histogram outputs of the individual fibers to produce a neurological pitch waveform. This waveform could then be delived to the pitch GSD's to generate pseudo autocorrelations, which could be evaluated for their utility for pitch extraction. Experiments by Sachs and Young and by Delgutte have worked with large populations of nerve fibers simultaneously, and hence data of this sort should be available.

13.4 Applying Pseudo Spectrum to Speech Recognition

An ultimate test for the utility of the pseudo spectral analysis proposed in this thesis is the performance of the resulting spectral representation in a speech recognition task. Such a step would require considerable care in the design of the supporting recognizer. It is difficult to translate even an idealized spectrum into a set of features from which reliable phonetic identification is possible. It is likely that, just as ear models should aid in determining a method for acoustic processing, brain models should aid in the design of a phonetic identifier. It may be appropriate to borrow from existing methods developed for vision. The pseudo spectrogram resembles a visual image. Strategies such as edge detection that have yielded some success in the field of vision [Marr, 1982] may also be appropriate in this context.

An attempt to automatically extract the frequencies of the first, second, and third formants is probably an inadequate approach. Such complexities as the individual harmonics of the pitch in the low frequency region for female speech, and the merger of F_2 with F_3 in certain phones such as /ɝ/, make it unlikely that any, even complex, peak-picking algorithm will succeed. A possibility is to represent steady-state phonetic elements by a specific spectral weighting function. A weighted average of the samples of the pseudo spectrum could be used to quantify the match to a particular phoneme. Positive weights would associate with regions where a peak is predicted, and negative weights would represent valleys. The actual values of the weights could be determined through some sort of automatic training procedure, including clustering to account for variability due to speaker or phonetic environment.

The pseudo spectrum will probably be a better choice than other available spectral representations because it in general has sharper formant peaks, exhibits stability over time, and is relatively insensitive to variations in energy level, both locally and globally. For example, when valleys are expressed in dB's they can show a wide range of absolute level. In the pseudo spectrum, valleys are

always near a zero level, and peaks typically range from 3.0 to 4.0. Thus there is a tight control over the absolute levels, a near necessity for this sort of weighting procedure to make sense.

Formant movements, such as the rapid rise of F_2 in a labial-front vowel transition, probably should be captured by a mechanism quite different from simply extracting a formant peak at each frame, and detecting its direction of movement. Information about the movement of the peak is available not only from the peak itself but also, for example, from its upper and lower edges, which move in the same direction as the peak. One possibility is a gradient strategy, whereby each sample located in frequency and time would "determine" which direction it could move in the time/frequency space with the least amount of "resistance", measured as change in amplitude. In other words, from the reference point of a particular time/frequency location, the amplitudes of all of the surrounding samples in the time/frequency space would be examined in turn. The one with the least amount of change in amplitude from the reference would be selected, and the direction of motion would then be defined by the direction in the space of the vector from the reference to the selected sample. Perhaps there would be two directions selected, one forward in time and one backward in time. A local majority rule would then decide whether there is a general trend upward or downward or straight across in the movement over time of a spectral energy concentration. Such a strategy would allow both the upper and lower edges of the peak to contribute to detecting upward movement, as well as the peak itself. Such formant movements are known to be very important, from our experiences with spectrogram reading, in the identification of diphthongs and of consonants adjacent to vowels. It is hoped that a gradient strategy would be more robust than methods based on peak picking; at least catastrophic errors are not anticipated.

As in the case of the recognition of steady-state phonemes, the pseudo spectrum should be more suitable than other spectral methods for the detection of formant movements. It is useful in this context to compare Figures 9.12, and 9.13 showing narrow-band and pseudo spectral analysis respectively of the sonorant region of the word "wish". The second formant rises very rapidly over the entire region, but the upward movement of the peak would probably be harder to track using the narrow-band representation, even with some smoothing to remove the harmonic structure, than using the pseudo spectrum. In the pseudo spectrum, the formant movement is quite clear to the eye, and the smooth representation should make gradient measures feasible.

Once information is available about trends in formant movements and steady-state formant patterns, the next step is to use this information to deduce goodness of fit to phonetic sequences. Ultimately, word proposals would appear, which would be strung together, using some grammatical constraints, for sentence recognition. The details of algorithms for these stages will not be addressed here.

Although it is important to evaluate the worth of the pseudo spectrum through studies of its performance in computer speech recognition devices, it is hard to devise an appropriate evaluation scheme. So much depends upon how well the recognition system is matched to the particular form of the spectral representation, as well as on how higher level aspects of the problem are handled. The task is so large that it will be necessary to break it down into subtasks such as identification at only the phoneme level, or recognition restricted to words containing a limited subset of the phonetic space. Nonetheless, it is an important area of research which must be attempted before we can hope to show that auditory modelling is at all useful for computer speech recognition.

