#1

Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features

Richard Scott Goldhor

Technical Report 505

February 1985

Massachusetts Institute of Technology
Research Laboratory of Electronics
Cambridge, Massachusetts 02139

Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features

Richard Scott Goldhor

Technical Report 505

February 1985

Massachusetts Institute of Technology
Research Laboratory of Electronics
Cambridge, Massachusetts 02139

This work has been supported in part by the National Science Foundation Grant MCS 81-12899

REPRESENTATION OF CONSONANTS IN THE PERIPHERAL AUDITORY SYSTEM: A MODELING STUDY OF THE CORRESPONDENCE BETWEEN RESPONSE PROPERTIES AND PHONETIC FEATURES

by

RICHARD S. GOLDHOR

Submitted to the Department of Electrical Engineering and Computer Science on February 1, 1985 in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Bioelectrical Engineering.

ABSTRACT

This thesis describes a model of the peripheral auditory system, and discusses certain properties of its response to speech signals. The work was motivated by a theory regarding the role of the peripheral auditory system in the perception of speech. The theory is that the peripheral auditory system transforms speech signals into neural firing patterns with properties which correspond to the phonetic content of the signal. Whereas the phonetic features of speech are encoded in the acoustic waveform in complex and often undecipherable ways, this theory suggests that the representation of speech in the discharge patterns of the auditory nerve correspond in a more direct and straightforward manner to those underlying phonetic features. Two testable hypotheses are implied by this correspondence theory. The first is that cochlear transformations to speech signals result in auditory neural firing patterns whose apparent properties are significantly different from the apparent properties of the acoustic signal. The second is that particular properties of the neural response to speech reflect particular phonetic features of the stimulus.

To test the correspondence theory a computer model of the peripheral auditory system was constructed. The model consists of three stages, each of which effects a transformation on the acoustic input signal. The spectral analysis stage

consists of a bank of filters whose characteristics reflect the frequency response of the basilar membrane. The transduction stage models the cochlea's transformation of signal intensity to expected neural firing rate. The adaptation stage produces a temporal transformation that mimics neural adaptation to sustained stimulation. The output of the model is the expected firing rate of a representative set of auditory fibers in response to the input signal.

The hypotheses regarding auditory neural response are tested by studying the model's response to signals. A number of simple signals exhibiting speech-like spectral structure, onset and offset characteristics, and durational properties, are used to demonstrate the model's behavior and to test the validity of the first hypothesis. Then two series of speech stimuli, one of which spans the perceptual boundary between stop consonants and glides, and the other of which spans the boundary between voiced and unvoiced intervocalic stop consonants, are used to test the second hypothesis.

The results tend to support the correspondence theory of peripheral auditory function. They demonstrate that acoustic patterns of speech-like signals are transformed by the peripheral auditory model into auditory patterns with quite different properties. They also show that certain perceptual judgements of the tested speech signals by human subjects conform closely with measurable properties of the model's response. These measurable properties correspond to changes in the overall expected firing rate in the mammalian auditory nerve in response to the speech signals.

Thesis Supervisor: Kenneth N. Stevens

Title: Lebel Professor of Electrical and Bioengineering

ACKNOWLEDGEMENT

It is both pleasant and appropriate that the writing of this doctoral dissertation should end, and its reading begin, with an acknowledgement of the many people whose support encompassed its creation. I did not write this thesis alone.

I acknowledge with deep gratitude the support, patience, wisdom, and encouragement of Prof. Kenneth Stevens, my thesis advisor. His support made it possible to begin this work, his faith sustained it through the difficult middle stages, and whatever is best and most true in the final result reflects his insight and wisdom. His extraordinary qualities as a speech scientist, teacher, and advisor made my doctoral studies a unique and wonderful experience.

Both of my readers contributed in important ways to my education and research. Prof. Campbell Searle nurtured my early interest in auditory modeling with his enthusiasm, informed opinions, and good ideas. He was always available and interested when I had a new wild idea to expound. Prof. Thomas Weiss taught me much of what I know about the auditory system. The precision, clarity, and thoroughness of his teaching, research, and writing is an inspiring example.

I thank Christine Shadle and Stephanie Seneff, my fellow graduate students, for their support and encourgement. Among other things, each semester on the day we had to petition to be removed from the degree list they were kind enough to drown their sorrows with me over lunch.

I thank Dr. Dennis Klatt for his advice on speech and signal processing issues. I thank Marie Southwick and Keith North, without whom life as we know it in the Speech Communication Group would be impossible, for their efficient and willing help, and good cheer. I thank Drs. Edith Maxwell, Karen Landahl, Patti Price, and Helen Simon for their kind permission to make use of the speech tokens and perceptual data discussed in Chapter Three. Thanks to Dr. Price also for her help wrestling alligators.

I thank Dr. Francis Ganong, Ray Kurzweil, and my many friends and colleagues at Kurzweil Applied Intelligence, for encouraging my studies and for believing for so long that I really would be done in another three months.

I thank my parents for instilling in me, by their example, an appreciation of the importance and honor of doing scholarly work.

I thank my wife, Pamela, for her unstinting love and support during the long days and nights of my graduate education. That that work has been successfully completed is due in very large measure to her offering of time and attention to me and to our children, at some considerable cost to her own scholarly activities. Our children Lucy, Robbie, and Andrew have lived with this thing called a thesis for many years. I thank them for their patience, and for celebrating its completion with us.

This work was partially supported by an NSF grant, and by a C.J. Lebel Fellowship.

ad majorem Dei gloriam

BIOGRAPHICAL NOTE

Richard Scott Goldhor was born on March 26, 1951, in Champaign, Illinois. He attended the University High School of the University of Illinois, and received his undergraduate and graduate education at MIT. He earned a B.S. in physics and a B.S. in Electrical Engineering in 1973, and an M.S. and E.E. in Electrical Engineering and Computer Science in 1976. While at MIT he served as a member of the Medical Advisory Board and the Committee on Educational Policy.

From 1978 to 1983 he was the senior partner in a consulting firm, doing software development for speech synthesis and recognition systems. During 1980 and 1981 he was a Research Associate at MIT's Center for Policy Alternatives, where he studied and wrote about technology transfer between universities and industry. He is currently the Assistant Director of Research at Kurzweil Applied Intelligence, Inc.

Mr. Goldhor is a member of Sigma Xi and numerous professional societies. He holds one patent.

PUBLICATIONS

1. 1983, "University-to-Industry Technology Transfer: A Case Study," Goldhor, R.S. and R.T. Lund, Research Policy, **12**(3).

2. 1983, "A Speech Signal Processing System Based on a Peripheral Auditory Model," Proceedings of the 1983 IEEE International Conference on Acoustics, Speech, and Signal Processing, Boston.

3. 1982, "Bringing up UNIX--One User's Experience," The UNIX Software List, **1**(4); January-March 1982.

4. 1980, "University to Industry Advanced Technology Transfer," IEEE Engineering Management Conference, Boston.

5. 1976, "Bound and Quasibound States of Alkali-Rare Gas Molecules," Goldhor and Pritchard, The Journal of Chemical Physics, **64**(3).

6. 1972, "A Scanned Infrared Light Beam Touch Entry System," Ebeling, Goldhor, and Johnson, Proceedings of the International Symposium of the Society for Information Display, San Fransisco.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

FIGURES AND TABLES FOR CHAPTER 1:

FIGURES AND TABLES FOR CHAPTER 2:

FIGURES AND TABLES FOR CHAPTER 3:

FIGURES AND TABLES FOR CHAPTER 4:

CHAPTER 1

INTRODUCTION: PERIPHERAL AUDITORY RESPONSE AND THE ANALYSIS OF SPEECH

1.1 MOTIVATION AND OVERVIEW

This thesis describes a model of the peripheral auditory system (PAS), and discusses certain properties of its response to speech signals. The study was motivated by a theory regarding the role of the peripheral auditory system in the perception of speech. This theory will be referred to as the correspondence theory. It states that the peripheral auditory system transforms speech signals into neural firing patterns with properties which correspond to the phonetic content of the signal. Whereas the phonetic features of speech are encoded in the acoustic waveform in complex and often undecipherable ways, this theory suggests that the discharge patterns with which speech is represented in the auditory nerve correspond in a more direct and straightforward manner to underlying phonetic features. It is the peripheral auditory system, and in particular the cochlea, that brings about this correspondence.

The correspondence theory does not assign any phonetic detection or discrimination task to the PAS, nor does it imply that the PAS is especially "tuned" to speech, or reacts differently to speech sounds than to other sounds. Rather, it

suggests that human speech has developed in a way that takes advantage of the characteristics of the peripheral auditory system; and that the neural signal characteristics which form the basis for the identification and discrimination of at least some phonetic contrasts are established by the PAS, and are apparent in the response patterns of the auditory nerve.

The PAS effects the transformation between the representation of speech as an acoustic pressure wave and its representation as a neural firing pattern. The correspondence theory argues that the form of this transformation is as important to speech perception as its nature, and that speech perception relies on the particular and peculiar transformations of spectral shapes, intensity profiles, and temporal relationships that acoustic signals undergo as they travel through the peripheral auditory system.

From the point of view of pattern recognition, the correspondence theory states that the peripheral auditory response to speech is a better--not merely different--representation than the acoustic waveform from which to attempt to recognize speech. By implication such alternative signal representations as short term Fourier transforms and linear predictive encoding, to the extent that they do not incorporate peripheral auditory transformations, are deficient as representations on which to attempt phonetic analysis. That is, in the peripheral auditory representation

phonetic features are more discriminable, and represented in fewer dimensions, than they are in such alternative representations.

The correspondence theory depends on two testable assumptions regarding the apparent properties of acoustic and auditory representations of speech. "Apparent" is used here in the literal sense of appearing, or visible. By apparent properties are meant appreciable, surface-level features of a representation, or simple relationships between such features.

The first assumption is that cochlear transformations of speech signals result in auditory neural firing patterns whose properties are significantly different from the properties of acoustic signals or their classical representation by spectrogram-like analysis techniques. This is a necessary, though not sufficient, condition for the correspondence theory to be true: clearly if the patterns of auditory neural response are insignificantly different from, say, the patterns of wideband spectrograms, the correspondence theory cannot be true in any important way.

The second assumption is that particular properties of the neural representation of speech signals reflect particular phonetic features of the stimulus. If this were not true, then even though the PAS might radically transform the signal properties of speech, the decoding of phonetic information

from the auditory nerve response would still appear to depend wholly on the activity of the higher auditory centers in the brain. In that case, perhaps any peripheral representation of speech that did not actually delete information contained in the acoustic signal could be interpreted equally well by the central nervous system. The correspondence theory, on the other hand, implies that the PAS plays a substantial and vital part in the perception of speech.

To test this theory a computer model of the peripheral auditory system was constructed. The model consists of three stages, each of which effects a transformation on the acoustic input signal. The spectral analysis stage consists of a bank of filters whose characteristics reflect the frequency response of the basilar membrane. The transduction stage models the cochlea's transformation of signal intensity to expected firing rate. The adaptation stage produces a temporal transformation that mimics neural adaptation to sustained stimulation. The output of the model is the expected firing rate of a representative set of auditory neurons in response to the input signal.

This Peripheral Auditory Model (PAM) is described in Chapter Two. The model is primarily based on published electrophysiological data of firing patterns observed in the auditory nerve. These data describe the neural response to acoustic stimuli, and the ways in which neural firing patterns

change as a function of the onset characteristics, duration, and intensity of the stimuli. Only those features of peripheral auditory response which were thought to play an essential role in the perception of consonant-like speech sounds have been modeled. No attempt has been made to accurately reproduce all aspects of the peripheral auditory system, such as the frequency response of the outer and middle ear, or the many nonlinear phenomena associated with the basilar membrane. Similarly, the PAM is not intended to be a physiologically valid model of the peripheral auditory system.

The two hypotheses regarding the nature of auditory neural response are tested by studying the model's response to signals. In Chapter Two a number of simple signals exhibiting speech-like spectral structure, onset and offset characteristics, and durational properties, are used as stimuli to demonstrate the model's behavior and to test the validity of the first hypothesis. The correspondence of patterns of peripheral auditory response with patterns of phonetic perception is demonstrated in Chapter Three. Two experiments are described which relate measurable properties of the PAM's response patterns to perceptual judgments of phonetic contrasts elicited from human subjects. In the first experiment, the speech stimuli span the perceptual boundary between stop consonants and glides. In the second, the stimuli span the boundary between voiced and unvoiced intervocalic stop con-

sonants.

In both experiments it is shown that an analysis using a single measurable property of the output of the PAM can do as well as, or better than, a classic "trading relation" analysis based on two or more measurable acoustic signal properties such as onset time and syllable duration. Evidence is presented that two well known and very general properties of neural firing patterns, namely spontanteous firing in the absence of stimulation, and adaptation of mean firing rate in the presence of sustained stimulation, may play an important role in conforming peripheral auditory response to phonetic content.

The results of these experiments tend to support the correspondence theory of peripheral auditory function. The results show that acoustic patterns of speech-like signals are transformed by the peripheral auditory model into auditory patterns with quite different properties. They also show that certain perceptual judgements of the tested speech signals by human subjects conform closely with measurable properties of the model's response. These measurable properties correspond to changes in the overall expected firing rate in the mammalian auditory nerve in response to the speech signals.

The final chapter discusses limitations of the peripheral auditory model, and suggests improvements that should enhance

further studies. It also presents a list of known acoustic phonetic effects for which perceptual data are available, and which would be interesting to re-analyze in the auditory domain.

1.2 THE ROLE OF PERIPHERAL AUDITORY MODELS IN SPEECH ANALYSIS

The study of phonetics has traditionally relied largely on acoustic instruments for experimental data. This has been particularly true since the invention in 1946 of the sound spectrograph (Koenig et al., 1946). This instrument has proved to be of such remarkable utility and longevity that its particular form of output, the sound spectrogram, might almost seem to show speech in its True Form. Indeed, almost all discussions of acoustic phonetics use, as descriptive terms, observable properties of visual speech patterns as displayed on spectrograms: bursts, formant peaks, stop gaps, voice bars, and so forth. Even the advent of sophisticated computer-based signal processing systems has often done surprisingly little to change the fundamental terms in which speech is discussed. A popular use of computers for speech analysis consists of the production of digital spectrograms whose characteristics are modeled after their analog predecessors.

At the same time that this representation of acoustic signals has been accepted as the natural form for speech analysis, the basis for most models of speech has been the

human speech production system. Production-based analysis has been advanced by such works as Fant's Acoustic Theory of Speech Production (1960) and research by Stevens and others on acoustic theories of vowel production (Stevens et al., 1961). In the realm of computer-based analysis techniques, linear predictive encoding (LPC) has become particularly popular. LPC is a method of analysis based on the representation of speech as the product of a linear time-varying system excited by either a quasi-periodic or random-noise source. (For reviews of LPC see Makhoul and Wolfe, 1972; or Rabiner and Schafer, 1978).

These production-based models of speech have been substantially successful in predicting and analyzing the characteristics of speech waveforms. This, however, encompasses only half the story. Although spectrograms reveal a significant amount of information about speech, they often display properties that are not perceptually relevant, while obscuring important, and perceptually clear, aspects of the speech signal (Klatt, 1982). The patterns of the spectrograph are not the speech patterns which humans hear. Speech signals which form different spectrographic patterns may sound the same to human listeners, and other signals which appear imperceptibly different on the spectrogram may elicit different phonetic responses. It remains true that speech is only perceived through the response patterns of the peripheral auditory sys-

Given our current state of ignorance, the auditory nerve represents a region of manageable complexity. From the terminus of the auditory nerve in the cochlear nucleus (CN) secondary fibers branch out in many directions to locations both within the CN and to other auditory nuclei in the brain. In addition to these ascending (afferent) fibers, a variety of descending (efferent) fibers terminate in the CN nucleus, creating complex feedback loops. To further complicate the situation, the central auditory system contains fibers with a variety of response patterns, much less uniform than the response patterns exhibited by the neurons in the auditory nerve. Thus the auditory nerve represents a pathway of relative simplicity before the so-far overwhelming complexity of CNS interconnection.

(This is not to say that no feedback circuits exist in the PAS. On the contrary, the PAS is well supplied with efferent fibers originating higher up in the nervous system. Since almost all auditory nerve studies have been performed on anesthetized animals whose state prevents the efferent auditory neurons from firing, little is known about the role these fibers, and the feedback loops they support, might play in the normal perception of speech.)

A fourth reason to study the output of the PAS is that many of the transformations that take place there are the result of non-neural processes: most prominently the frequency

tem. It is through this system, consisting of the outer, middle, and inner ear, that some transformation of the acoustic pressure waves of speech sounds reach the central nervous system. In particular, it is through the firing patterns of neurons in the auditory nerve that the CNS receives all of its information about the speech signal it must interpret. Any information that is not somehow encoded in that auditory firing pattern is, ipso facto, not perceivable. This is the first reason for directing some attention to the PAS: it is the input to the auditory perceptual system of the brain.

The second reason for attending to the patterns of auditory neural response has to do with experimental practicality: the response patterns of the auditory nerve are accessible. Using micropipettes implanted in the auditory nerve of a variety of animals (most commonly cats), it is possible to directly record the firings of representative neurons that terminate on sensory cells located within the cochlea. Through the pioneering work of such researchers as Kiang and his associates (Kiang, Watanabe, Thomas, and Clark, 1965), electophysiological measurements have become relatively routine, and an extensive amount of data regarding the discharge patterns of single auditory neurons to a wide variety of both non-speech and speech-like stimuli are now available.