References

1. Allen, J.B. (1983) "Magnitude and Phase-Frequency Response to Single Tones in the Auditory Nerve," JASA **73**, 2071-2092.

2. Allen, J.B. (1984) "A Hair Cell Model of Neural Response," Submitted for publication to JASA.

3. Anderson, D. J. (1973) "Quantitative Model for the Effects of Stimulus Frequency on Synchronization of Auditory Nerve Discharges," JASA **54**, 361-364.

4. Bilsen, F.A. (1966) "Repetition Pitch: Monaural Interaction of a Sound with the Repetition of the Same, but Phase Shifted, Sound" Acustica **17**, 295-300.

5. Bilsen, F.A., and Ritsma, R.J. (1967) "Repetition Pitch Mediated byTemporal Fine Structure at Dominant Spectral Regions" Acustica **19**, 114-116.

6. Blomberg, M., R. Carlson, K. Elenius, and B. Granstrom (1983) "Auditory Models in Isolated Word Recognition," Paper Number 17.9 presented at ICASSP '84 in San Diego, CA.

7. Carlson, R., G. Fant, and B. Granstrom (1975) "Two-formant Models, Pitch and Vowel Perception," in **Auditory Analysis and Perception of Speech**, G. Fant and M.A.A. Tatham, Ed., Academic Press, London, New York, San Francisco, 55-82.

8. Carlson, R., B. Granstrom, and D. Klatt (1979) "Vowel Perception: The Relative Perceptual Salience of Selected Acoustic Manipulations," Speech Transmission Laboratory Quarterly Progress Report, STL-QPSR 3-4, p. 73-83.

9. Chistovich, J. A., Grostrem, M. P., Kozhevnikov, V. A., Lesogor, l. W., Shupljakov, V. S., Taljasin, P. A., and Tjulkov, W. A. (1974) "A Functional Model of Signal Processing in the Peripheral Auditory System," Acustica **31**, 349-353.

10. Chistovich, L. A., R. L. Sheikin, and V. V. Lublinskaja [1979], ' 'Centres of Gravity' and Spectral Peaks as the Determinants of Vowel Quality," in /bf Frontiers of Speech Comminication Research, B. Lindblom and S. Ohman, Eds., Academic Press, London, 143-157.

11. Colburn, H.S. (1973) "Theory of Binaural Interaction based on Auditory-Nerve Data I. General Strategy and Preliminary Results on Interaural Discrimination," JASA **54**, 1458-1470.

12. Dave, R.V. (1977) "Studies in Gujarati Phonology and Phonetics," PhD Thesis, Cornell University, Ithaca, N.Y.

13. Davis, H. (1958) "A Mechano-Electrical Theory of Cochlear Action," Ann. Otol. Rhinol. Laryngol. **67**, 789-801.

14. deBoer, E. (1956) "Pitch of Inharmonic Signals," Nature (Lond.) **178**, 535-536.

15. deBoer, E. (1967) "Correlation Studies Applied to the Frequency Resolution of the Cochlea," J. Aud. Res. 7, 209-217.

16. deBoer, E., (1974) "On the 'Residue' and Auditory Pitch Perception," Chapter 13 in **Handbook of Sensory Physiology**, Vol. 5/1, W.D. Keidel and W.D. Neff, ed., Springer, Berlin.

17. deBoer, E., and H. R. Jongh (1978) "On Cochlear Encoding: Potentialities and Limitations of the Reverse-Correlation Technique," JASA **63**, 115-135.

18. Delgutte, B. (1980) "Representation of Speech-like Sounds in the Discharge Patterns of Auditory-nerve Fibers," JASA **68**, 843-857.

19. Delgutte, B. (1984) "Speech Coding in the Auditory Nerve: II. Processing Schemes for Vowel-like Sounds," JASA **75**, 879-886.

20. Delgutte, B. and N.Y.S. Kiang (1984) "Speech Coding in the Auditory Nerve: I. Vowel-like Sounds," JASA **75**, 866-878.

21. Dolmazon, J. M., Bastet, L., and Schupljakov, V. S. (1977) "A Functional Model of Peripheral Auditory System in Speech Processing," IEEE Acoust. Speech and Signal Process. Rec., April, 261-264.

22. Dubnowski, J.J., R. W. Schafer, and L.R. Rabiner (1976) "On the Use of Autocorrelation Analysis for Pitch Detection," IEEE Trans. Acoust. Speech and Signal Proc., **ASSP-24**, 2-8.

23. Duifhuis, H., L.F. Willems, and R.J. Sluyter (1982) "Measurement of Pitch in Speech: An Implementation of Goldstein's Theory of Pitch Perception," JASA **71**, 1568-1580.

24. Evans, E.F. (1977) "Frequency Selectivity at High Signal Levels of Single Units in Cochlear Nerve and Nucleus," in **Psychophysics and Physiology of Hearing**, E.F. Evans and J.P. Wilson, Eds., Academic Press, New York and London, 185-192.