There is a third reason why the auditory nerve is an attractive place to study auditory response to speech stimuli.

response of the outer and middle ear, the macro- and micro-mechanical properties of the basilar membrane, and the electrochemical properties of the receptor cells. While it is clear that further radical transformations in the speech signal take place in the central nervous system, these transformations may be expected to differ in kind from the transformations observable in the PAS.

An important potential benefit that may result from an understanding of the representation of speech signals in the auditory nerve is the ability to restore hearing to people whose deafness is due to sensory loss in the cochlea. The clinical problems of direct electrical stimulation of the auditory nerve via cochlear implants are being steadily overcome, but unless adequate models of the signal processing characteristics of the peripheral auditory system are developed, the resulting auditory sensations will continue to be highly artificial, and of little use in understanding speech.

A final reason to focus on the response patterns of the PAS is that there is no evidence to suggest that the PAS in man is specialized for the perception of speech. Indeed, most of the fundamental characteristics of PAS response appear to be qualitatively similar for a wide range of mammals. Thus an investigation of the role of the PAS in speech perception can help answer the question of the extent to which speech

perception depends on speech-specific perceptual mechanisms, and to what extent it is carried out by more general auditory mechanisms. In addition, of course, any aspect of speech perception that depends on the peripheral auditory system is certain to be language independent. Therefore studying peripheral auditory response can help separate language independent from language dependent aspects of speech perception.

1.3 THE NATURE OF PERIPHERAL AUDITORY RESPONSE

This section contains a review of the major structures of the peripheral auditory system, and the ways in which they contribute to the transformation of acoustic stimulation into auditory neural response. The contribution of some of these structures to peripheral auditory response is well represented in the PAM. Other structures are represented less precisely, or not at all. The degree of representation in the PAM will be mentioned in the discussion below.

Figure 1.1 shows the anatomical structure of the outer, middle, and inner ear. Acoustic pressure waves in air are funneled through the outer ear to the eardrum, where they are converted into motion of the middle ear ossicles. The mechanical linkage of these ossicles drives the base-plate of the stapes with a piston-like motion in the oval window of the cochlea. As a result, traveling waves are set up in fluid-filled cavities and membranes of the cochlea. The resulting

motion of the cochlear membranes stimulate sensory cells enervated by the auditory nerve (the eighth cranial nerve), which carries information to the higher auditory centers of the brain.

The transformations of the outer ear are primarily ones of frequency- and direction-dependent amplification of the free-field acoustic signal (see, for example, Shaw, 1974). Among other things, the pinnae selectively enhance high-frequency sounds originating in front of the head. The concha and ear canal enhance frequencies in the mid- to high-frequency range for speech signals: roughly 2 to 7 KHz (Mehrgardt and Mellert, 1977). For a fixed orientation between the sound source and the head of the listener, the outer ear can be modeled as a linear time-invariant filter with roughly bandpass characteristics.

The middle ear also acts as a bandpass filter, enhancing frequencies in the range of 200 to 2000 Hz (Zwislocki, 1975). A nonlinear element is present in the action of the ossicular muscles, which contract in response to loud (approximately 80 dB SL) sounds, thereby reducing the transfer of pressure through the ossicular chain. Presumably this acoustic reflex action serves to protect the delicate inner ear from potentially damaging levels of stimulation. It is unlikely that this reflex plays an important role in acoustic response at intensity levels typical of conversational speech.

The transformations of the outer and middle ear are not represented in the PAM in a principled way. Instead, a simple first-difference preemphasis filter is used to counteract the 6 dB/octave roll-off of voiced speech sounds at high frequency. The acoustic reflex of the ossicular muscles is not represented at all.

Sound waves are transferred to the inner ear via the piston-like motion of the stapes in the oval window of the vestibule to the cochlea. The resulting fluid displacement results in a vibratory pattern of motion in the structures surrounding the scala media, including the basilar membrane. This sequence of events is shown schematically in Figure 1.2A, where the coclear duct is shown as an unrolled tube. The cross sectional structure of the cochlear duct is shown in more detail in Figure 1.2B. Because the width and flaccidity of the basilar membrane increases steadily from the base of the cochlea to its apex, different points along the basilar membrane respond preferentially to different frequencies of vibration. This structural variation provides the primary frequency analysis of the peripheral auditory system.

The PAM equivalent to the frequency analysis of the basilar membrane is a filter bank with 20 channels, roughly corresponding to 20 equally-spaced locations along the basilar membrane. Although the basilar membrane exhibits a number of nonlinear effects, it is not known how important those effects

are to the perception of speech. The PAM filter bank is completely linear.

The sensory organ of the auditory system in the organ of Corti, mounted on the medial surface of the basilar membrane. Figure 1.2C shows the structure of this organ, including its three rows of outer hair cells and single row of inner hair cells. These cells transduce mechanical motion into changes in internal electrical potential. Emerging from the upper surface of each of these transducer cells are rows of stiff rod-like cilia. As the basilar membrane rocks in response to acoustic stimuli, each rigidly-attached haircell body drags its cilia through the fluid of the scala media, causing the cilia to pivot about their base. This angular displacement, in turn, causes electrochemical changes within the haircell, which can be measured as changes in the haircell receptor potential.

The PAM contains a very simple model of this mechanical to electrical transduction. The model is a minimal model, in the sense that it simply posits a smooth monotonic saturating nonlinear relationship between input and output, and derives the DC transfer function of such a transduction device when driven by sine waves of various amplitude. Despite the simplicity of the model, it displays at least four important characteristics of haircell rate-level relationships. The nature and behaviour of this transduction model will be

discussed at length in Chapter Two.

Each inner haircell of the organ of Corti is innervated by as many as a dozen or more afferent auditory nerve fibers. The outer haircells are innervated almost exclusively by efferent fibers. Since a particular afferent fiber terminates at the base of only a single haircell on the basilar membrane, each fiber is preferentially stimulated by a single range of frequencies, and the frequency analysis of the basilar membrane is reflected in the firings of the auditory neurons. Each channel of the PAM is intended to be a model of the expected firing rate of an afferent fiber terminating at a haircell located at one of twenty equally spaced locations along the basilar membrane.

Neural firings are generated by the release of neurotransmitters by an inner haircell. The rate of release is a function of the receptor potential in the haircell, and is reflected in the firing rate of the afferent neurons terminating at that haircell. Little is known about the role of outer haircells and efferent neurons in mammalian auditory systems. Art and Fettiplace (1984) have shown that, in the turtle, strong stimulation of efferent fibers at their origins in the central nervous system seems to result in elevated auditory thresholds and a broadening of the frequency response of associated afferent fibers.

Neural firings are random, spike-like electrical discharges, and can be described using the statistics of point events. These neural responses to the quasi-periodic changes in the internal potential of the haircell (caused by acoustic stimulation) can be characterized in two ways: the short term average number of firings for a given level of stimulation, and the temporal distribution of those firings relative to the phase of the stimulation. Post stimulus time (PST) histograms of neural discharges in response to an acoustic stimulus can be used to study the effect of stimulus parameters, such as frequency, intensity, and duration, on both the average rate of firing and the temporal characteristics of the response.

Some of the more obvious characteristics of the average response rate include: a range of spontaneous firing rates (for different neurons) in the absence of stimulation; a correlated variation in the threshold of response as the intensity of stimulation is increased; a limited dynamic range; a saturation in firing rate at high stimulus intensities; and an adaptation of firing rate in response to prolonged stimulation. All of these phenomena are well represented in the PAM, and will be discussed in detail in Chapter Two.

The limited dynamic range of neural firings, and the saturation in the firing rate of many auditory neurons at stimulus intensities below those typically found in speech,

have led many researchers to doubt that spectral information, particularly of vowels, can be adequately encoded in firing rate (see, for instance, Sachs and Young, 1979). An attractive alternative makes use of the synchrony of neural firings to a particular phase of the stimulating acoustic wave. Neurons terminating at locations along the basilar membrane that respond preferentially to frequencies below about 3 KHz show a statistical preference for firing during a particular phase of basilar membrane motion. This preference can ultimately be traced to a non-symmetrical relationship between the orientation of cilia displacement and the resulting change in the haircell receptor potential. This preference is not seen above 3 KHz. possibly because of capacitive low-pass filtering effects in the haircell.

A variety of generalized synchrony measures have been proposed and implemented (Young and Sachs, 1979; Delgutte, 1981, 1984b; Seneff, 1985). Applying these measures to auditory response data shows that the degree of phase synchrony increases as a function of stimulus intensity, beginning below the threshold of increased firing rate, and continuing well beyond the intensity at which expected firing rate has saturated. It has been shown that the spectra of vowels is well represented by such synchrony measures, albeit with significant differences from acoustic spectra. It is clear that models of auditory synchrony are important, and active

research continues.

The PAM does not model auditory synchrony. That aspect of auditory response was omitted from the model, which is primarily intended to model temporal response patterns resulting from stimulation of the PAS by consonant-like sounds. Such sounds often include substantial high frequency energy, typically have less prominent spectral peaks than vowel-like sounds, and are characterized by abrupt changes in spectral shape and amplitude. However, a strategy for expanding the PAM to include neural synchrony, and other methods of increasing the effective dynamic range of the model, are discussed in Chapter Four. These extensions would make the PAM more appropriate for investigating peripheral auditory response to vowels.

1.4 PERIPHERAL AUDITORY RESPONSE AND THE ISSUE OF INVARIANCE AND VARIABILITY IN SPEECH

The relationship between phonetic features and observable properties of the speech signal is a major issue in the study of acoustic phonetics. One theory of speech perception--the theory of acoustic invariance--states that phonetic perception depends on the detection by the auditory system of invariant properties in the acoustic signal of a speech segment (Stevens and Blumstein, 1981), properties which are directly related to the distinctive features which combine to specify the identity

of a word.

However, despite decades of work aimed at articulating a "unified theory" of the acoustics and phonetics of speech, it is still difficult in many cases to explain the way in which the distinctive features which describe the phonetic perception of speech are encoded into measurable properties of the speech waveform: properties observable using acoustic instruments such as the spectrograph. On the contrary, acoustic-phonetic studies have shown that, if such invariant properties do exist, they at least do not correspond to discrete attributes located in particular regions of traditional time-frequency-amplitude representations of the acoustic signal. Acoustic properties such as formant frequency or amplitude, onset rate or duration of bursts, etc., do not represent invariant correlates of phonetic features.

Therefore any invariant correlates must take the form of more complex properties of the acoustic signal, presumably spread over particular time and frequency ranges. In the case of consonants, for instance, the place and manner of articulation affects the entire spectral shape, as well as the evolution of that shape throughout some interval of time. It has become clear that many, and perhaps all, distinctive phonetic contrasts, such as the contrast between voiced and unvoiced consonants, or the contrast between continuant and noncontinuant consonants, are affected by the relative strengths of

several--sometimes many--acoustic properties. In the case of voicing, for instance, over a dozen acoustic properties have been identified which are salient to stop recognition (Edwards, 1981). Much work has been dedicated to cataloging and studying the nature of these "trading relationships."

The transformations of the peripheral auditory system are of particular interest in light of the difficulty of identifying acoustic correlates of phonetic features. The question arises, what are the auditory correlates of phonetic identity in speech? Are there auditory response properties which can be identified with phonetic features? These questions, and investigations into the nature of auditory response to speech, suggest a new point of view on the issue of invariance and variability in speech. This point of view might be called a theory of auditory invariance, and could be summarized in the following steps:

1. invariant phonetic perception develops from variable speech signals as the result of transformations which occur to the signals as they travel through the peripheral and central auditory systems;

2. these transformations result in neural response patterns which correspond to the phonetic pattern of linguistic information in the signal;

3. these "correlating transformations" may well occur at different points in the auditory system for different

phonetic features;

4. the proper question is not so much whether invariance exists, but where in the auditory system the "correlating transformations" are located for each phonetic feature.

To what extent do the transformations of the peripheral auditory system serve as the "correlating transformations" out of which invariant phonetic perception emerges? This thesis represents a first step in answering that question.

FIGURE CAPTIONS FOR CHAPTER 1

FIGURE 1.1

The human external, middle, and inner ear. From Moore (1982) and Lindsey and Norman (1972), by permission.

FIGURE 1.2

The structure of the cochlea. From Pickles (1982), by permission.
A: The cochlea shown as an unrolled tube. Acoustic stimulation results in vibration of the stapes in the oval window. This vibration sets up traveling waves in the basilar membrane and the other structures which surround the scala media.
B: A cross-sectional view of the cochlear duct. The afferent fibers of the auditory nerve terminate at haircells in the organ of Corti, which sits on the basilar membrane.
C: A detailed view of the organ of Corti. Almost all of the afferent fibers of the auditory nerve terminate on inner haircells. The cilia of these haircells project into the scala media, and may touch the tectorial membrane.

FIGURE 1.1

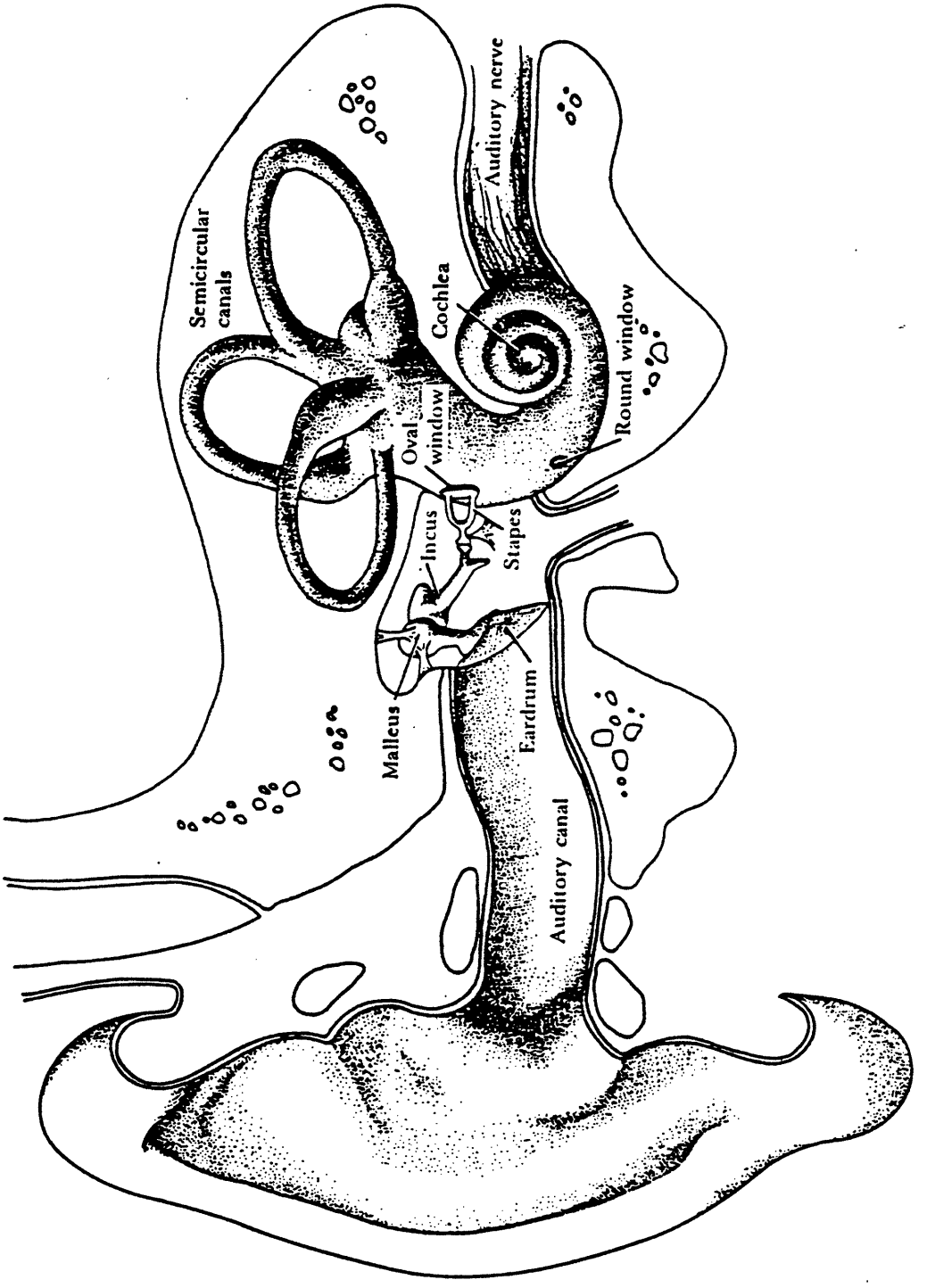
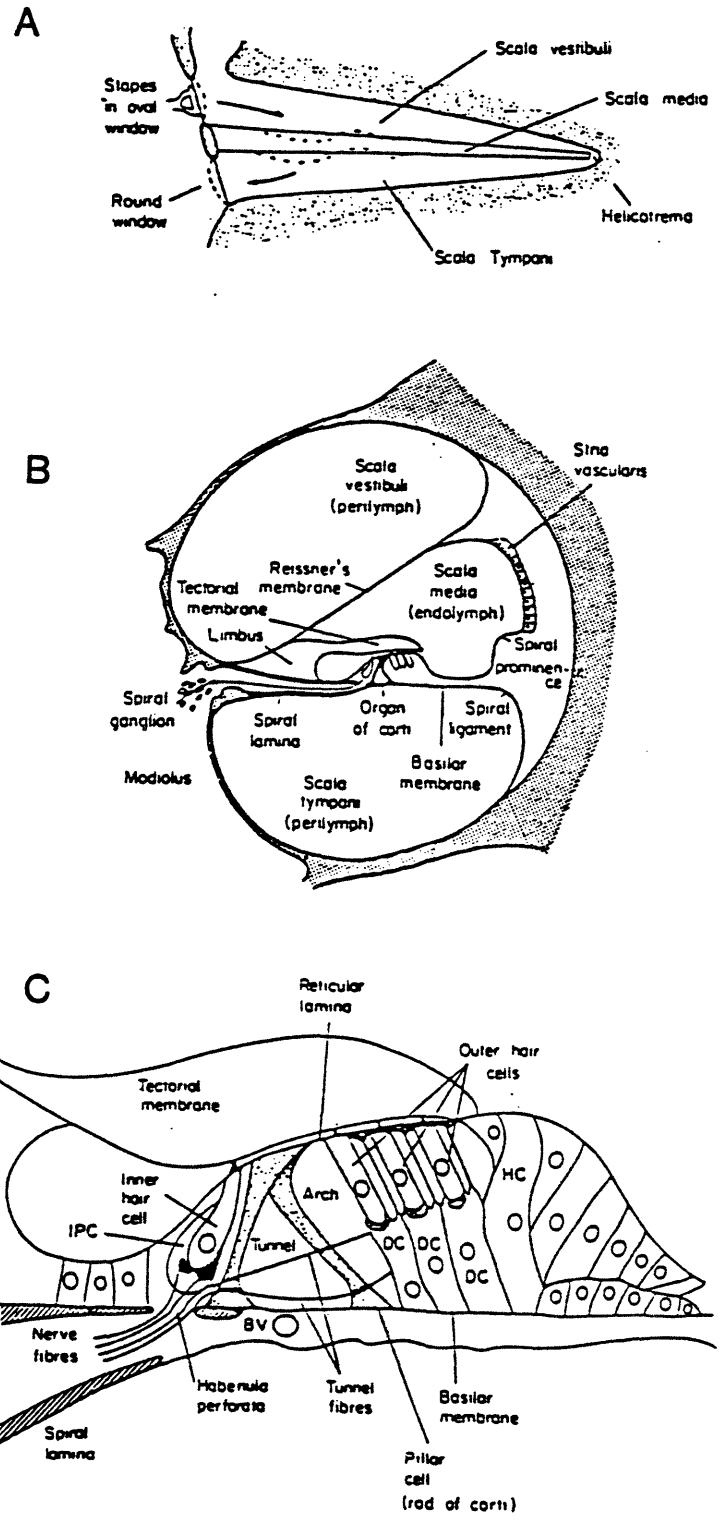

Semicircular canals
Auditory nerve
Cochlea
Round window
Oval window
Incus
Stapes
Malleus
Eardrum
Auditory canal

FIGURE 1.2


A
Scala vestibuli
Scala media
Stapes in oval window
Round window
Helicotrema
Scala Tympani
B
Stria vascularis
Scala vestibuli (perilymph)
Reissner's membrane
Tectorial membrane
Scala media (endolymph)
Limbus
Spiral prominence
Spiral ganglion
Spiral lamina
Organ of corti
Spiral ligament
Basilar membrane
Modiolus
Scala tympani (perilymph)
C
Reticular lamina
Outer hair cells
Tectorial membrane
Inner hair cell
IPC
Arch
Tunnel
DC
HC
Nerve fibres
BV
Habenula perforata
Tunnel fibres
Basilar membrane
Spiral lamina
Pillar cell (rod of corti)

CHAPTER 2

PAM: A PERIPHERAL AUDITORY MODEL

2.1 CONSIDERATIONS IN CONSTRUCTING AND USING THE PERIPHERAL AUDITORY MODEL

This chapter describes a model of the peripheral auditory system. The model, which will be referred to as the PAM, reproduces some of the response characteristics of the mammalian auditory system. The PAM is limited in two ways. First, it does not include a model of the outer and middle ear, nor of the central auditory system. The PAM only models the activity of the inner ear, and the auditory nerve. Second, it focuses on a limited number of phenomena in these parts of the auditory system.

In the experiments described in this thesis the PAM is normally used in conjunction with a preprocessor and a postprocessor that take the place of, and perhaps in a primitive way imitate, the parts of the auditory system that precede and follow the inner ear. A preprocessing filter acts as a frequency shaper to assist in the alignment of speech spectra within the limited amplitude "window" of the PAM. A set of postprocessing filters pick out certain properties of the response patterns of the PAM so that they can be more easily examined. This distinction is shown in Figure 2.1. Whereas the characteristics of the PAM are based on psychoacoustic and

physiological data, the characteristics of the pre- and post-processor are the result of experimental necessity or invention. The preprocessor is a typical speech processing filter, but at least one of the postprocessors is based on an unconventional, and hopefully somewhat provocative, treatment of the PAM output.

2.2 IMPLEMENTATION

The PAM has been implemented as a set of programs in the C language. The programs make use of a general signal processing package called SigPak, and have been implemented under two operating systems, UNIX and VMS. The execution of the PAM programs, and the acquisition of the signals used to test them, was carried out on a variety of general purpose computers, including a PDP-11/60, a VAX-11/750, and an LSI-11/23.

2.3 THE PREPROCESSOR

2.3.1 Signal Acquisition

The speech signals described in this thesis were either digitized from analog tapes or synthesized using the Klatt (1980) speech synthesis program. The digital signals have 14 bit precision sample values, and a sampling frequency of 12,987 Hz. (sample period of 77.000 microseconds). The analog signals were low-pass filtered at 6.4 KHz, using a Rockland

752A programmable elliptic filter with a frequency response that falls off at 115 dB/octave above the cut-off frequency.

2.3.2 Signal Preemphasis

Before being submitted to the PAM, speech signals are normally preprocessed by a high frequency preemphasis filter. The input-output equation of the filter is:

$$y(n) = Gain * (x(n) - x(n-1)/Zero)$$

where x(i) and y(i) are the i'th input and output samples, and Gain and Zero are adjustable parameters. This filter has the system function:

$$H(z) = Gain * (1.0 - 1.0/(Zero*z))$$

The magnitude of the frequency response of this filter is shown in Figure 2.2, for Gain = 0.78 and Zero = 0.80. As can be seen, this filter gives a gentle boost to mid and high frequency components of the input signal. The magnitude of the frequency response at 100 Hz, which is the center frequency of channel 1 in the PAM, is 20 dB less than the magnitude of the frequency response at 6400 Hz, which is the center frequency of channel 20 in the PAM. The use of this preemphasis filter was found necessary to compensate for the 6 dB/octave intensity loss in typical speech signals, combined with the limited

dynamic range of the PAM. (As will be seen, the high frequency channels of the the PAM provide an inherent 6 dB/octave boost to high frequency broadband signals because of the filters' increasingly wide bandwidths. The additional amplification of this preprocessor is required in the mid frequency range in order to keep the higher formants of many voiced speech signals within the dynamic range window of the PAM.)

2.4 THE PERIPHERAL AUDITORY MODEL

The components of the peripheral auditory model are shown in Figure 2.3. The input to the PAM is a preprocessed signal representing the waveform of a speech (or other) sound at the oval window of the cochlea. The output of the PAM is a set of 20 filter-bank-like signal channels. Each channel models the expected firing rate of a representative auditory neuron. The neurons represented by the PAM channels have identical tuning characteristics, thresholds, and spontaneous firing rates. The 20 channels correspond to neurons terminating at 20 equally spaced intervals along the basilar membrane. Like auditory neurons, the 20 channels respond to signal energy within different frequency bands, centered at equal intervals along the Bark frequency scale (Zwicker, 1961). The frequency response of a particular channel roughly corresponds to the frequency response of the auditory system to the channel's center frequency, as derived from psychoacoustic data.

As shown in Figure 2.3, the PAM consists of three stages. Each stage effects a transformation on the speech signal: the spectral analysis stage produces a frequency transformation similar to a cochlear frequency analysis; the transduction stage produces an amplitude transformation that models the ear's conversion of signal intensity to average neural firing rate; and the adaptation stage produces a temporal transformation that mimics neural adaptation to stimulation. The intent in this model is to reproduce a few selected characteristics of the response of the peripheral auditory system, which were considered most important in determining the PAS's response to consonantal sounds. (Chapter Four discusses how this model could be expanded.) The following sections will discuss each stage in detail, and specify the auditory phenomena that are being modeled.

2.4.1 The Spectral Analysis Stage

The spectral analysis stage consists of a linear filter bank, with bandpass filter characteristics that produce a cochlear-like frequency mapping. Twenty filters are used to cover the frequency range of DC to approximately 6500 Hz.

2.4.1.1 Characteristics of the Frequency Transformation

The PAM bandpass filter spacings and frequency responses are derived from psychophysical data. The magnitude of the

frequency response of three adjacent filters are shown in Figure 2.4. The important features of these filters are:

1. Center frequencies occur at approximately integral values of frequency in Bark. This places each filter at approximately one critical bandwidth from its neighbors on both sides. Table 2.1 lists the center frequencies of these filters in Hz and Bark. As can be seen, 1 Bark intervals correspond to approximately equal intervals below 1 KHz, and to 1/6th octave intervals for frequencies above 1 KHz. Note that the number of channels that have been used (20) is not based on auditory data: mammalian cochleas contain several thousand inner hair cells, each with a different characteristic frequency. The number of channels was choosen to minimize computational complexity while maintaining a reasonably smooth composite frequency response between 100 and 6000 Hz.

2. The 3dB bandwidth of each filter is 1 Bark, or one critical bandwidth (Zwicker, 1961). This corresponds to a center frequency to 3 dB bandwidth ratio of 6 for filters with center frequencies above 1 KHz.

3. The tops of the filters are parabolic for gain measured in dB, corresponding to the Gaussian-shaped (for gain measured on a linear scale) masking curves measured by Patterson (1976).

4. The low frequency filter skirts fall off at 10 dB/Bark, and the high frequency skirts fall off at 25 dB/Bark. For frequencies above 1200 Hz or so, this corresponds to 60 dB/octave and 150 dB/octave respectively. For frequencies below 1000 Hz, 30 dB/octave and 60 dB/octave are typical values for low and high frequency skirts. These skirt characteristics are derived from masking data (Zwicker, 1970).

5. The phase response of the filters is linear in Hz, with a group delay which is identical across all channels. This choice of a linear phase characteristic is very close to

that reported by Pfeiffer and Kim (1975), who measured the phase and amplitude response of populations of auditory neurons as a function of their characteristic frequency. Their phase responses come very close to being linear in Hz, with a slope that is a function of characteristic frequency. Equalizing the slopes is equivalent to adjusting the group delay so that the peak response to bursts occurs at the same time in all channels. Figures 2.5 and 2.6 present two different views of the phase and magnitude components of the frequency responses of even-numbered channels.

6. The gain at each filter's center frequency is equal to one. Since the bandwidth increases linearly with center frequency (in Hz), channels with center frequencies above 1 KHz have an effective 6 dB/octave preemphasis for wide band signals. Figure 2.7 shows the magnitude of the frequency response of the individual filters, and Figure 2.8 shows the composite frequency response, both with and without the preemphasis filter.

7. Since the bandwidth of the filters is a function of center frequency (for filters above 1 KHz), high frequency filters provide good temporal resolution at the expense of spectral resolution, while low frequency filters provide good spectral resolution at the expense of temporal resolution. The temporal resolution of the filters can be inferred from Figure 2.9, which shows the magnitude of the impulse response of each filter. As Searle (1979) and Klatt (1979) point out, the auditory variation in temporal and frequency resolution produces conditions favorable for formant tracking at low frequencies and good temporal resolution at high frequencies.

The frequency response of the band pass filters is based on psychophysical masking data rather than physiological data--that is, neural tuning curves such as those presented by Kiang (1965). The psychophysical shapes were used because of

their greater simplicity and uniformity. It is from masking data that the Bark frequency scale is derived. One Bark corresponds to one critical bandwidth of masking noise, beyond which increasing bandwidth no longer serves to increase the effectiveness of a masking signal on a pure-tone test signal. This critical bandwidth is wider at high frequencies than at low, and is generally assumed to correspond to a "natural" frequency scale reflecting the change of characteristic response frequency with position along the basilar membrane.

Because most of the post-processing used in these experiments involves recombining all of the PAM channels, it is unlikely that any of the results reported here depend on the exact shape of the PAM filters. On the other hand, both basilar membrane and auditory fiber tuning curves exhibit a sensitive, sharply-tuned "tip" (Kiang, 1965; Khanna and Leonard, 1982) that could provide a much greater frequency selectivity than psychophysical masking curves indicate. This frequency selectivity might well be important to model in any investigation of the auditory response properties of vowels.

2.4.1.2 Implementation Issues

Throughout the frequency transformation stage, the following conversion formula was used to convert from Hz to Bark:

```
X = H * Ø.Ø1;          for H <= 5ØØ.Ø.
  = 1.5 + H * .ØØ7;    for 5ØØ.Ø < H <= 12ØØ.Ø
```

$$= -32.6 + 6.0 * \log_{10}(H); \qquad \text{for } 1200.0 < H$$

Here X is a particular frequency in Bark, and H is the corresponding frequency in Hz. This formula is continuous for all values of H, and deviates from Zwicker's tabular values (1961) by less than 3% between DC and 10 KHz. Throughout this range its deviation is less than Schroeder's formula (1979) which differs from Zwicker's values by more than 5% for frequencies less than 200 Hz or more than 7 KHz.

Another advantage of the formula is that it provides an explicit center-frequency-to-bandwidth ratio of 6 to 1 at high frequencies for filters, like ours, that have 1 Bark bandwidths.

Each bandpass filter in the filter bank was implemented by convolving the channel's impulse response with the input signal. The impulse response was calculated from the frequency response, whose magnitude and phase characteristics were calculated from the specifications presented above. The calculated impulse response was windowed and smoothed with a 256 point Hamming window before being used for filtering.

The frequency band between DC and 6493.5 Hz (sampling rate of 12987 divided by 2) was spanned using a 256 point DFT. As a result, the phase and magnitude of the frequency response of each of the filters was specified every 50.7 Hz. The center frequencies of the 20 channels were adjusted to fall on

multiples of this frequency.

The magnitude of the frequency response of each PAM bandpass filter can be expressed analytically as a function of distance from the center frequency of the channel. The analytic expression is independent of channel number when frequency is measured in Bark. With frequency in Bark and gain in dB, the magnitude function has a symmetric parabolic top, and straight skirts: the low frequency skirt rises at 10 dB/Bark, and the high frequency skirt falls at 25 dB/Bark. In the PAM, values below -60 dB were approximated as zero. The straight skirts were fitted smoothly to the parabolic center. As a function of DelX, distance in Bark from the center frequency, the filter attenuation in dB was calculated using the following formula:

```
A = -1.92 + 10.0 * (DelX + 0.4);   for DelX <= -0.4
  = -12.0 - 25.0 * (DelX - 1.0);   for 1.0  <= DelX
  = -12.0 * DelX * DelX;           for -0.4 < DelX < 1.0
```

The phase of the frequency response of each PAM bandpass filter can also be expressed analytically as a function of distance from the center frequency of the channel. In contrast with the magnitude function, the phase function is independent of channel number when frequency is represented in Hz. With frequency in Hz, the phase function is linear: as a function of DelF, distance in Hz from the center frequency, the phase is calculated using the following formula:

```
P = -PI * DelF / (SamplingRate / DFTSize)
```

this choice of slope provides a group delay equal to 1/2 the size of the impulse response: 128 samples, or 9.856 msecs. This means that the peak response to an impulse occurs in all channels at a delay of one half the width of the impulse response, as evident in Figure 2.9.

The impulse response of each PAM channel was calculated using the following procedure. First, the value of the frequency response of each channel was calculated at 256 equally spaced points around the unit circle from the magnitude and phase functions. Then a high-pass filter was simulated by setting to zero the DC and 50.7 Hz magnitude components of the frequency response of each channel. A 256-point complex impulse response was then calculated for each channel using an inverse DFT, as follows:

```
I(n) = H(n) * (sum (for i = 0 to 255) of T(n, i))
T(n, i) = F(i) * W(i*n)
```

where I(n) is the n'th sample of the impulse response, H(n) is the n'th coefficient of a 256-point Hamming window, F(i) is the i'th complex DFT coeffecient, and W(k) is the value of a point on the unit circle in the complex plane, at an angle of (-2*pi*k/256) radians. I, T, F, and W are all complex numbers. Figure 2.9 shows the resulting impulse response of the PAM filter bank (not including the high frequency preem-

phasis of the preprocessing stage).

The PAM filter bank itself is implemented by convolving the input signal to the PAM with the impulse response for each channel. Each output sample of each channel is the result of convolving the most recent 256 samples of the input signal with the 256 samples of the channel's impulse response. Since the impulse response sample values are complex, the filter bank output values are complex.

2.4.1.3 Examples of The Filter Bank Response

As part of the test procedure for the PAM spectral analysis stage, a test signal was constructed that consists of the sum of two sine waves of equal amplitude. The frequency of one sine wave is 405.8 Hz, or precisely the center frequency of channel 4. The frequency of the other sine wave is 1674 Hz, or precisely the center frequency of channel 12. The signal is preceded by 50 msecs of silence. The signal's amplitude at the onset increases smoothly over 5 msecs. The amplitude remains constant for 100 msecs, and then decreases smoothly to zero over 5 msecs.

Figure 2.10 shows the magnitude of the response of the PAM filter bank to this signal. Time increases from left to right. Low frequency channels are at the top, and high frequency channels at the bottom. Channel center frequencies are

shown to the nearest Bark on the left, and to the nearest Hz on the right.