25. Evans, E.F. and P.G. Nelson (1973) "The Responses of Single Neurones in the Cochlear Nucleus of the Cat as a Function of their Location and Anaesthetic State," Exp. Brain Res. **17**, 402-427.

26. Evans, E.F. and A.R. Palmer [1975] "Responses of Single Units in the Cochlear Nerve and Nucleus of the Cat to Signals in the Presence of Bandstop Noise," J. Physiol. (Lond.) **252**, 60-62P.

27. Evans, E.F., and Wilson, J.P. (1973) "The Frequency Selectivity of the Cochlea" in A. R. Moller (Ed.), **Basic Mechanisms in Hearing**, New York and London, Academic Press.

28. Fant, G. (1970) **Acoustic Theory of Speech Production**, Mouton & Co. N.V., The Hague.

29. Flanagan, J.L. (1962) "Models for Approximating Basilar Membrane Displacement; Part II: Effects of Middle-Ear Transmission and some Relations between Subjective and Physiological Behavior," Bell System Tech. J. **41**, 959.

30. Flanagan, J.L. (1972) **Speech Analysis Synthesis and Perception**, Springer-Verlag, Berlin.

31. Fletcher, H. (1953) **Speech and Hearing in Communication**, D. Van Nostrand Company, Inc., Princeton, N.J.

32. Fletcher, H. and W.A. Munson (1933) "Loudness, Definition, Measurements, and Calculation," JASA **5**, 82-108.

33. Fletcher, H. (1940) "Auditory Patterns," Review of Modern Physics, **12**, 47-65.

34. Godfrey, D.A., N.Y.-S. Kiang, and B.E. Norris (1975) "Single Unit Activity in the Posteroventral Cochlear Nucleus of the Cat," J. Comp. Neurol. **162**, 247-268.

35. Gold, B. and L.R. Rabiner (1962) "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," JASA **46**, 442-448.

36. Goldhor, R. (1983) "A Speech Signal Processing System Based on a Peripheral Auditory Model," Paper No. 28.11 presented at ICASSP '83, Boston, MA.

37. Goldhor, R. (1985) "Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features," Ph.D Thesis, Dept. of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA.

38. Goldstein, J. (1967) "Auditory Nonlinearity," JASA **41**, 676-689.

39. Goldstein, J. (1973) "An Optimal Processor Theory for the Central Information of the Pitch of Complex Tones," JASA **54**, 1496-1516.

40. Goldstein, J.L., A. Gerson, P. Srulovicz, and M. Furst (1978) "Verification of the Optimal Probabilistic Basis of Aural Procesing in Pitch of Complex Tones," JASA **63**, No. 2, 486-497.

41. Green, D. M. (1976) **An Introduction to Hearing**, Lawrence Erlbaum Associates, Hillsdale, N. J.

42. Hall, J.L. (1980) "Cochlear Models: Two-tone Suppression and the Second Filter," JASA **67**, 1722-1728.

43. Harris, D.M., and P. Dallos (1979) "Forward Masking of Auditory Nerve Fiber Responses," J. Neurophys. **42**, 1083-1107.

44. Hawkins, S. and K. Stevens (in Press) "Acoustic and Perceptual Correlates of the Nasal-Nonnasal Distinction for Vowels," to be published in JASA.

45. Helmholtz, H. L. F. (1863) **Die Lehre von den Tonempfindungen als Physiologische Grundlage fur die Theorie der Musik**, First ed. Brunswick:Vieweg. Eng. trans. by A. J. Ellis, 1885.

46. Hudspeth, A. J. and D.P. Corey (1977) "Sensitivity, Polarity, and Conductance Change in the Response of Vertebrate Hair Cells to Controlled Mechanical Stimuli,"Proc. Natl. Acad. Sci. U.S.A. **74**, 2407-2411.

47. Itakura, F. (1975) "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, and Signal Process., ASSP-23, 67-72.

48. Jeffress, L.A., H.C. Blodgett, T.T. Sandel, and C.L. Wood III (1956) "Masking of Tonal Signals", JASA **28**, 416-426.

49. Jeffress, L.A. (1972) "Binaural Signal Detection: Vector Theory," Chapter 9 in Tobias, Vol. II.

50. Johnson, D.H. (1970) "Statistical Relationships between Firing Patterns of Two Auditory-Nerve Fibers," S.M. and S.B. Theses, Dept. of Electrical Engineering, M.I.T., Cambridge, MA.

51. Johnson, D.H. (1974) "The Response of Single Auditory-Nerve Fibers in the Cat to Single Tones: Synchrony and Average Discharge Rate," Ph.D Thesis, Dept. of Electrical Engineering, M.I.T., Cambridge, MA.