Below the output response is shown the input signal, to the same time scale. Since the filter bank has approximately a 10 msec delay, the input signal has been delayed by an equal amount, so that the input and output signals are time aligned. The filter bank clearly resolves the two components of the test signal.

Figure 2.11 shows a single spectral slice of the response shown in Figure 2.10. The ordinate of the top panel is gain in linear units, similar to Figure 2.10. The ordinate of the bottom panel is gain in dB, similar to Figure 2.4, with which it should be compared. This figure verifies the effect of high frequency preemphasis, and the basic shape of the band pass filters. The high frequency tails of the "formants" reflect the low frequency slopes of the bandpass filters, and the low frequency tails reflect the steeper high frequency slopes of those filters.

2.4.2 The Transduction Stage

The transduction stage of the PAM contains the primary nonlinear component in the auditory model. As shown in Figure 2.12A, it consists of an envelope detector followed by a saturating memoryless nonlinearity, which is used to convert

the detected envelope amplitude into the proper output value. This system is analogous to the transduction of basilar membrane motion by hair cells. The input to this stage is the band pass filter channel signals generated by the previous stage. As indicated in Figure 2.3, this transformation is applied independently to each filter bank channel, using an identical detector and nonlinearity for each channel.

2.4.2.1 Characteristics of the Amplitude Transformation

The amplitude characteristics of the transduction stage are motivated by electrophysiological rate-level data relating mean neural firing rate to stimulus intensity. The characteristics of the rate-level data that have been modeled are:

1. Some auditory neurons exhibit spontaneous firing in the absence of input stimulation. The spontaneous firing rate can range from Ø to approximately 25 percent of the adapted saturated firing rate (Liberman, 1978).

2. Auditory neurons exhibit a range of relative thresholds (input amplitudes above which the output exceeds the spontaneous by a certain percent) of 3Ø to 4Ø dB (Liberman, 1978).

3. A strong inverse relationship exists between spontaneous rate and relative threshold (Liberman, 1978). The population of low threshold fibers is generally co-extensive with the population of fibers exhibiting a range of spontaneous rates greater than 1 spike per second, while the population of fibers with a spontaneous rate less than 1 spike per second exhibits a wide range of relative thresholds.

4. In a region between threshold and saturation, auditory fibers respond to exponential increases in the amplitude of stimulation with roughly linear increases in firing rate. The width of this region, which can be considered a measure of the dynamic range of the fiber, is typically 20 to 30 dB, but can be as large as 60 dB for fibers with high thresholds (Liberman, 1978; Evans and Palmer, 1980).

2.4.2.2 Constructing the Transduction Model

In this section we will justify the construction of the transduction stage from an envelope detector followed by a static memoryless nonlinearity, and we will show how the nonlinear function itself can be constructed in a well-motivated way. To do this, we will raise six considerations, drawing a conclusion from each consideration in turn. These six conclusion explain and justify our transduction model.

Consider the specifications, listed above, of the desired characteristics of the transduction stage. Each of them specifies something about the steady state DC characteristics of the auditory system. None of them specify any phase relationships between the acoustic stimulus and the resulting firing rate. (As mentioned in the Introduction, the PAM was not designed to model auditory synchrony.)

Conclusion: The specifications above can be satisfied by correctly establishing the steady-state DC characteristics of the PAM.

Consider the use of the transduced output signals. As shown in Figure 2.3, the output of the transduction stage is further processed by the adaptation stage. As we will see, the specifications for this later stage call for the degree of adaptation to depend only on the long-term average firing rate of the channel in response to sustained stimulation. The fastest time constant that will be involved is 15 msecs, implying that only low frequency components of the transduction stage output will affect adaptation. We will also see that high frequency components of the output will be passed through the adaptation stage essentially unaffected.

Furthermore, the output of the PAM is defined to be expected firing rate. Although this term has not be precisely defined, it can be taken to mean the average firing rate over a period of at least one or two msecs. In fact, the post processors that are used smooth the output signal even more.

Conclusion: The following stage of the PAM, and the PAM postprocessors, depend only upon the low frequency components of the output of the transduction stage. Taken in conjunction with the first conclusion, this means that the transduction stage may include a low frequency filter that removes signal components above 100 or 200 Hz.

Consider the characteristics of the input signal to the transduction stage. For each channel, this signal is the out-

put of a PAM band pass filter. All of these filters have a three-dB-bandwidth of one Bark. For all but the lowest channels, the filter bandwidth is considerably smaller than the filter center frequency. Therefore, as an approximation, the output of each PAM filter channel can be represented as a periodic signal at the channel center frequency, modulated by a slowly varying envelope signal:

$$O(t) = A(t) * \exp(i*pi*f*t)$$

where O(t) is the complex-valued output signal of the channel; A(t) is the complex-valued, slowly varying, envelope signal; and f is the center frequency of the channel in Hz. It is important to realize that the frequency components of A(t) are limited to one half the bandwidth of the channel's band pass filter. Thus for low frequency channels A(t) only has frequency components below 50 Hz, while for high frequency channels A(t) may contain components as high as 500 Hz.

Conclusion: The input signal to the transduction stage for each channel can be approximated by a slowly modulated sine wave.

Consider the configuration for an idealized transduction system shown in Figure 2.12B. Note that this is not the PAM transduction stage shown in Figure 2.12A. Instead, it is an even simplier transduction model whose properties we would like to investigate. If we can show that the DC

characteristics of this idealized model has desirable properties, and meets all of our specifications, the model's extreme simplicity justifies its use--at least until we add some more specifications that it can not meet! What we will attempt to do is to show that under the input and output conditions described above, the PAM model in panel A is equivalent to the idealized model in panel B.

This idealized model consists only of a static memoryless saturating nonlinearity. Adding the constant bias B to the input signal is a way of collapsing a family of nonlinear functions B + T(i) into a single function T(i). Consider the output that will result from applying the transduction model to a sinusoidal input signal A*sin(f*t). Because the idealized model involves a memoryless nonlinearity, the output signal will contain components at DC, at the fundamental frequency of the input, and at higher harmonics. For such an input signal, the output characteristics of the model are completely specified by a set of transfer functions that relate the amplitude of each component to the input amplitude A. There is a DC transfer function, and a set of AC transfer functions, one for each harmonic. The shape of these functions do not depend on the frequency of the input signal: only on its amplitude A.

Conclusion: For the input signals we are dealing with in the PAM, the DC characteristics of the response of the

idealized transduction model depend only on the magnitude of A(t), and the bias B.

Consider the system shown in Figure 2.12C. This configuration is a method for determining the DC transfer functions of the idealized transduction model, as a function of the input amplitude A and the constant offset bias B. The box labeled "AVERAGE" measures the average output of the model over one period of the input sine wave. Figure 2.13 shows a number of different transduction functions that might be used in this idealized model. Figure 2.14 shows the DC transfer functions that result from using the raised hyperbolic tangent function. Each line in Figure 2.14 corresponds to the use of a different bias B. The lines are labeled with the values of B that were used. We will consider the characteristics of these transfer functions in more detail below.

Figure 2.15 shows the DC transfer functions that result from using each of the transduction functions of Figure 2.13. Although specific details of the functions change from panel to panel, the broad characteristics of the functions, such as their dynamic range, their saturation value, and their relative threshold, are very insensitive to the shape of the underlying transduction function from which they were generated. The one exception to this is that none of the DC transfer functions generated from the step function have non-zero lower asymptotes. The reason for this will be discussed

in a moment.

Conclusion: The specific shape of the underlying transduction function does not have a strong effect on the shape of the resulting DC transfer function.

Consider again the implementation of the PAM transduction stage shown in Figure 2.12A. Assume that the nonlinear function used in the model is a DC transfer function generated from the test setup shown in Panel C of that figure: for instance, one of the curves shown in Figure 2.14. The envelope detector extracts the short term channel signal amplitude from the filter bank output. This amplitude is then transduced using the appropriate DC transfer function. Thus the PAM transduction stage synthesizes the low frequency component of the idealized model's output signal from the input signal's amplitude and the appropriate DC transfer function.

Conclusion: The PAM transduction model, when used with a calculated DC transfer function, will generate a slowly varying output signal that is the DC component of the output of the idealized transduction model.

2.4.2.3 Analysis of the Transduction Model

In the previous section the underlying nonlinearity used in the idealized model was referred to as the "transduction function". The nonlinearity used in the PAM transduction

stage is the DC transfer function of the idealized model, and is referred to either as the "transfer function" or as the "amplitude nonlinearity". These designations will be maintained in this and following sections. It should be emphasized that the only use that is made of the idealized model shown in Figure 2.12B is to generate transfer functions for use in the PAM transduction model. Only the PAM stage, shown in Figure 2.12A, is used to process speech.

The transfer functions generated by the idealized model are the PAM equivalent of physiological rate-level curves. In a moment we will compare the characteristics of our functions with those curves, but first it is worth while examining more closely the dependence of the shape of the DC transfer functions on the characteristics of the underlying transduction functions that are used to generate them.

Looking again at the examples of potential transduction functions shown in Figure 2.13, we can see that all of the functions have a negative asymptote of zero, and a positive asymptote of two. All of the functions have an output of one for an input value of zero. Each function also has a finite "transition region" around zero, a region within which the function varies from a small amount (say Ø.1) above the negative asymptote to the same amount below the positive asymptote.

The value of the negative and positive asymptotes were selected to provide "normalized" transfer functions. The negative asymptote of zero ensures that resulting transfer functions will themselves have a value of zero for sufficiently small input signal amplitudes and large negative bias values B (see Figure 2.14, B = -16, input amplitudes less than 2Ø dB). The positive asymptote of two ensures that the "saturation value" of the resulting transfer function (output level for large input amplitude) will be one (see Figure 2.14, any curve, input amplitudes greater than 8Ø dB).

Changing the value of the negative asymptote of the transduction function simply adds an offset to the negative asymptote of the resulting transfer functions, while changing the value of the positive asymptote adds an offset to the saturation value of the transfer functions. However, neither of these changes effect the dynamic range of the transfer functions: only the vertical scale and offset of those functions.

Similarly, changing the horizontal scale of the transduction function (the width of the transition region, and hence the dynamic range of the transduction function), only changes the horizontal offset (the relative threshold) of the generated transfer functions, and not their dynamic range.

The only two remaining aspects of the transduction functions that might effect the shapes of the resulting transfer functions are the shape of the nonlinearity, and the DC offset (the operating bias B) of the input sine wave. In the idealized model, only bias values less than or equal to zero yield meaningful transfer functions. The effect of varying the operating bias is shown in Figure 2.14, using a raised hyperbolic tangent function as the transduction nonlinearity. The effect of varying the shape of the transduction nonlinearity is shown in Figure 2.15. Each curve in these figures is the amplitude nonlinearity resulting from a given transduction nonlinearity and a given operating bias B.

Figure 2.14 shows that the characteristics of the DC transfer functions have much in common with observed rate-level curves. First of all, for operating bias values that fall outside of the transition region, doubling the bias increases the relative threshold of the resulting transfer function by 6 dB (observe, for example, the 12 dB change in threshold between the B = -16, B = -64, and following curves). This makes sense, because doubling the bias means that the input amplitude must be twice as large before the most positive input values reach the left side of the transition region, and are transduced into output values significantly greater than zero.

Once the input amplitude reaches the transition region, it continues to grow until, for input amplitudes much larger than those needed to reach the right side of the transition region, the positive half of the input cycle is transduced into values close to two, while the negative half of the input cycle is transducted into values close to zero. Thus for input amplitudes much larger than the width of the transition region, any transduction function looks like a step function. The average output value of one input cycle is the saturation value of the transfer function, which approaches one.

For operating bias values within the transition region of the transduction nonlinearity, the transduced DC output value for a sine wave of zero amplitude is greater than zero. This "quiescent value" of the DC transfer function corresponds to a nonzero spontaneous firing rate in auditory neurons. The quiescent value of the transfer function is precisely the value of the transduction function at the operating bias B. This quiescent value is the sole source of the final non-zero spontaneous rate in the output of the PAM. The later adaptation stage transforms the value of this quiescent level, but does not, in itself, give rise to any spontaneous output. The presence of a transition region ensures that there are values of B which result in transfer functions with non zero quiescent values. (In Figures 2.13 and 2.15, only the step function does not have a transition region.)

Note that all of the transfer functions with nonzero quiescent values have approximately the same relative threshold, while all of the transfer functions with higher relative thresholds have a quiescent value equal or very close to zero. This satisfies item three in our original list of desirable characteristics.

Thus changing the operating bias B has a distinct effect on the shape of the resulting transfer function. In fact, varying B from zero to much less than the half-width of the transition region causes the relative threshold and quiescent value of the resulting transfer function to vary in the same manner that relative thresholds and spontaneous rates of auditory fibers are seen to vary.

The dynamic range of a transfer function can be defined as the range of input amplitudes for which the transfer function varies between, say, 10 and 90 percent of the difference between its quiescent and saturation values. The dynamic range of each of the transfer functions shown in Figure 2.11 is around 20 to 30 dB.

The one remaining characteristic of the transduction function that might be varied is its shape. Experimentation with a variety of possible transduction nonlinearities indicates that the basic characteristics of the transfer functions shown in Figure 2.15 are relatively insensitive to the

specific nonlinearity chosen. Changing the transduction nonlinearity only has an effect on the shape of the resulting transfer functions around threshold. The shape of the transfer functions around their saturation level is independent of the transduction function shape, because the input amplitude is so high that any transduction function with a transition region of the size shown in Figure 2.13 essentially acts as a step function.

2.4.2.4 Selecting Values for the Transduction Parameters

We have shown that the only characteristics of the idealized transduction model that significantly affect the shape of the resulting transfer function is the presence of a transition region in the transduction function, and the value of the offset bias B. The raised hyperbolic transfer function shown in Figure 2.13 was selected as the standard transduction function in the PAM. This leaves the question of how many bias values B to select, and what their values (and thus the relative thresholds and quiescent values of the resulting transfer functions) should be. Delgutte (1982) has suggested that variations in relative threshold can be employed in auditory models as a principled way to extend the apparent dynamic range of the model, by assuming that different auditory neurons have different operating points, and modeling the output of an ensemble of neurons with a weighted sum of three curves

with different thresholds and dynamic ranges. For the sake of simplicity, however, the PAM was implemented using only a single transfer function, modeling a single auditory rate-level curve. A value of -1.1 was selected as the standard PAM bias value. This bias results in a quiescent response of 20 percent of the saturation value, and a dynamic range (as measured between 10 percent and 90 percent of maximum) of 22 dB. The resulting amplitude nonlinearity is shown in Figure 2.14.

2.4.2.5 Implementation Issues

The DC transfer functions generated by the raised hyperbolic tangent function discussed above were calculated for a set of bias values and stored in a data file. The value of each transfer function was calculated at one dB steps of input signal amplitude, between -20 and 100 dB (re 1.0).

When the PAM is processing signals, the amplitude nonlinearity function for the bias value chosen (typically -1.1) is read from the data file and stored in an array. For each output channel of the filter bank, the magnitude of the channel response is calculated and converted to dB (re 1.0). The resulting amplitude value is then used as an index into the amplitude nonlinearity array. The linearly interpolated value of the transfer function at that amplitude becomes the output, for that channel, of the transduction stage of the PAM.

The input signal is down-sampled by a facter of six in the transduction stage. The resulting sampling rate is 2164.5 samples/second, or .462 msecs/sample. This rate is the Nyquist rate for the envelope signal of the highest channel, which has a pass band that is 1082 Hz wide at the -3dB level. Since all of the lower channels have narrower pass bands, they continue to be oversampled at this rate.

2.4.2.6 Examples of the Transduction Stage Response

Figure 2.16 shows the response of the combined spectral analysis and transduction stage to the test signals described earlier. The most obvious feature of the response pattern is the significant quiescent output (for zero amplitude input). Note that the transduction stage effectively converts the channel amplitudes to a "log magnitude" scale. As a result, the apparant width of the two "formants" is greater in Figure 2.16 than in Figure 2.10, and the onsets and offsets appear faster.

Figure 2.17 compares a spectral slice of the output magnitude of the PAM filter bank to the output of the transduction stage. In the top panel the output of the filter bank has been converted to dB. In the bottom panel the output of the transduction stage is shown in linear units. The limited dynamic range of the transduction stage is clearly evident.

2.4.3 The Adaptation Stage

The adaptation stage of the PAM acts as a constant-gain amplifier which subtracts a time-varying DC offset from the input signal. The amount of offset depends on the average input level in the recent past. Negative output levels are clipped to zero, since the output of this stage represents the expected firing rate of auditory neurons. As indicated in Figure 2.3, a separate adaptation amplifier is associated with each channel of the PAM, and it is the amplitude of stimulation in each channel that determines the instantaneous offset of that channel's amplifier. However, the inherent characteristics of all of the channel amplifiers are identical.

2.4.3.1 Characteristics of the Temporal Transformation

The temporal transformation models the decay of auditory neural firing rates over time in response to sustained acoustic stimulation. The characteristics of adaptation that have been modeled are:

1. The maximum firing rate for an abrupt amplitude onset is approximately 2.5 to 4 times the adapted maximum, which is around 200 spikes/second (Smith and Zwislocki, 1975; Delgutte, 1980).

2. A range of adaptation time constants can be observed in a single neuron. These vary from so called "rapid" adaptation, with time constants in the 5 to 15 msec range (Delgutte, 1980), to short term adaptation, with time constants around 40 ms. (Smith and Zwislocki, 1975;

Delgutte, 1980), to slower adaptations, often referred to as fatigue or long term adaption (Kiang et al., 1965; Young and Sachs, 1973; Abbas and Gorga, 1981) with adaptation times in the hundreds of msecs up to many seconds or even minutes.

3. A step increase in signal intensity results in a step increase in firing rate, and the magnitude of the response increment is independent of the degree of adaptation (Smith and Zwislocki, 1975).

4. Following the abrupt offset of an intense signal of sufficient duration to cause adaptation, spontaneous activity is initially absent, and then increases until it reaches its nominal level. The time constant of this recovery is somewhat longer than the time constant of adaptation (Harris and Dallos, 1979), and although different animals exhibit variations in both adaptation and recovery time constants, the ratio of recovery to adaptation time may be fairly constant across species (Harris and Dallos, 1979).

5. A signal of sufficient intensity and duration to cause adaptation can mask another signal which follows the offset of the first signal within a period of time on the order of the recovery time of the haircell. This masking takes the form of a lowered firing rate in response to the second signal, or even a complete lack of response to that signal.

6. For a level signal with an abrupt onset, the ratio of onset firing rate to adapted firing rate is constant (after subtracting spontaneous rate from both) and independent of input level (Smith and Zwislocki, 1975; Smith, 1979).

2.4.3.2 The single time-constant model

A circuit that exhibits most of the desired properties is shown in Figure 2.18. This is a single time constant adaptation model (STCAM). The independent (input) variable is the voltage V_{in} and the dependent (output) variable is the current I_{out} through the diode. The input to this circuit is the output of the transduction stage of one channel of the PAM, and the output of this circuit is the final output for that channel of the PAM. Except for the diode, the circuit is a linear time invariant high pass filter.

2.4.3.2.1 Analysis of the STCAM

The incremental resistance of the circuit to a step increase in V_{in} is the parallel resistance $R_a \,||\, R_s$. However, as capacitor C_a charges up, the incremental current through the diode falls off to its steady state level with an "adaptation" time constant:

$$Ta = C_a * R_a$$

The steady state value of Vc, the voltage across C_a, is equal to V_{in}: all of the current flows through R_s and the diode, and I_{out} is equal to V_{in} / R_s. Note that the steady state incremental current is smaller than the peak transient incremental current by an amount

$$f_a = R_a / (R_a + R_s)$$

The constant f_a, which will be referred to as the adaptation ratio, is an important ratio which occurs repeatedly throughout the analysis of this model and its successor, the multiple time constant adaptation model.

If V_{in} is suddenly reduced, V_{out}, the voltage across the diode, is reduced. If V_{in} is reduced so that

$$V_{in} < Vc * R_a / (R_a + R_s) = Vc * f_a$$

V_{out} will become positive. In that case the diode no longer conducts, I_{out} becomes zero, and the capacitor begins to discharge through R_s and R_a in series. The time constant of this "recovery" is:

$$Tr = C_a * (R_s + R_a)$$

This recovery time constant is longer than the adaptation time constant, which correctly models the physiological data described in item four above. The recovery time constant is related to the adaptation time constant by the adaptation ratio:

$$Ta = Tr * f_a$$

It would be interesting to see if this predicted relationship really exists between the incremental and steady state firing rates in auditory neurons, and the adaptation and recovery

times of those same neurons.

The diode remains turned off, and the output from the model is zero, until the capacitor has discharged sufficiently that

$$Vc < V_{in} / f_a$$

at which point V_{out} becomes negative, the diode begins to conduct, and Vc completes its discharge through R_a alone.

2.4.3.2.2 Selecting Values for the Circuit Elements in the STCAM

The values of C_a, R_a, and R_s are completely specified by the desired values for the adaptation time constant, the adaptation ratio, and the overall gain of this stage of the PAM. Based on physiological data, an appropriate value for the adaptation time constant is 40 msecs, and 0.3 for the adaptation ratio. The overall gain is arbitrary, and is set to unity in the PAM. These values model short term adaptation only, and leave the response in normalized units, such that a channel output value of 1.0 represents the maximum adapted expected firing rate of the neurons modeled by that channel.

2.4.3.2.3 Response of the STCAM to Simple Signals

Figure 2.19 shows the response of the single time constant adaptation circuit to an input stimulus whose envelope

consists of a quiescent level followed by a step onset followed by a step offset back to the original quiescent level. Panel A shows the input signal, which in the PAM would be the output of the transduction stage, while Panel B shows the output of the adaptation model itself. (The computer model of the adaptation circuit has been initialized with the capacitor charged to the full voltage it will achieve in response to the quiescent input level. Thus the response of the circuit to that level is its steady state response.) When the input increases in a step-wise manner, the circuit responds with a sharp transient which falls off exponentially to a steady state level higher than the original level. When the input returns to its quiescent level, the response falls to zero for an interval during which time the capacitor is discharging. We shall refer to the state of zero output as the "masked state", because during this interval the input is masked by the relatively high input that preceded it. Following masking, the output current builds back up to its pre-stimulus level.

The dotted line in Panel B shows the value that I_{out} would have at each point in time if the diode were to be short circuited for an instant. In other words, when the diode is conducting, the current through it is equal to

$$I_{out} = (V_{in}/R_s) + (V_{in}-Vc)/R_a$$

The dotted line shows this value even during that time when the diode is not conducting. As we will see, this value can be thought of as a hypothetical baseline quiescent response on top of which incremental responses to stimuli are built.

Figure 2.20 demonstrates an important property of the adaptation circuit: its constant incremental gain. A series of eight input signals is shown. Each signal consists of an initial step onset followed by a secondary step increase. The interval between the initial onset and the secondary increase ranges from 0 to 280 msecs. The intensity of the primary and secondary steps are adjusted so that their amplitudes following the transduction stage are equal. Panel A shows the response of the transduction stage, and Panel B shows the response of the adaptation stage. As can be seen, the size of the incremental response of the adaptation stage to the secondary step is the same regardless of the time delay between the primary and secondary step. This demonstrates that the adaptation circuit is not an automatic gain control circuit: if it were, the effect of the primary step would be to reduce the response to the secondary step, and the circuit would respond differently to a stimulus with the two steps occurring together, than to a stimulus with the second step occurring after the circuit had adapted to the first step.

Figure 2.21 is intended to elucidate the sources of nonlinearity in the PAM. The PAM responses to eight signals are

shown. Panel A shows the input to the transduction stage in dB, panel B shows the output of that stage, and panel C shows the output of the adaptation stage. All eight signals have an initial intensity of -30 dB. At 50 msecs there is a step increase to -6 through 30 dB in 6 dB increments in seven of the signals. One of the signals continues at -30 dB. By comparing Panels B and C, it can be seen that the nonlinearity in the onset transient is due solely to the saturation of the transduction stage. On the other hand, the masking period which follows the offset of the step increase is solely due to the adaptation stage. Note that the more intense the preceding stimulus, the longer the duration of the masking period.

Figure 2.22 demonstrates the masking effect in the response of the adaptation circuit. The input is a 150 msec intense stimulus, followed by a series of 1 msec probe stimuli. The output of the transduction stage is shown in Panel A, and the output of the adaptation stage in Panel B. The dotted line in Panel B shows the hypothetical baseline response level as it builds back up to zero. Following the offset of the masking stimulus, the adaptation circuit's response falls to zero for a period of time, and then builds back up to its initial quiescent level. During this time, its response to the probe stimuli is greatly reduced. In fact, there is no response at all to the probe signal occuring 10 msecs after the offset of the masking signal. Only after 250

msecs of recovery time does the circuit response with full intensity to the probe signal. It can be seen that the reduced response is due to the negative level of the hypothetical baseline quiescent response, upon which all incremental responses are built.

Figure 2.23 shows the effect of masker amplitude on the degree of masking of following probe signals. As expected, higher amplitude signals cause greater adaptation, which in turn results in smaller responses to probe signals that occur within the recovery period of the adaptation circuit.

2.4.3.3 The Multiple Time Constant Adaptation Model

Although the single time constant model meets many of the criteria set forth for the adaptation stage, it does not exhibit the variety of adaptation time constants that can be seen in auditory electophysiological data. If the STCAM's time constant is set around 40 msecs it can be used as a model of short term adaptation, but then it does not model either rapid or long term adaptation. Obviously, a model with multiple adaptation time constants is required.

Presumably the way to do this is to add additional RC pairs to the circuit in Figure 2.18. There are many ways of adding these additional circuit elements. We have made use of the following two conjectures from a survey of electropysio-

logical data in choosing our method:

1. Before and during adaptation, all RC pairs appear to act in parallel: the response rate associated with each time constant appears to be added together. To put it another way, as the transient period for rapid, short, and long term adaptation is in turn exceeded, a certain portion of the response rate--associated with those adapted portions--is subtracted from the overall response, until at last only the steady state portion of the response remains. In circuit terminology, the resistors of the RC pairs appear to be in parallel.

2. During periods of recovery, all RC pairs appear to act in series: the shape of the response deficit appears to be the sum of the deficits associated with each of the time constants. As each of the adaptation portions recover, the response rate is boosted toward the spontaneous rate. In circuit terminology, the capacitors of the RC pairs appear to be in series.

Based on these two hypotheses, we have augmented our single time constant model as shown in Figure 2.24. The voltage source and diode remain unchanged. There is still a shunt resistor R_s through which the steady state current I_s flows, generating the nonzero steady state response of the circuit. Although three RC element pairs are shown, the model can be implemented with as many as desired, to model any number of adaptation time constants. The element values must be chosen such that

$$Ta_{\emptyset} = R_{\emptyset} * C_{\emptyset}$$

is the fastest adaptation time constant,

$$Ta_1 = R_1 * C_1$$

is the second fastest, and so forth.

2.4.3.3.1 Analysis of the MTCAM

Consider a version of the MTCAM with n+1 adaptation time constants. Let R_n and C_n be the resistor and capacitor that produce the longest time constant, and R_i and C_i be the elements which produce the i'th time constant. Similarly, let I_i and Vc_i be the current through R_i and the voltage across C_i, and I_s be the current through R_s. Then the operation of this model can be understood by considering its incremental response to a step increase in V_{in}. The incremental response is determined by replacing the capacitors with short circuits. Then the incremental current flow through the diode equals the increment in V_{in} divided by $R_\emptyset$ through R_n, and R_s, in parallel. Thus the relative contributions of each time constant, and f_a, the relative size of the steady state response to the transient response, is determined solely by the relative values of the resistors in the circuit. In all of the cases we shall discuss, the transient associated with each time constant contributes equally to the overall response. Consequently all of the adaptation resistors have the same value.

As current flows through the adaptation resistors in response to the increment in V_{in}, the adaptation capacitors

charge up, with the voltage across $C_{\emptyset}$ increasing fastest, then C_1, etc. The capacitors charge up until the voltage across them equals the voltage drop across the next highest adaptation resistor. (e.g., $C_{\emptyset}$ charges up until $Vc_{\emptyset}$, the voltage drop across it, equals the voltage drop across R_1.) When that happens, current stops flowing through that RC element pair, except for a small negative current as the voltage across the next highest resistor begins to drop in response to its capacitor charging up. Eventually, Vc_n, the voltage across the highest capacitor, equals V_{in}; the voltages across all of the other capacitors equal zero; none of the adaptation resistors are carrying current; R_s is carrying all of the current; and V_{out}, which equals $Vc_n - V_{in}$ when the adaptation resistors are not conducting, is zero.

If now V_{in} suddenly drops (becomes smaller), Vc_n creates a voltage drop across the adaptation resistors. If the decrease in V_{in} is large enough, V_{out} can be driven above zero. This occurs when

$$V_{in} < Vc_n * f_a$$

where $$f_a = R_a / (R_a + R_s)$$

R_a is the parallel combination of all of the adaptation resistors:

$$R_a = R_{\emptyset} \,||\, R_1 \,||\, \ldots \,||\, R_n$$

When V_{in} falls below this threshold, V_{out} becomes positive, and the diode stops conducting. As long as the diode is not conducting, C_n will discharge itself through the series combination of R_s and R_a.

As before, each of the capacitors will charge up in its turn, until its voltage equals the voltage across the adaptation resistor above it in the ladder. Then it will discharge slowly as that RC pair charges up. During recovery, however, the time constant of charging is longer than during adaptation, because the resistance through which C_i is charged is the adaptation resistor R_i plus R_s in series. Therefore the recovery time constant Tr_i will be longer than the corresponding adaptation time constants Ta_i:

$$Tr_i = C_i * (R_i + R_s) > Ta_i = C_i * R_i$$

To demonstrate these relationships, consider the two time constant circuit shown in Figure 2.25. Figure 2.26 shows the way this circuit responses to a signal with a step increase in intensity followed some time later by a step decrease. Such a signal is shown in Panel A of Figure 2.26. The model is initialized so that $Vc_1 = V_{in}$, and $Vc_\emptyset = \emptyset$, at $t = \emptyset$. Thus during the first 50 msecs, the circuit is in steady state, all of I_{out} flows through R_s, and $I_{out} = I_s$ (Panel B).

When the step increase in V_{in} occurs (at t = 50 msecs), I_s continues to be proportional to V_{in}. Indeed, as long as the diode is conducting, it must be. However, current now begins to flow through the adaptation resistors, resulting in the current spikes visible in Panels B, C and D. Since $C_\emptyset$ is relatively small, $I_\emptyset$ quickly falls to zero as $Vc_\emptyset$ rises to match the voltage drop across R_1. When this state is reached (at about t = 60 msecs), $Vc_\emptyset$ begins to fall again, tracking the decreasing level of I_1 (Panel D). Note that $I_\emptyset$ goes slightly negative at this stage. Vc_1 continues to rise until it equals V_{in}. At this point the circuit is almost fully adapted, and almost all of the current is flowing through R_s. When V_{in} drops back to its initial level (at t = 200 msecs), several things happen. V_{out} becomes positive, causing the diode to stop conducting. C_1 now starts discharging through the series combination of R_s and R_a, the parallel combination of $R_\emptyset$ and R_1. Once again, $Vc_\emptyset$ quickly charges to the voltage drop across R_1, although it takes about twice as long (20 msecs) as during adaptation, because $R_\emptyset$ and R_s are now in series.

After t = 220 msecs, a negligible amount of current flows through $R_\emptyset$, and C_1 must continue discharging through R_1 and R_s in series. When Vc_1 has dropped sufficiently that V_{out} is equal to zero, the diode begins to conduct again, and I_s once more is proportional to V_{in}.

Note that at this point Vc_1 is still larger than V_{in}. C_1 will continue to discharge until Vc_1 equals V_{in}, but at a slightly faster rate, because the diode effectively shorts out R_s as far as the discharge current is concerned.

2.4.3.3.2 Selecting Values for the Circuit Elements in the MTCAM

The desired steady state gain between the input and output of the multiple time constant adaptation circuit determines the value of R_s:

$$R_s = V_{in} / I_{out}$$

For use in the PAM, the overall gain has been set to one, so that the implicit dimensions of the output values of the PAM channels are "maximum steady state expected firing rate". For example, a value of Ø.5 indicates an expected firing rate for that channel that is one half of the maximum possible steady state firing rate.

During transient responses to onsets, expected firing rates can go as high as 2 or 3 times the maximum steady state rate. The desired value of the adaptation ratio f_a establishes the size of the incremental peak transient response to a step increase in V_{in}, relative to the incremental steady state response to that same increase. This ratio determines the ratio between R_s and R_a, the parallel resistance of all of

the adaptation resistors:

$$R_a = (f_a / (1-f_a)) * R_s$$

The adaptation ratio is normally set to be Ø.3 for the PAM.

The values of the adaptation resistors in the circuit are determined by the desired partial contribution f_i of the i'th RC pair to the difference between the size of the transient response to an onset and the steady state response to the same onset:

$$R_i = R_a / f_i$$
$$f_i = I_i / I_a$$

where I_i is the incremental current through R_i, and I_a the incremental current through R_a in response to a step onset.

In the PAM, we normally choose to use equal contributions from each RC pair, so that $f_i = 1/3$ for the standard three time constant model. This gives us, finally:

$$R_s = 1.Ø$$
$$R_i = R_s * (f_a / (1-f_a)) / f_i$$
$$= 1.Ø * (Ø.3 / (1 - Ø.3)) / (1/3)$$
$$= 1.286$$

Once the values of the resistors in the adaptation circuit have been chosen, the values of the capacitors are deter-

mined by the desired adaptation time constants. If the i'th desired time constant is Ta_i, the correct value for the i'th capacitor i is:

$$C_i = Ta_i / R_i$$

For most of the use of the PAM discussed in this study, three ranges of adaptation time constants were used: 5 to 15 msecs for modeling rapid adaptation; 40 to 60 msecs for modeling short term adaptation; and 200 to 500 msecs for modeling long term adaptation. These values are within the appropriate ranges for modeling adaptation in auditory neurons (Kiang et al., 1965; Young and Sachs, 1973; Smith and Zwizlocki, 1975; Smith, 1979; Delgutte, 1980; Abbas and Gorga, 1981).

2.4.3.3.3 Response of the MTCAM Circuit to Simple Signals

Figure 2.27 shows the response of the multiple time constant adaptation circuit to signals with ramp onsets of various slopes. In this figure three signals, and the responses of the transduction and adaptation stages to them, are overlaid. Each signal has a ramp onset from an initial value of -24 dB to +24 dB. The durations of the ramps are 5, 20, and 60 msecs for Signals 1, 2, and 3 respectively. 150 msecs after the beginning of the onset, each signal abruptly drops back to -24 dB. A second repetition of each signal begins after a 5 msec gap. After the second offset of each signal

there is another gap, this time of 30 msecs duration, followed by a third repetition of the signals. The following gap is 100 msecs long, and precedes a final repetition of each signal. Panel A shows the input to the transduction stage; Panel B the output of that stage; and Panel C the output of the adaptation stage.