52. Johnson, D. H. (1980) "The Relationship between Spike Rate and Synchrony in Responses of Auditory-nerve Fibers to Single Tones," JASA **68**, 1115-1122.

53. Johnson, D.H., and N. Y-S. Kiang (1976) "Analysis of Discharges Recorded Simultaneously from Pairs of Auditory Nerve Fibers," Biophys. J. **16**, 719-734.

54. Johnstone, B.M., and J.J.F. Boyle (1967) "Basilar Membrane Vibrations Examined with the Mossbauer Technique," Science **158**, 390-91.

55. Johnstone, B.M., Taylor, K.J. and Boyle, A.J. (1970) "Mechanics of Guinea Pig Cochlea," JASA **47**, 504-509.

56. Kane, E.C. (1973) "Octopus Cells in the Cochlear Nucleus of the Cat: Heterotypic Synapses on Homeotypic Neurons," Intern. J. Neurosci. **5**, 251-279.

57. Khanna, S. M. and D.G.B. Leonard (1982) Basilar Membrane Tuning in the Cat Cochlea, Science **215**, 305-306.

58. Kiang, N. Y-S., T. Watanabe, E.C. Thomas, and L.F. Clark (1965) **Discharge Patterns of Single Fibers in the Cat's Auditory Nerve**, Research Monograph No. 35, The M.I.T. Press, Cambridge, Mass.

59. Klatt, D.H. (1980)"Software for a Cascade/parallel Formant Synthesizer," JASA **67**, 979-995.

60. Licklider, J.C.R. (1954) "Periodicity Pitch and Place Pitch," Paper presented at 47th Meeting of th Acoustical Society of America, JASA **26**, 945.

61. Licklider, J.C.R. (1959) "Three Auditory Theories" in **Psychology: A Study of a Science**, S. Koch, Ed., McGraw-Hill, New York.

62. Lyon, R. (1982) "A Computational Model of Filtering, Detection, and Compression in the Cochlea," IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings, pp1282-1285.

63. Lyon, R. (1983) "A Computational Model of Binaural Localization and Separation," Paper No. 24.9, IEEE International Conference on Acoustics, Speech and Signal Processing, Boston, MA, April 14-16.

64. Makhoul, J. (1975) "Linear Prediction: A Tutorial Review," Proc. IEEE **63**, 561-580.

65. Markel, J.D. and A. H. Gray, Jr. (1976) **Linear Prediction of Speech**, Springer-verlag, New York.

66. Marr, D. (1982) **Vision**, W. H. Freeman and Co., San Francisco.

67. Miller, M.I. and M.B. Sachs (1981) "Temporal Representations of CV Syllables in Populations of Auditory Nerve Fibers," JASA **70** Suppl. 1, S9.

68. Moore, B.C.J. (1982), **An Introduction to the Psychology of Hearing**, Second edition, Academic Press.

69. Moore, B.C.J. and S.M. Rosen (1978) "Tune Recognition with Reduced Pitch and Interval Information," Q. J. Exp. Psychol. **31**, 229-240.

70. Moorer, J.A. (1974) "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech," IEEE Trans on Acoustics, Speech, and Sign. Proc, **ASSP-22**, 330-338.

71. Moushegian, G., A. Rupert, and M. Whitcomb (1972) "Processing of Auditory Information by Medial Superior-Olivary Neurons," Chapter 7 in Tobias, Vol II.

72. Nordmark, J. (1963) "Some Analogies Between Pitch and Lateralization Phenomena," JASA **35**, 1544-1547.

73. Oppenheim, A.V., and R.W. Schafer (1975) **Digital Signal Processing**, Prentice-Hall, Inc., Englewood Cliffs, N.J.

74. Pickles, J.O. (1982) **An Introduction to the Physiology of Hearing**, Academic Press, 154-193.

75. Patterson, R. D. (1973) "The Effects of Relative Phase on the Number of Components in Residue Pitch,", JASA **53** No. 6, pp1565-1572.

76. Patterson, R. D. (1976) "Auditory Filter Shapes Derived with Noise Stimuli," JASA **59**, 640-654.

77. Plomp, R. (1965) "Tonal Consonance and Critical Bandwidth," JASA **38**, 548-560.

78. Rabiner, L.R., M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal (1976) "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. Acoust. Speech, and Signal Proc., **ASSP-24**, 399-418.

79. Rabiner, L. R., and R. W. Schafer (1978) **Digital Processing of Speech Signals**, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

80. Rhode, W.S. (1971) "Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys using the Mossbauer Technique," JASA **49**, 1218-1231.