The model responds differently to the different onset rates, as can be seen in Panel C. Signal 1, with the fastest onset, produces the strongest response from the adaptation stage in Repetition 1, before any adaptation takes place. The slower onsets produce significantly lower peak responses, and the peaks occur later in time. However, the slopes of the onset at the output of the adaptation stage show much less variation than the slopes of the inputs: all three onsets are very steep, approximately parallel, and step-like. This is due both to the limited dynamic range of the transduction stage, and the high-pass filter characteristics of the adaptation stage.

The responses to Repetition 2 are very different from the responses to Repetition 1. The very short gap between the two repetitions leaves the adaptation circuit strongly adapted. As a result, the peak responses are significantly lower for all three signals, but proportionally much lower for Signal 1--the fast rise time signal--than for Signal 3. In fact, the peak response rates to Repetition 2 are approximately the same

for the three signals. The peak response rate in Repetition 1 for Signal 3 (no previous adaptation; 60 msec rise time) is approximately equal to the peak response rate in Repetition 2 for Signal 1 (significant previous adaptation; 5 msec rise time).

The gap between Repetitions 2 and 3 is 30 msecs. This recovery time is sufficiently long that the peak response to Signal 1 is somewhat higher that the peak responses for Signals 2 and 3. (The reason for the relatively strong response to Signal 1 is that its peak is due to the rapid adaptation circuit, which is the first to recover.) However, the peak response to all three signals is still substantially lower than for Repetition 1.

Before Repetition 4 the gap is 100 msecs, and significant rapid and short term recovery has occurred. However, the three previous repetitions have caused substantial long term adaptation, from which the model has not yet recovered. As a result the peak responses are still less than for Repetition 1.

In examining the signals shown in Figure 2.28, and the PAM responses to them, we are interested in considering the relationship between the PAM responses to speech signals at the transition between two speech events, and the ways in which those responses depend on the intensities of the signals

that immediately precede and follow the transition point. Two signals are shown in Figure 2.28. These signals are constructed from three levels: a LOW level (-24 dB); a MID level (6 dB); and a HIGH level (18 dB). To make the figure clearer, the LOW level for Signal 1 is actually at -21 dB, and for Signal 2 the MID level is 4.5 dB and the HIGH level 16.5 dB. Panel A shows the input to the transduction stage; Panel B shows the output of the transduction stage; and Panel C shows the output of the adaptation stage. All possible transitions between preceding and following LOW, MID, and HIGH signal levels can be found in Figure 2.28.

Beginning on the left in Panel C, consider the LOW->MID transition. The PAM response to such a transition shows a sharp peak because of the rapidity of the transition. At t = 90 msecs, a visual comparison can be made between the response to a MID->HIGH transition and a LOW->HIGH transition. The peak response to the HIGH onset is greater when it is from a LOW level than from a MID level, because of the adaptation to the MID level. After 40 msecs, the response to the HIGH level has fallen off to the point that it is not stronger than the initial transient response to the LOW->MID transition. Just as the LOW->HIGH transition is greater than the MID->HIGH transition, the HIGH->HIGH transition (at t =490 msecs) is lower than either LOW->HIGH or MID->HIGH.

In the middle of Panel C (t = 29Ø msecs) it is possible to see the difference between the PAM response to a LOW->MID transition and a MID->MID transition. Because of adaptation, the response to the MID level is substantial different depending upon whether the previous level was LOW or MID. A LOW->MID transition results in a stronger response than a MID->MID transition. Furthermore, at t = 49Ø msecs, it can be seen that the response to a HIGH->MID transition is even lower than for MID->MID.

Figure 2.29 shows another characteristic of the PAM: its transient response exhibits a larger dynamic range than its steady state response, for signals with rise times on the order of 1Ø to 4Ø msecs. This characteristic has also been observed in auditory nerve fibers (Smith and Brachman, 198Ø). Three signals are shown in Figure 2.29. All three have rise and fall times of 2Ø msecs. Signal 1 has a maximum intensity of 25 dB; Signal 2 has a maximum intensity of 5Ø dB; and Signal 3 has a maximum intensity of 8Ø dB. The input to the transduction stage is shown in Panel A; the output to that stage is shown in Panel B; and the output of the adaptation stage is shown in Panel C. The steady state response to the three signals are equal. However, the transient response to Signal 1 is only 85% of the transient response to Signal 3. This phenomemom is due to the limited dynamic range of the transduction stage, which (as can be seen in Panel B) has the

effect of sharpening the transitions of the more intense signals, so that their effective rise times are substantially shorter than their acoustic sources. This in turn increases the peakiness of the transient responses of the adaptation circuit. Note also that the more intense the signal is the wider the PAM response to it: another effect of the limited dynamic range of the transduction stage.

2.4.3.4 Implementation Issues

In the PAM a three time constant adaptation model is implemented using standard difference equation techniques. Circuit element values are calculated from specifications of parameters in the electrophysiological domain: maximum steady state firing rate; ratio of maximum transient to steady state rate; and adaptation time constants. State variables (the charge on each capacitor) are maintained separately for each PAM channel, and are updated for each sample of each channel. These input sample values are used as the sampled values of V_{in}, the voltage source shown in Figure 2.24. In test mode the output of the circuit can be set by the user to be any of the voltages or currents of the MTCAM, but in normal use the output is taken to be I_{out}, the current through the diode.

2.4.3.5 Complete Example of the Adaptation Stage Response

Figure 2.30 shows the response of the complete PAM to the two-sine-wave test signal discussed earlier in this chapter in the sections on the filter bank and transduction stages. Many of the important PAM transformation are evident: frequency analysis; non-zero spontaneous output; adaptation; masking; and recovery of spontaneous output. The effects of adaptation are shown even more clearly in Figure 2.31, which, in Panel B, shows spectral slices taken from Figure 2.30 before, during, and following the test signal. The locations of the spectral slices shown in Figure 2.31 are marked in Figure 2.30 by arrows above channel 1.

The spectral slice taken at 55 msecs (before the onset of the test signal) shows a level spontaneous output. Immediately following the onset, the channels near the "formant" frequncies, which are positioned on channels 4 and 12, respond strongly, duplicating the response of the transduction stage shown in Panel A, albeit with a higher amplitude. As the signal continues, the size of the PAM response declines, until it reaches a near steady state value at 140 msecs, 80 msecs after the onset. The curve for 171 msecs shows the output 1 msec after the offset of the test signal. Masking is evident for several channels on either side of the formant channels. As time passes, these channels begin to recover their quiescent response level, with the channels closest to the formant

peaks recovering most slowly.

2.5 POSTPROCESSORS: MEASURING PROPERTIES OF THE PAM RESPONSE

To review, the output of the PAM is a 20-channel signal. The sampling rate is one sample per channel every 462 microseconds. The value of the signal in each channel corresponds to the expected firing rate of an auditory neuron terminating on a haircell at one of 20 equally spaced locations along the basilar membrane. Thus the PAM models the transformation of acoustic signals into expected firing rate, but includes no mechanism for explicit property or feature detection.

This section describes some simple postprocessing filters that can be applied to the output of the PAM. These filters have the effect of simplifying and extracting some properties of the PAM response. They are used, in different combinations, in the two experiments described in the next chapter.

2.5.1 A Masking Detector

The PAM models the response of auditory neurons with non-zero spontaneous firing rates. Such neurons normally fire even in the absence of acoustic stimulation. However, following the offset of stimulation of sufficient intensity and duration, these fibers exhibit a complete absence of firing

for a period of time, followed by a second period during which spontaneous firing occurs, but at a lower rate than before stimulation. During this recovery period the fiber can be said to be in a masked state.

In the output of the model, a channel sample value that is less than the quiescent output rate represents an expected firing rate that is less than the spontaneous rate for the fiber modeled by the channel, and ipso facto, represents a masked state. A postprocessing filter was designed to detect such a state. This filter operates on each channel of the PAM output independently, and produces a signal with the same sampling rate as the input. The possible output values of this filter are 1 and Ø. The relationship between the channel input and output is as follows:

```
y(n) = 1        if x(n) is less than THRESHOLD;
     = Ø        otherwise.
```

where y(n) is the n'th output value, and x(n) is the n'th input value, for a particular channel. THRESHOLD is a value that can reasonably be set anywhere between zero and the quiescent output value. A lower value of THRESHOLD corresponds to a more conservative definition of the masked state. In normal use, with PAM response signals having a quiescent value of Ø.2, THRESHOLD was set to Ø.Ø1.

2.5.2 Two Smoothing Filters

The width of the impulse responses, and therefore the effective time windows, of the bark filters in the filter bank vary from 10 msecs for the low frequency channels to 1 or 2 msecs for the high frequency channels. This resolution is useful for many purposes, but needs to be smoothed to obtain expected firing rates averaged over one or more glottal pulses: times on the order of 10 to 20 msecs.

Two smoothing filters were used on the output of the PAM. Both filters convolve the output of each channel with a half hamming window. Since the left half of the hamming window is used, the effect of the convolution is to smooth the channel response, weighting recent samples within the window more than earlier ones. In the first experiment to be described in Chapter 3 (the STOP/GLIDE experiment), the non-zero width of the smoothing window is 20 msecs. In the second experiment (the VOICED/UNVOICED experiment), the width is 5 msecs. In both cases, the output of the smoothing filters is down-sampled to one sample per msec per channel.

2.5.3 Averaging across Channels

In both experiments, the final postprocessing step is to average the output from all the channels. This was done because there were no grounds for treating channels dif-

ferently from each other, and because no more sophisticated treatment was found necessary. The group delay of all channels (including preprocessing, the PAM processing proper, and postprocessing) are identical, so that multi-channel events such as glottal pulses occur at the same time in each channel. Furthermore, the preprocessing filter and the limited dynamic range of the PAM ensure that the output from each channel is weighted approximately equally.

Thus in both experiments, the final output is a single-channel signal, derived by adding the individual channel values and dividing by their number. The output signal has a 1 msec sampling rate, and models the whole-nerve expected value of either firing rate (experiment 1) or degree of masking (experiment 2) in the auditory nerve.

2.6 SOME OTHER AUDITORY MODELS

The field of auditory modeling has a long and rich history. Furthermore the label "auditory model" has been applied to systems with a wide variety of design goals and performance characteristics. In this section we will discuss some auditory models that are related to the PAM by their design objectives, response characteristics, or intended use. The systems discussed are important in their own right, and can also be considered to be examples of many other auditory models that have been implemented or proposed. Most, but not all, of the

systems are based on physiological or psychological characteristics of the peripheral auditory system. Most, but not all, are intended to be used in studying auditory response to speech.

The system developed by Searle, Jacobson, and Rayment (Searle et al., 1979) is an example of a system that combines cochlear-based preprocessing of speech with postprocessing designed to extract phonetic information from the resulting response pattern. The frequency analysis in this system is performed by an analog filter bank employing 1/3 octave filters whose characteristics were chosen to match auditory fiber tuning curves. Each bandpass filter is followed by a wide dynamic range envelope detector (low pass filter) and a logarithmic amplifier. These elements play much the same role as the transduction stage in the PAM. The digitized output of this filter bank system represents Searle's model of the "information flowing down the eighth nerve to the brain."

The second half of this system consists of feature abstraction algorithms which, for CV input signals, measure the values of such psychophysically motivated features as high frequency rise time, voice onset time, and spectral peak location. In their article, Searle et al. describe a recognition experiment in which features abstracted from CV input signals were subjected to a statistical discriminant analysis in order to measure the discrimination accuracy of the entire system.

Following training on 148 syllables, the system identified the consonant in a further 148 syllables with an overall accuracy rate of 77 percent.

A system that is somewhat similar in structure is that described by Zwicker, Terhardt, and Paulus (Zwicker et al., 1979). This system also combines an analog auditory model with a digital recognition postprocessor. The front end attempts to directly measure psychoacoustic properties of speech signals, such as loudness, pitch, roughness, and subjective duration. An analog filter bank based on a Bark frequency scale is followed by halfwave rectification, smoothing with low pass filters, and conversion to the log domain. The resulting set of "raw" signals are combined in various ways to produce the psychoacoustically-related measures of signal loudness and roughness both within each frequency band (so-called "specific" loudness and roughness), and overall ("total" loudness and roughness). As part of this process the raw signals are fed through a nonlinear circuit which combines a fast non-adapting onset response to a step signal with a two time constant decay following signal offset. Signal qualities related to timbre are measured by calculating three moments of the specific-loudness patterns. The resulting numbers appear to measure spectral balance, and perhaps "peakiness".

The postprocessing in this system consists of what are described as "standard pattern recognition procedures."

Accuracy results of various recognition experiments are reported, most of which are unremarkable, perhaps because the system is reported to be in a "preliminary state".

Chistovich and her colleagues have implemented what appears to be an all-analog system that includes frequency analysis, transduction, and adaptation stages (Chistovich et al., 1974). The frequency analysis stage models the frequency response of auditory fibers, including slopes of low and high frequency skirts, time delays, and bandwidths. This system uses a series-connected sequence of second order low pass filters, with taps between each filter, to implement the filter bank. Each tapped output is halfwave rectified and nonlinearly transformed. The resulting signal is fed to an adaptation circuit which models neural adaptation. Two such circuits are proposed in the 1974 article: one appears to be an automatic gain control circuit in which the output equals the input divided by a coefficient whose value depends upon the amount of energy in the input signal during a preceding period of time. The other circuit appears to be similar to the PAM adaptation stage, in that the output equals the input minus an offset whose value depends upon the energy in the preceding input signal. The researchers confirm that speech perception experiments support the validity of this second adaptation model over the first.

These researchers and others at the Pavlov Institute of Physiology in Leningrad have combined this peripheral model with a central auditory model. (Zhukov et al., 1974; Chistovich et al., 1982). One of the characteristics of this central model is that it detects onsets and offsets in the peripheral frequency channels in such a way as to mark phonetic segment boundaries.

Another approach to auditory modeling is direct manipulation of FFT spectra. A group of researchers at KTH in Stockholm have described this approach in several papers (Carlson and Granstrom, 1982; Blomberg et al., 1982, 1984). Beginning with a standard FFT spectrum, a series of transformations can produce spectra that represent successive approximations to an "auditory spectrum". These transformations include converting from a Hz to a Bark frequency scale, converting the amplitude scale from dB to phons or sones, and finally, converting to a pseudo-spectrum in which the height of each point along the spectrum represents the width of the surrounding region whose response is "dominated" by the nominal frequency at that point. Thus a vowel spectrum would appear as a number of sharp spikes at the formant frequencies, because these frequencies would dominate wide surrounding regions. A final model includes an unspecified technique for modeling adaptation.

These researchers used a standard isolated word recognition system to test the utility of these various representations for male and female vowel and consonant recognition. The standard FFT representation was better than any of the auditory representations, especially for consonant recognition. The model which included adaptation had one of the lowest accuracy rates. In reporting this work, the authors emphasize that it is by no means proven that auditory modeling provides the best representation from which to attempt speech recognition, at least using the standard pattern matching techniques which have proven successful with FFT- and LPC-based representations.

An important issue in auditory modeling is the extent to which the model is intended to be physiologically valid. An example of a model that is intended to directly incorporate physiological knowledge about auditory haircells is the one presented by Allen (Allen and Sondhi, 1979; Allen, 1983). This model combines a cochlear frequency analysis stage with a haircell transduction stage. The transduction stage attempts to model the variation of haircell receptor potential with cilia displacement, and the resulting diffusion of calcium ions across the haircell wall. In the model this diffusion current is taken as a measure of the expected firing rate of auditory fibers terminating at the base of the haircell. In addition to testing the model's response to simple test sig-

nals, Allen is using it in combination with a central model to study auditory response to speech signals (personal communication).

Any auditory model, such as the PAM, that includes a transduction stage with the same limited dynamic range that auditory fibers demonstrate (20 to 30 dB) must deal with the problem of modeling speech with its much wider dynamic range. Delgutte solves this problem in his auditory model by using a parallel combination of three transduction nonlinearities with differing thresholds and dynamic ranges (Delgutte, 1982). These nonlinearities model fibers with correspondingly varying response characteristics. The composite dynamic range of the three nonlinearities is close to 80 dB. The Delgutte model uses a large number of bandpass filters with shapes that more closely reproduce the shapes of auditory tuning curves than many of the other models discussed here. The model also includes a two time constant adaptation stage that is used to model rapid and short term adaptation.

The models discussed above calculate the expected (i.e., mean) firing rate of auditory neurons in response to acoustic stimuli. In a model developed by Seneff (1983, 1985) the primary intent is to reproduce the synchronicity of auditory response to stimulus waveform for frequencies in the lower half of the speech range (below about 3 KHz). Seneff's system includes a peripheral and a central model. The peripheral

model uses a bank of bandpass filters which were designed to have frequency responses similar to auditory tuning curves. The output of each filter is passed through an "envelope compressor" which uses a two time constant AGC circuit to limit the maximum signal level. The two time constants are approximately three and forty msecs, so these compression circuits can be said to model neural adaptation. Following envelope compression the signal in each channel is halfwave rectified, so that the response is always nonnegative. The resulting signal is taken to be a model of the fine time firing rate of fibers in the auditory nerve.

Seneff's central model makes use of a number of "Generalized Synchrony Detector" elements. Each of these elements measures the degree of periodicity of its input signal to a particular reference frequency. In Seneff's central model each auditory fiber response signal is processed with a GSD element tuned to the fiber's characteristic frequency. The resulting pattern, described as a "pseudo spectrogram," seems to do a good job of representing vowel-like speech signals over a wide range of input amplitudes, and in the presence of noise. Seneff also describes a scheme for using an array of the same kind of GSD elements as a pitch detector that has psychoacoustically appropriate response characteristics.

The output signals produced by all of the auditory models we have discussed have values which represent, to one extent

or another, the probability of neural firings. In a model developed by Lyon (1982, 1984) the output is a stream of bits which represent the presence (1) or absence (0) of a neural discharge in a particular fiber at a particular moment in time. This bit pattern is produced as the output of a "primary auditory neuron model" which takes as input the output of a haircell model copied after Allen's transduction model. Lyon typically models the firing of more than 2200 neurons with a sampling period of 50 usecs. With bit-stream signals such operations as auto-correlation and cross-correlation can be performed with simple logic elements in Lyon's model and, by strong implication, in the central nervous system. These operations, in turn, can be used to model such pyschoacoustic functions as pitch detection and binaural lateralization.

The models discussed above demonstrate the richness and variety of the auditory modeling that is currently underway. All of these models, including the PAM, suffer from the fact that they are partial models of a complex system, designed with limited objectives and from incomplete data. Nevertheless, the field of speech perception is likely to be considerably advanced through the use of such models to study peripheral and central auditory response to speech.

FIGURE CAPTIONS FOR CHAPTER 2

FIGURE 2.1

Block diagram of preprocessor, PAM, and postprocessor. The preprocessor shapes the speech spectra to fit within the amplitude window of the PAM, while the postprocessor measures particular properties of the PAM response pattern.

FIGURE 2.2

Magnitude of the frequency response of the preemphasis filter. The frequency axis is linear in Hz. The vertical axis shows linear gain. The diagonal line shows the frequency response of a perfect 6 dB/octave preemphasis filter.

FIGURE 2.3

Block diagram of the Peripheral Auditory Model. The spectral analysis stage is a 20-channel filter bank. The following stages are shown for only one channel of the filter bank. Each channel has a similar and independent set of transduction and adaptation stages.

FIGURE 2.4

Graph of the magnitude of the frequency response of three adjacent PAM filter bank filters. The horizontal axis shows frequency shift from the center frequency of the center channel in Barks. The vertical axis shows magnitude in dB of frequency response. These curves apply for any of the PAM bandpass filters, except for the lowest and highest channels, where the frequency responses are truncated at DC and the Nyquist frequency.

TABLE 2.1

List of center frequencies of the PAM bandpass filters. The center frequencies nominally appear at integral Bark values between 1 and 20, but are adjusted so that they fall on the closest multiple of 50.7 Hz, the frequency spacing of the 256-point DFT that is used to construct the impulse response of the filter bank channels.

FIGURE 2.5

Gain and phase of even numbered PAM filter bank channels. Frequency is shown in Barks, and gain is linear. Note that the magnitude characteristics of all of the filter banks are identical on this frequency scale, while the slope of the phase curves increase with channel number.

FIGURE 2.6

Gain and phase of even numbered PAM filter bank channels. Frequency is shown in Hz, and gain is in dB. Note that the bandwidths of the filters increase with channel number on this frequency scale, while the slopes of the phase curves are constant.

FIGURE 2.7

Magnitude of frequency response of individual PAM filters. Frequency is shown in Hz, and gain is in dB. Notice that on this scale, the filters appear almost symmetrical, although the low frequency slopes rise at 10 dB/Bark, and the high frequency slopes fall off at 25 dB/Bark.

FIGURE 2.8

The composite frequency response of the PAM filter bank. The solid line is for the PAM filter bank alone. The dashed line includes the preemphasis filter.

FIGURE 2.9

Impulse response of PAM filter bank. The magnitude of the impulse response of each filter is shown. The peak response to the impulse occurs simultaneously in all channels. Each horizontal tick is one msec. The ringing in channel 19, and the anomalous width of channel 20, are artifacts resulting from the abrupt high frequency truncation of the frequency responses of those channels. These curves include the effect of windowing the impulse response with a Hamming window, but do not include the effect of the preprocessing stage shown in Figure 2.1.

FIGURE 2.10

PAM filter bank response to two sine waves. The center frequencies of each channel in Hertz and Bark are listed along the sides. The vertical dimension is output signal magnitude. The vertical distance between the zero lines of adjacent channels corresponds to a magnitude of 45. The input signal, which consists of the sum of a 406 Hz and a 1674 Hz sine wave, is shown at the bottom, delayed by 9.9 msecs. This is the group delay of the PAM filters, so that the input and output signals are vertically aligned.

FIGURE 2.11

Spectral slices of the PAM filter bank response. This figure shows two views of the response of the PAM filter bank to the same test signal shown in Figure 2.10. In panel A, the vertical dimension is gain in linear units. In panel B, the vertical dimension is gain in dB (re 1.0). Panel B should be compared with Figure 2.4.

FIGURE 2.12

PAM transduction stage and its derivation. Panel A shows the structure of the PAM transduction stage. This model synthesizes the low frequency response of the idealized transduction model shown in Panel B. The nonlinearity in Panel A is the DC transfer function of the model shown in Panel B. This

transfer function is generated using the measuring setup shown in Panel C. See the text for a detailed explanation.

FIGURE 2.13

Four possible transduction functions. The four functions shown are a raised hyperbolic tangent, a raised half cosine, a ramp, and a step function. Notice that the transition region of the first three functions are all approximately equal, but that the step function has a zero width transition region.

FIGURE 2.14

DC transfer functions. These functions are the result of using the the raised hyperbolic tangent transduction function shown in Figure 2.13. Each curve is the result of setting the bias to a different value, which is shown next to each curve.

FIGURE 2.15

DC transfer functions for the four transduction functions shown in Figure 2.13. Notice that the shape of the underlying transduction functions has little effect on the shape of the resulting transfer functions, especially near the saturating amplitudes of transfer functions. It is clear from the characteristics of the transfer functions resulting from the step function that the presence of a transition region in the transfer function gives rise to transfer functions with non-zero quiescent values.

FIGURE 2.16

PAM transduction stage response to two sine waves. The input signal, and picture format is the same as in Figure 2.10. The vertical dimension is transduced channel amplitude. The vertical distance between the zero lines of adjacent channels corresponds to an amplitude of one.

FIGURE 2.17

Comparison of the PAM filter bank and transduction stage responses. The input signal is the same as in Figure 2.16. Panel A, which is duplicated from Panel B of Figure 2.11, shows the output of the filter bank stage with channel amplitude in dB. Panel B shows the output of the transduction stage: another view of the response shown in Figure 2.16. In panel A, the vertical dimension is gain in dB (re 1.0). In panel B, the vertical dimension is transduced amplitude in linear units.

FIGURE 2.18

Single time constant adaptation circuit. The input variable is V_{in}, the source voltage, and the output variable is I_{out}, the current through the diode.

FIGURE 2.19

Response of the STCAM to a step onset and offset. Panel A shows V_{in} as a function of time. Panel B shows I_{out}. Although the current through the diode is zero when the voltage across the diode is negative, the dotted line in Panel B shows the current that would flow if the diode were to be momentarily short circuited at any given instant.

FIGURE 2.20

Demonstration of the constant incremental gain characteristics of the STCAM. Panel A shows V_{in}, and Panel B shows I_{out}. A series of eight overlapping signals are shown. Each signal begins at the quiescent value, increases stepwise to the first plateau, and after a delay of 0 to 280 msecs, rises again to a higher level. Note that the size of the incremental response to the second step is independent of the delay between the first and second step, and independent of the degree of adaptation that has occurred during that delay. The size of the input steps were adjusted for equal amplitude AFTER the transduction stage.

FIGURE 2.21

Origins of nonlinearities in the PAM response. Panel A is the input to the transduction stage. Note that the vertical axis in Panel A is in dB. Panel B is the output of the transduction stage. Note the quiescent response level, and the saturation in response to higher and higher input levels. Panel C is the output of the STCAM. Note the additional nonlinearites created by the diode in the adaptation circuit. The transduction stage is responsible for nonlinearities in the response to onsets, and the adaptation stage is responsible for nonlinearities in response to offsets.

FIGURE 2.22

Demonstration of forward masking in the responses of the STCAM. Panel A shows the input to the STCAM. Panel B shows the output of the STCAM. The dotted line in Panel B is the hypothetical baseline quiescent response. The initial (masking) signal is intense enough and long enough to cause almost complete adaptation. This reduces or eliminates the response to the following (probe) signals.

FIGURE 2.23

Another demonstration of masking phenomenom, and the effect of the intensity of the masking signal. Panel A is the input to the STCAM, and Panel B is the output.

FIGURE 2.24

Multiple time constant adaptation circuit. The input variable is V_{in}, the source voltage, and the output variable is I_{out}, the current through the diode.

FIGURE 2.25

Adaptation circuit with two time constants.

FIGURE 2.26

Responses of the circuit shown in Figure 2.2Ø. Panel A shows V_{in}. Panel B shows I_{out} (including its hypothetical negative portion) and I_s. Panel C shows $Vc_Ø$ and $I_Ø$, the response of the 15 msec time constant elements. Panel D shows Vc_1 and I_1, the responses of the 4Ø msec time constant elements.

FIGURE 2.27

Response of the MTCAM to signals with ramp onsets of various slope. Three overlapping signals are shown, with rise times of 5, 2Ø, and 6Ø msecs. Panel A shows the input to the adaptation stage, Panel B the output of that stage, and Panel C the output of the adaptation stage. The gaps between adjacent signals are 5, 3Ø, and 1ØØ msecs respectively.

FIGURE 2.28

Effect of previous signal levels on the size of the response of the MTCAM to onsets and offsets. Two different overlapping signals are shown. Each signal contains transitions between a LOW, MID, and HIGH level. The timing of the transitions, and their exact size, is adjusted to assist in the interpretation of the figure. Panel A is the input to the transduction stage, Panel B the output of the transduction stage, and Panel C the output of the adaptation stage.

FIGURE 2.29

Demonstration of the dynamic range of the transient and steady state response of the MTCAM. Three overlapping signals are shown. Panel A shows the input to the transduction stage. Each signal has a rise and fall time of 2Ø msecs Panel B shows the output of the transduction stage. Note that the response to the signals has saturated, but that the response to the highest signal is wider than the response to the lowest signal. Panel C show the output of the adaptation stage. The incremental response to the three signals are different, even though the output of the transduction stage has saturated. This is because the onset of the highest signal is more abrupt

after it has been windowed though the transduction stage. The greater abruptness results in a stronger transient response.

FIGURE 2.30

PAM adaptation stage response to two sine waves. The input signal, and picture format is the same as in Figure 2.10. The vertical dimension is normalized channel amplitude. The vertical distance between the zero lines of adjacent channels corresponds to a normalized amplitude of three. The arrows mark the times at which the spectral slices shown in Figure 2.31 occur.

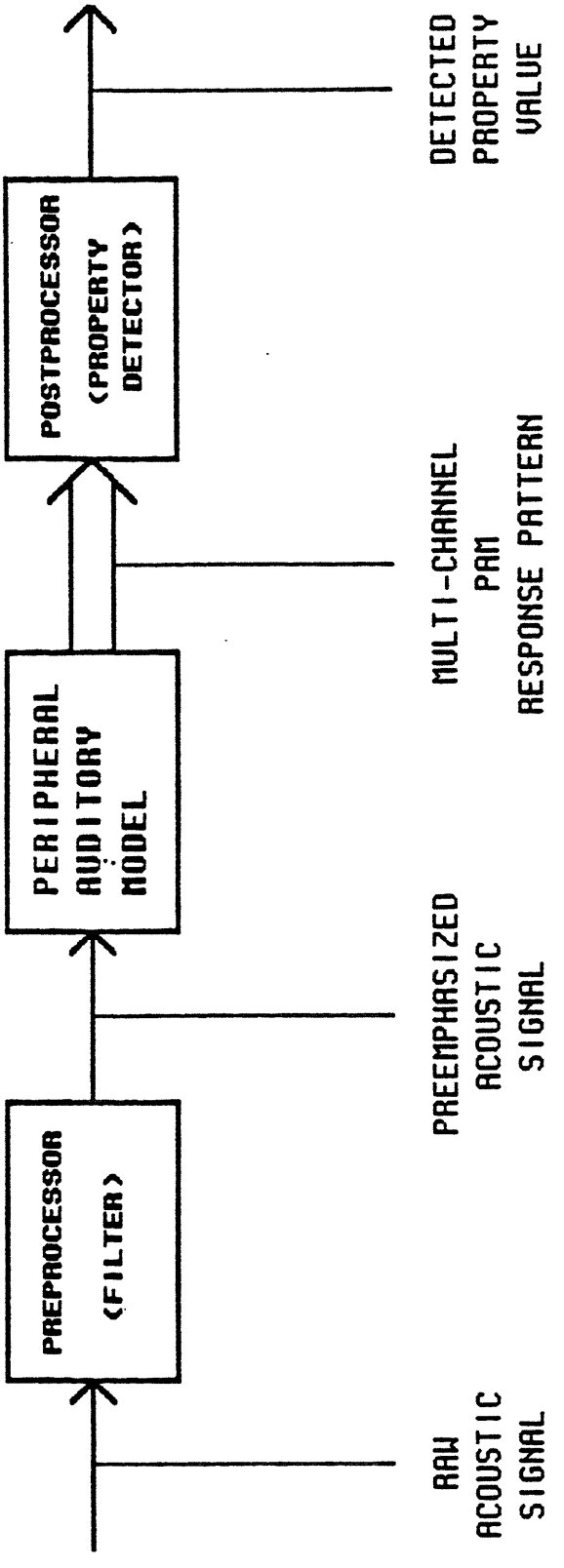
FIGURE 2.31

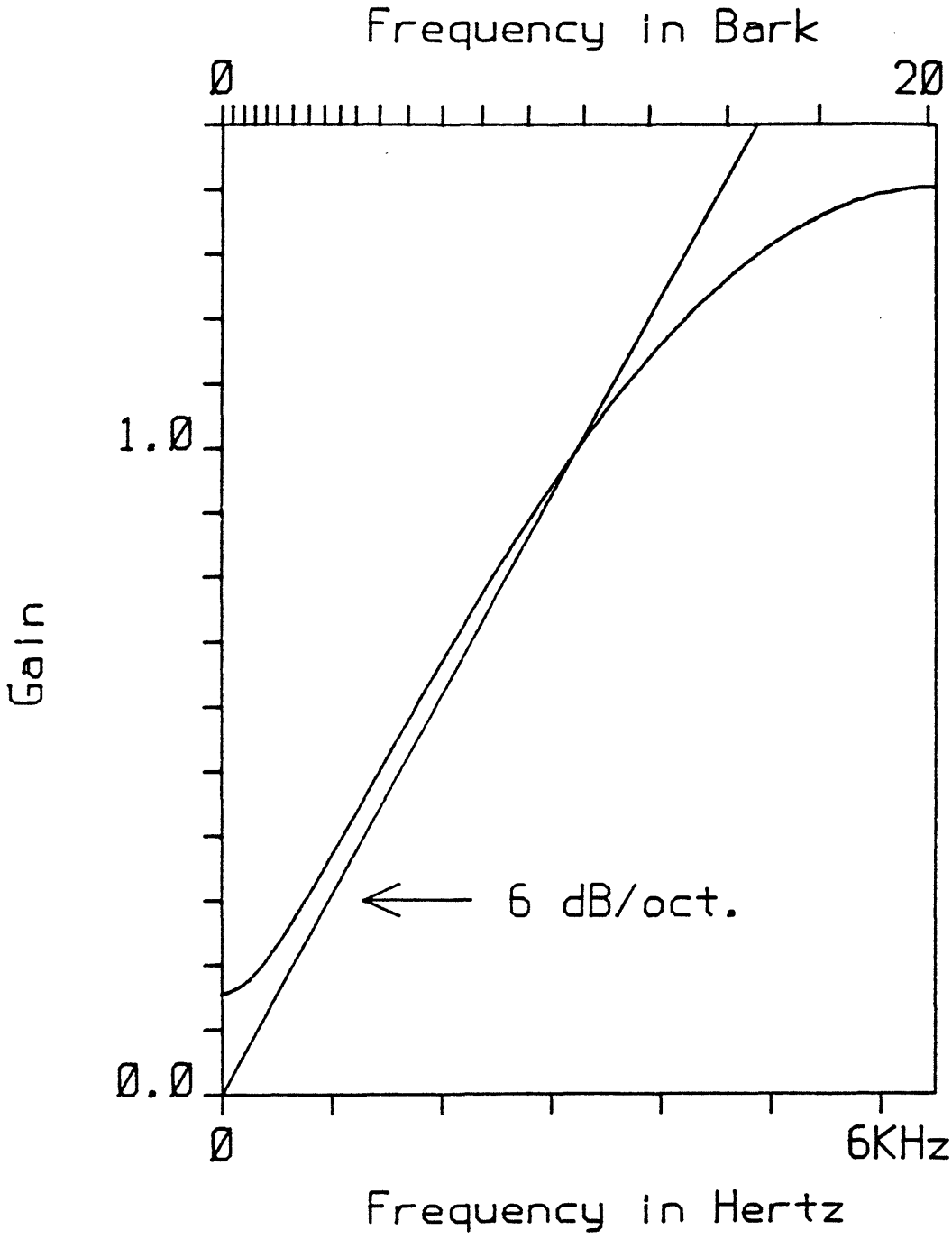
Comparison of the transduction stage response, and adaptation stage response at various times. The input signal is the same as in Figure 2.30. Panel A, which is duplicated from Panel B of Figure 2.17, shows the output of the transduction stage at 80 msecs. Panel B shows the output of the adaptation stage at 55 msecs (before the onset of the signal), 65, 70, 80, and 140 msecs (during the signal interval), and 171, 210, 270, and 390 msecs (after the offset of the signal). In both panels, the vertical dimension is normalized amplitude.