81. Ritsma, R.J. (1962) "Existence Region of the Tonal Residue," JASA **34**, 1224-1229.

82. Ritsma, R.J. (1967) "Frequencies Dominant in the Perception of the Pitch of Complex Sounds," JASA **42**, 191-198.

83. Ritsma, R.J., and F. L. Engel (1964) "Pitch of Frequency-Modulated Signals," JASA **36**, No. 9, 1637-1644.

84. Rose, J. E., L.M. Kitzes, M.M. Gibson, and J.E. Hind (1974) "Observations on Phase-Sensitive Neurons of Anteroventral Cochlear Nucleus of the Cat: Nonlinearity of Cochlear Output," J. Neurophysiol. **37**, 218-253.

85. Sachs, M.B. and N.Y.S. Kiang (1968) "Two-Tone Inhibition in Auditory-Nerve Fibers," JASA **43**, 1120-1128.

86. Sachs, M.B. and Young, E.D. (1979) "Encoding of Steady-state Vowels in the Auditory-nerve: Representation in Terms of Discharge Rate," JASA **66**, 470-479.

87. Sachs, M. B., and E. D. Young (1980) "Effects of Nonlinearities on Speech Encoding in the Auditory Nerve," JASA **68**, 858-875.

88. Sachs, M.B., E.D. Young, and M.I. Miller "Encoding of Speech Features in the Auditory Nerve," manuscript in preparation.

89. Scheffers, M.T.M. (1983) Sifting Vowels, Auditory Pitch Analysis and Sound Segregation, Doctoral Thesis, Institute for Perception Research, Eindhoven, the Netherlands.

90. Schouten, J.F. (1938) "The Perception of Subjective Tones," Proceedings Kon. Acad. Wetensch. (Neth.) **41**, 1086-1094.

91. Schouten, J.F. (1940a) "The Residue, a New Component in Subjective Sound Analysis," Proc. Kon. Acad. Wetensch. (Neth.) **43**, 356-365.

92. Schouten, J.F. (1940b) "The Residue and the Mechanism of Hearing," Proc. Kon. Acad. Wetensch. (Neth.) **43**, 991-999.

93. Searle, C.L., J. Z. Jacobson, and S.G. Rayment (1979) "Stop Consonant Discrimination based on Human Audition," JASA **65**, 799-809.

94. Searle, C.L., J.Z. Jacobson, and B.P. Kimberley (1980) "Speech as Patterns in the 3-Space of Time and Frequency," Chapter 3 in **Perception and Production of Fluent Speech**, R. Cole, Ed.

95. Seebeck, A. (1843) "Uber die Sirene," Ann. Phys. Chem. **60**, 449-481.

96. Seneff, S. (1978) "Real-time Harmonic Pitch Detector," IEEE Trans. Acoust., Speech, Signal Proc. **26**, 358-365.

97. Shaw, E.A.G. (1974) "The External Ear," in **Handbook of Sensory Physiology**, Vol. 5/1, W.D. Keidel and W.D. Neff, ed., Springer, Berlin, 455-490

98. Siebert, W.M. (1973) "What Limits Auditory Performance?" Symposial paper in International Biophysics Congress, Academy of Sciences of the USSR, 399-413.

99. Smith, J.C., J.T. Marsh, S. Greenberg, and SW.S. Brown (1978) "Human Auditory Frequency-Following Responses to a Missing Fundamental," Science, **201**, 639-641.

100. Smith, J.C., and J.J. Zwislocki (1975) "Short-Term Adaptation and Incremental Responses of Single Auditory-Nerve Fibers," Biol. Cybernetics **17**, 169-182.

101. Sondhi, M.M. (1968) "New Methods of Pitch Extraction," IEEE Trans. Audio and Electroacoustics, **AU-16**, 262-266.

102. Srulovicz, P. and Goldstein, J.L. (1983) "A Central Spectrum Model: a Synthesis of Auditory-nerve Timing and Place Cues in Monaural Communication of Frequency Spectrum," JASA **73**, No. 4, 1266-1276.

103. Terhardt, E.,G. Stoll, and M. Seewann (1982) "Algorithm for Extraction Pitch and Pitch Salience from Complex Tonal Signals," JASA **71**, 679-688.

104. Tobias, J.V., Ed. (1972) **Foundations of Modern Auditory Theory, Vol. II.**, Academic Press, New York.

105. Vogten, L. L. M. (1974) "Pure Tone Masking; a New Result from a New Method," in **Facts and Models in Hearing**, E. Zwicker and E. Terhardt, ed., Springer-Verlag, Berlin.

106. Voigt, H.F. and E.D. Young (1980) "Evidence for Inhibitory Interactions between Neurons in Dorsal Cochlear Nucleus," J. Neurophysiol. **44**, 76-96.

107. von Bekesy, G. (1928) Zur Theorie des Hoerens; Die Schwingungsform der Basilarmembran," Physik. Zeits., **29**, 793-810.

108. von Bekesy, G. (1960) **Experiments in Hearing** (edited and translated by E.G. Wever) McGraw-Hill, New York.

109. Weiss, T.F. (1966) "A Model of the Peripheral Auditory System," Kybernetik **3**, 153-175.

110. Weiss, T.F. and R. Leong, "Model for Signal Transmission in the Alligator Lizard Ear: IV. Mechanoelectric Transduction," manuscript in preparation.

111. Whitfield, I.C. (1970) "Central Nervous Processing in Relation to Spatio-Temporal Discrimination of Auditory Patterns," in **Frequency Analysis and Periodicity Detection in Hearing**, Plomp, R. and Smoorenburg, G.F., Eds., Sijthoff, Leiden, the Netherlands.

112. Wightman, F.L. (1973) "The Pattern-transformation Model of Pitch," JASA **54**, 397-406. Wightman, F.L. and Green, (1974) "The Perception of Pitch," American Scientist, **62**, 208-215.

113. 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," JASA 'bf 66, 1381-1403.

114. Zwicker, E. (1961) "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," JASA **33**, 248-249.

Appendix A

Filter Design Strategy

A.1 Overview

The filter design strategy was motivated by the following specific goals:

1. Each filter should have a bandwidth between 3dB points set to agree with psychophysical tuning curves [Zwicker].

2. Slopes on rise and fall should be similar to those measured from physiological studies of nerve fibers.

3. The phase should be essentially linear in the frequency region below resonance, and the phase curve should steepen as resonance is approached [Rhode,1971], passing through a shift of approximately 2π through resonance [Fletcher,1953].

It was decided to generate the resonance at center frequency by means of a double complex pole at a radius near the unit circle. This decision was mainly motivated by the observation that phase response passes through an approximate 2π shift as the input frequency crosses through resonance (Fletcher,1953). In addition to the double complex pole, zeros are included somewhat arbitrarily on the unit circle and on the real axis to improve the slope characteristics of the filter.

The filter design procedure is to first process the original speech sampled at 16kHz through an FIR filter to remove high frequency components and DC, and then process the output of this filter through a parallel bank of filters set up to achieve approximately critical bandwidths, with a steep slope above resonance (nominally 25dB/Bark) and a lesser slope on the low side (10dB/Bark).

Each filter consists of the double complex pole pairs as mentioned above plus a double complex zero pair positioned on the unit circle above center frequency, necessary to increase the slope on the high frequency side, and a fourth-order zero on the x-axis, which is near $x = -1$ for low frequency filters and near $x = +1$ for the highest frequency filters. The x-axis zeros serve a dual role: for the low f_c filters they further attenuate the unwanted high frequencies, and for the high f_c filters they provide additional losses below center frequency.

The difficult part of the filter design is to determine a positioning of the zeros that will give slopes in the low and high ends that correspond to the measured slopes in nerve fiber responses, and to determine a value for the radius of the double pole at the filter center frequency that will give an approximately critical bandwidth. The approach is to first consider the response of the system at frequencies local to the center frequency, f_c, when the double pole at $+f_c$ is not included, but the one at $-f_c$ is included, with the radius approximated as 1.0. Thus only the poles and zeros

distal to the center frequency are included, in order to decide where to locate the 3dB points above and below f_c. The difference between the response of this incomplete system at f_c + 1/2 Bark and the response at f_c - 1/2 Bark is used as an approximation to the slope of the response, which, in turn, is used to locate two frequencies, x and y, defined as the frequencies above and below f_c respectively where the 3dB requirement should be met. For example, if the slope near f_c is strongly negative, then x should be almost one Bark below f_c and y should be just barely above f_c. The total distance between x and y should always be one Bark.

By neglecting the curvature of the unit circle near f_c and assuming a linear slope on the response of the filter when the double pole at f_c is not included, it is possible to determine x, y and finally r such that x and y are both 3dB points. The last step is to determine a gain term for the filter, defined as the reciprocal of the response of the filter at f_c, once the double pole has been reinstated with the proper value for r. With the gain term included, the response at f_c would then be 1.0.

A.2 Step-by-Step Procedure

1. Let $f_l = f_c - 1/2$ Bark, $f_h = f_c + 1/2$ Bark, and $BW = f_h - f_l$, in Hz, using the following formula for converting from Hertz scale to Bark scale:

   Let B = frequency in Barks and f = frequency in Hz.

   Then,

$$B(f) = \begin{cases} .01f, & 0 \le f < 500 \\ .007f + 1.5, & 500 \le f < 1220 \\ 6\ln f - 32.6, & 1220 \le f \end{cases}$$

2. Compute response at f_l, f_c, and f_h, by including the distal double pole at - f_c, but excluding the proximal double pole at f_c. Label these g_l, g_c, and g_h respectively. [Assume the distal pole has a radius of 1.0 for simplicity.]

3. Compute the normalized slope, m, of the response function, local to f_c, approximated by

$$m = \frac{g_h - g_l}{BW g_c}$$

4. Using a linear approximation to the arc of the unit circle near f_c and the requirement that the gain should be down by 3dB at both x and y, solve for $a (= f_c - x)$, according to the following formula, which takes into account the measured slope (See later for derivation).

   Let $k = \sqrt{2}$ be the ratio of the amplitude at the center frequency to the amplitude at the edge of the critical band.

   Define:

$$K = \frac{k-1}{|m|\,k}$$

$$z = K + BW/2 - \sqrt{(BW/2)^2 + K^2}$$

Then,

$$a = \begin{cases} z, & m \geq 0 \\ BW - z, & m < 0 \end{cases}$$

$$\delta = \frac{BW - 2a}{mk}$$

$$r = 1.0 - \delta$$

A.3 Derivation

A.3.1 Derivation of Formulas for Setting Critical Bandwidth Radius:

To compute the amplitude response, G, at a specified radian frequency, the contribution by each pole and zero is determined using the triangle rule. For each pole at the location $r \exp j\gamma$ in the z-plane a term is included in the denominator of the form:

$$\sqrt{1 + r^2 - 2r\cos(\theta - \gamma)}$$

where $\theta = 2\pi f/sr$, and sr is the sampling rate. For each zero at location $r \exp j\gamma$, an identical term is included in the numerator.

Let G_x, G_c, and G_y be the amplitude response of the total system at frequency locations f_x, f_c, and f_y respectively, when the double pole at f_c is **included**.

With reference to Figure A.1, the following equations are evident:

[Note: The partial system has been pre-normalized to have a gain at f_c of 1.0.]

$$G_x = \frac{1 - ma}{\delta_x^2} \tag{A.1}$$

$$G_c = \frac{1}{\delta^2} \tag{A.2}$$

$$G_y = \frac{1 + mb}{\delta_y^2} \tag{A.3}$$

Given the 3dB requirements at x and y, we also know that

$$G_x = G_y = G_c/k \tag{A.4}$$

$$a + b = BW \tag{A.5}$$

Also, if a linear approximation to the unit circle arc near f_c is assumed:

$$\delta_x^2 = a^2 + \delta^2 \tag{A.6}$$

$$\delta_y^2 = b^2 + \delta^2 \tag{A.7}$$

Hence,

$$G_x = G_y = \frac{1}{k\delta^2} = \frac{1+mb}{\delta_y^2} = \frac{1+mb}{\delta^2 + b^2}$$

Substituting $(BW - a)$ for b and simplifying,

$$G_x = \frac{1 - ma + mBW}{\delta^2 + (BW - a)^2}$$

Substituting $G_x\delta_x^2$ for $(1 - ma)$, and then $1/(k\delta^2)$ for G_x and $(\delta^2 + a^2)$ for δ_x^2:

$$\frac{1}{k\delta^2} = \frac{(\delta^2 + a^2)/k\delta^2 + mBW}{\delta^2 + (BW - a)^2}$$

Solving for δ:

$$\delta^2 = \frac{BW - 2a}{mk} \tag{A.8}$$

Now we need to find an expression for a independent of δ.

From Equations A.1, A.2, A.4, and A.6:

$$a^2 + \delta^2 = k\,\delta^2(1 - ma)$$

$$\delta^2 = \frac{a^2}{k - ma + 1} \tag{A.9}$$

Equating RHS's of Equations A.8 and A.9 and simplifying:

$$a^2 - 2a\left(\frac{BW}{2} + \frac{k-1}{mk}\right) + BW\left(\frac{k-1}{mk}\right) = 0$$

Letting $K = (k-1)/mk$, and solving for a using the quadratic formula:

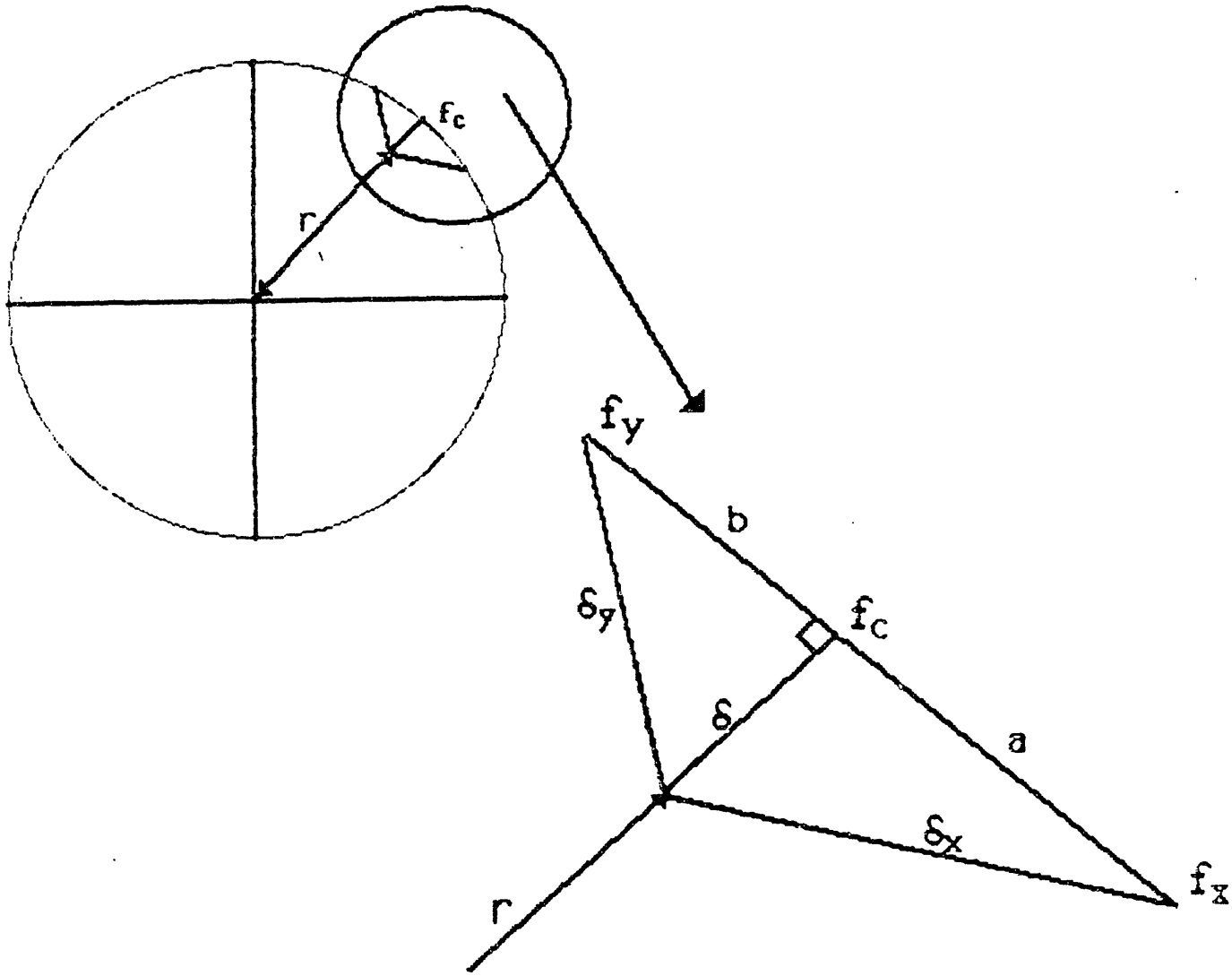

Figure A.1: Geometric configuration of linear approximation to unit circle used to estimate radius of pole at resonance frequency, f_c. f_x and f_y are the frequencies of the two 3dB points, to be determined from the geometry. The arc of the unit circle from f_y to f_x has been approximated by a straight line, allowing the equations to be simplified enormously.

$$a = \frac{BW}{2} + [K + \sqrt{K^2 + \frac{BW^2}{2}}]$$

$$b = \frac{BW}{2} - [K + \sqrt{K^2 + \frac{BW^2}{2}}]$$

By considering the case $m > 0$, which constrains a to be less than b and K to be positive, we can observe that the part in brackets will only be less than zero if we choose the negative square root. Likewise, if m is negative we should choose the positive square root.

Hence, we can solve for a, then plug into Equation A.8 to solve for δ, and finally obtain

$$r = 1 - \delta$$

the radius of the double pole.

A.3.2 Determining the Overall Gain Term:

The gain can be forced to be 1.0 at f_c by computing the total gain of the system, using the triangle rule as described above for each pole and zero, and then introducing a constant gain term in the system equal to the reciprocal of the computed gain. For this computation, the exact value of r, as determined above, for the double complex pole pair at resonance, should be used for both the proximal and the distal pole pairs. The desired overall gain of the system can then be achieved by adjusting the constant term according to specifications.

Biographical Note

Stephanie Seneff was born in Columbia, Missouri, on April 20, 1948. She received the B.S. Degree in Biophysics from M.I.T. in June, 1968. For the next ten years she worked at the M.I.T. Lincoln Laboratory, first as a scientific programmer, and, after two years, as a member of the technical staff. Her research efforts included computer speech recognition, speech synthesis, voice encoding techniques, models for the transmission of speech over digital networks, and digital seismic signal processing. In 1978, she returned to M.I.T. full time, receiving the M.S. and E.E. Degrees in Electrical Engineering in 1980. The Master's thesis is titled "Speech Transformation System [Spectrum and/or Excitation] without Explicit Pitch Extraction." She is the author of several technical journal articles, and has presented her work at conferences on numerous occasions. Ms. Seneff lives in Winchester, MA, with her husband, Victor Zue, her mother-in-law, Lily Zue, and her four sons, Michael and Gregory McCandless, and Timothy and Cory Zue.