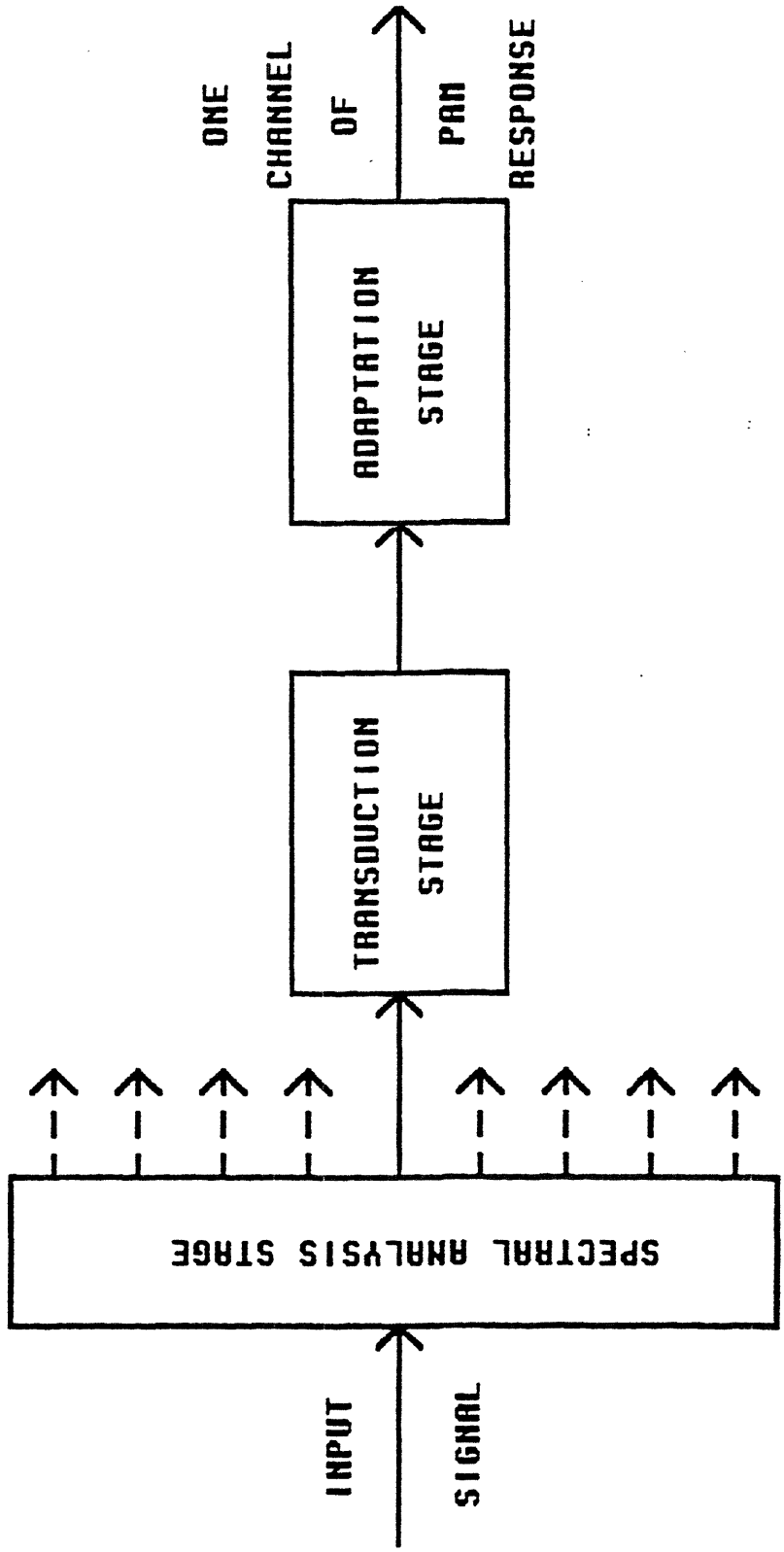
FIGURE 2.32

Masking postprocessor response to two sine wave test signal. The input signal, and picture format is the same as in Figure 2.30. Each channel output value is zero if the corresponding value in Figure 2.30 is greater than zero, and one if the corresponding value in Figure 2.30 is zero. The arrows mark the times at which the post-offset spectral slices shown in Figure 2.31 occur.

FIGURE 2.1


RAW ACOUSTIC SIGNAL
PREPROCESSOR (FILTER)
PREEMPHASIZED ACOUSTIC SIGNAL
PERIPHERAL AUDITORY MODEL
MULTI-CHANNEL PAM RESPONSE PATTERN
POSTPROCESSOR (PROPERTY DETECTOR)
DETECTED PROPERTY VALUE


Frequency in Bark
0
20
1.0
0.0
Gain
6 dB/oct.
0
6KHz
Frequency in Hertz



PERIPHERAL AUDITORY MODEL

FIGURE 2.4

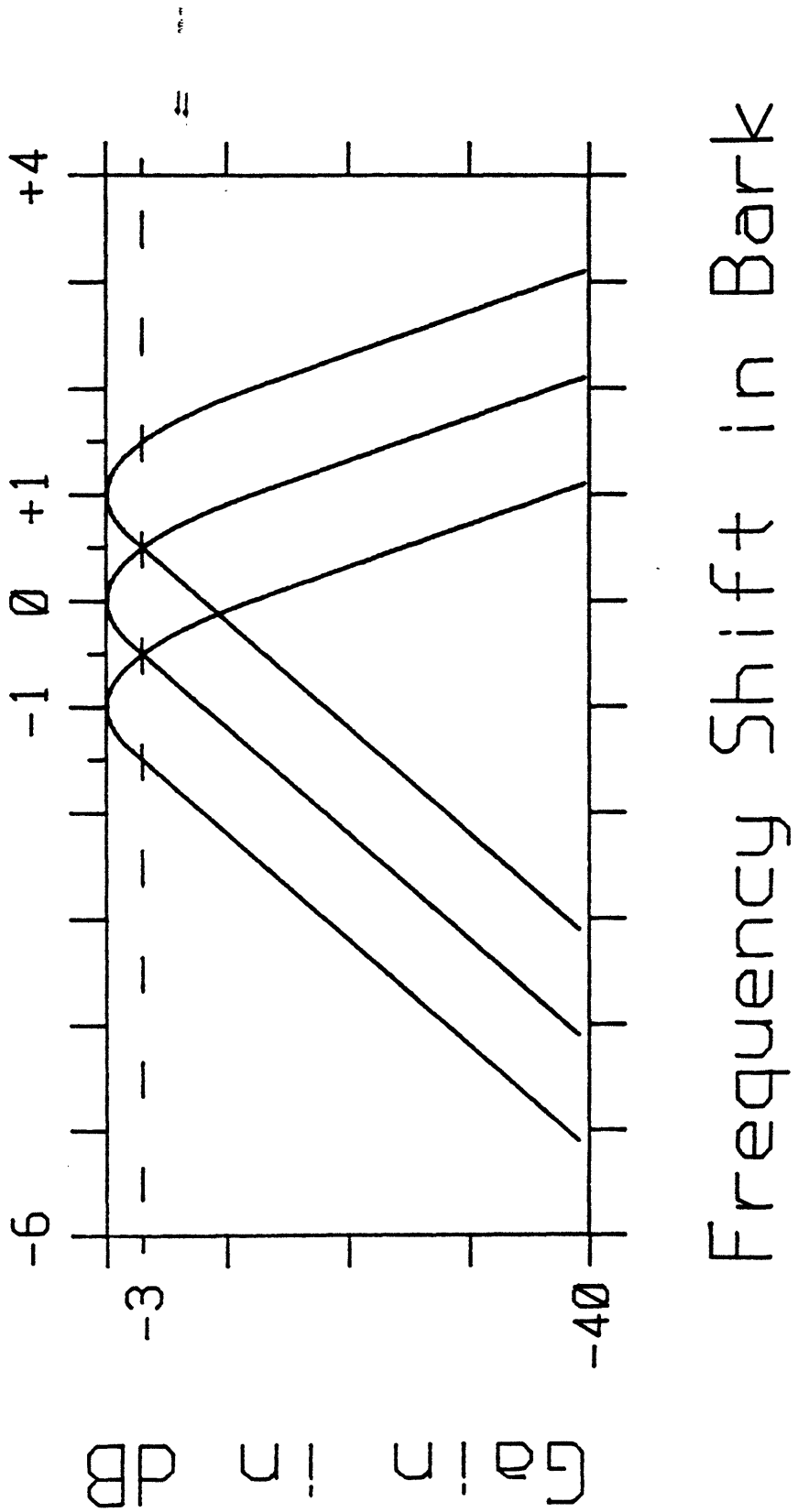


TABLE OF PAM FILTER CENTER FREQUENCIES

Filter Number	Center Frequency in Barks	Center Frequency in Hertz
1	1.015	101.5
2	2.029	202.9
3	3.044	304.4
4	4.058	405.8
5	5.051	507.3
6	6.117	659.5
7	6.827	761.0
8	7.892	913.1
9	8.957	1065.
10	10.02	1218.
11	10.95	1421.
12	11.94	1674.
13	12.94	1979.
14	14.06	2384.
15	15.00	2790.
16	16.01	3298.
17	17.02	3906.
18	18.02	4617.
19	19.00	5428.
20	19.98	6392.

FIGURE 2.5

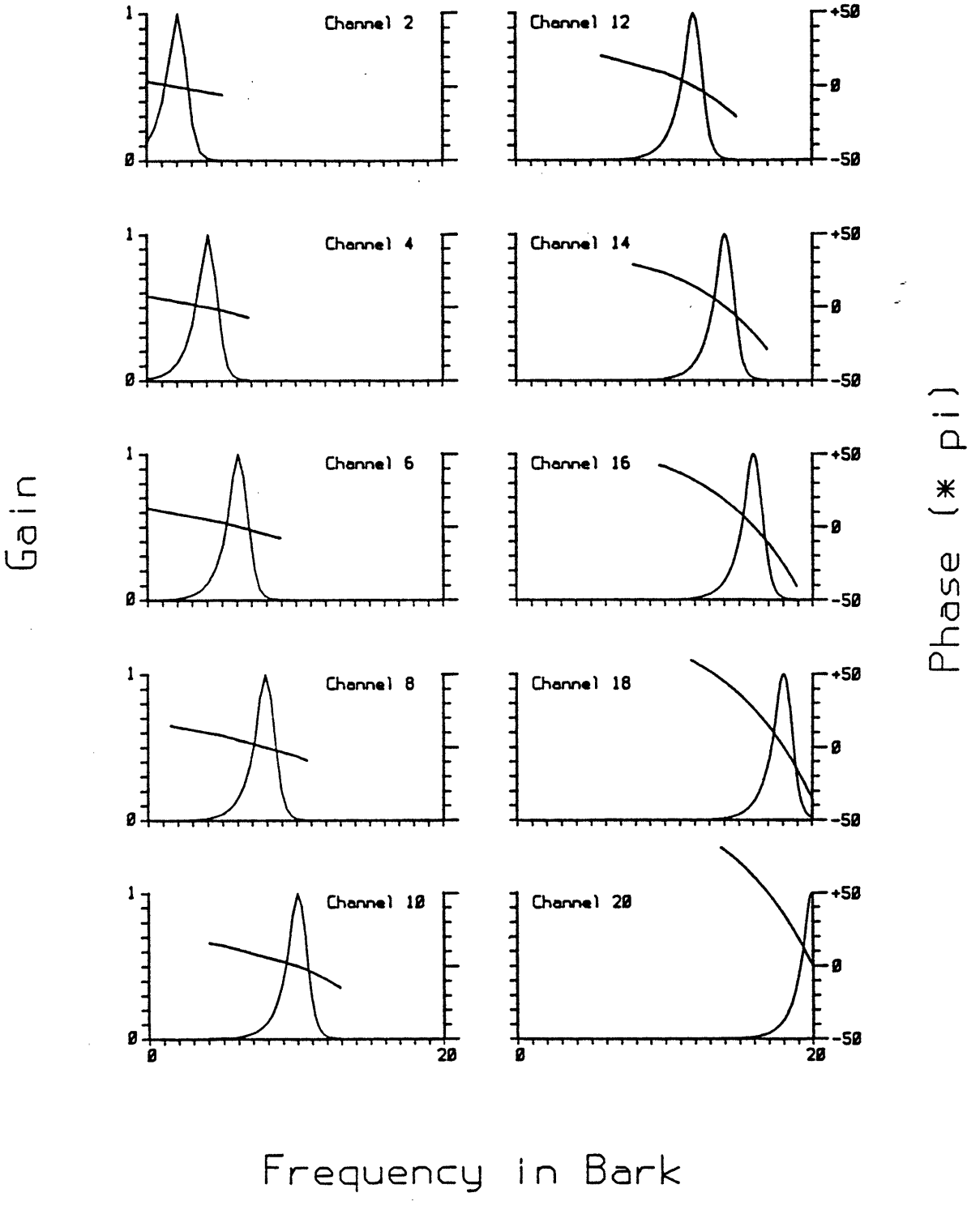

Channel 2
Channel 12
Channel 4
Channel 14
Channel 6
Channel 16
Channel 8
Channel 18
Channel 10
Channel 20
1
0
+50
-50
20
Gain
Phase (* pi)
Frequency in Bark

FIGURE 2.6

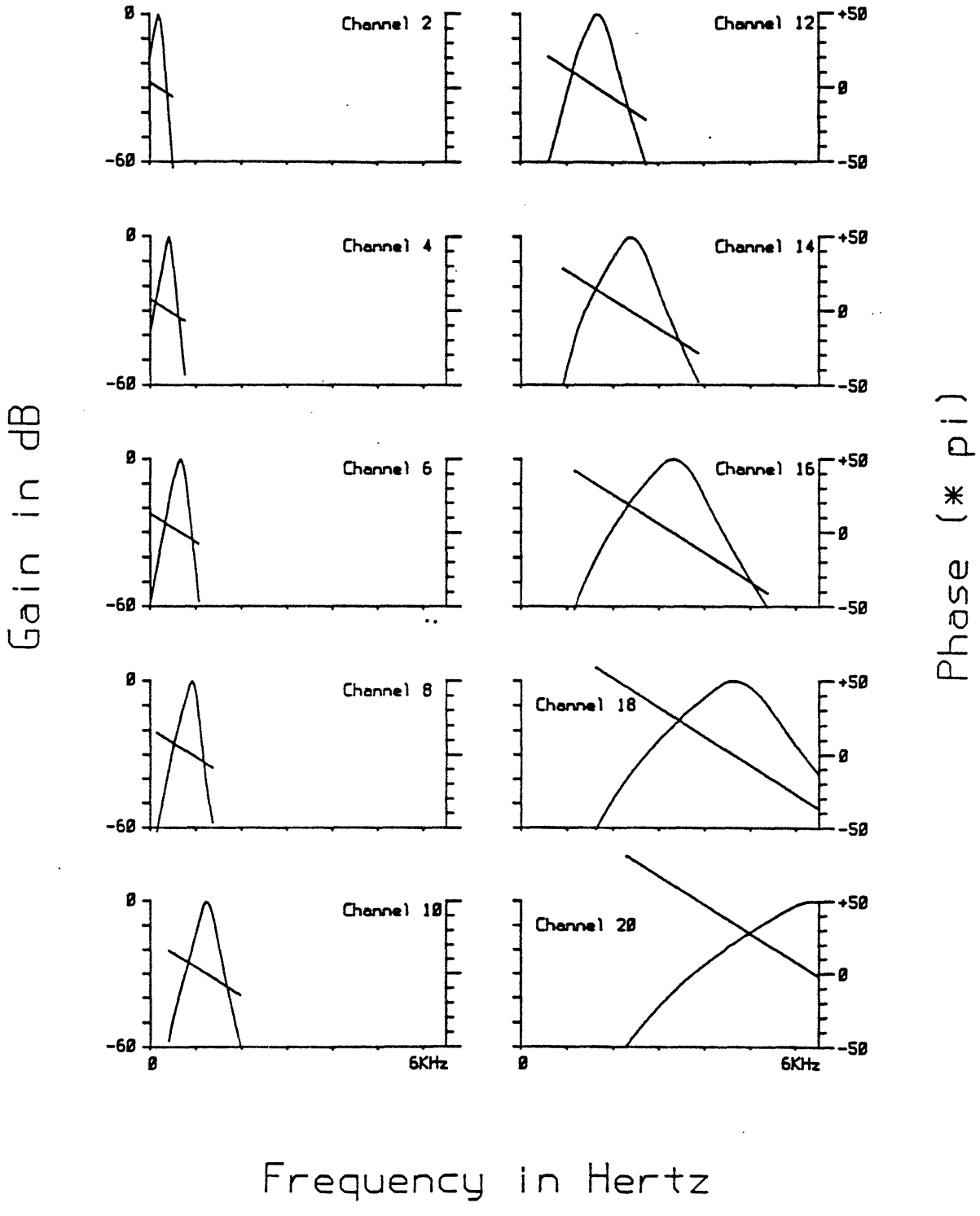

Channel 2
Channel 12
Channel 4
Channel 14
Channel 6
Channel 16
Channel 8
Channel 18
Channel 10
Channel 20
0
-60
+50
-50
6KHz
Gain in dB
Phase (* pi)
Frequency in Hertz

FIGURE 2.7

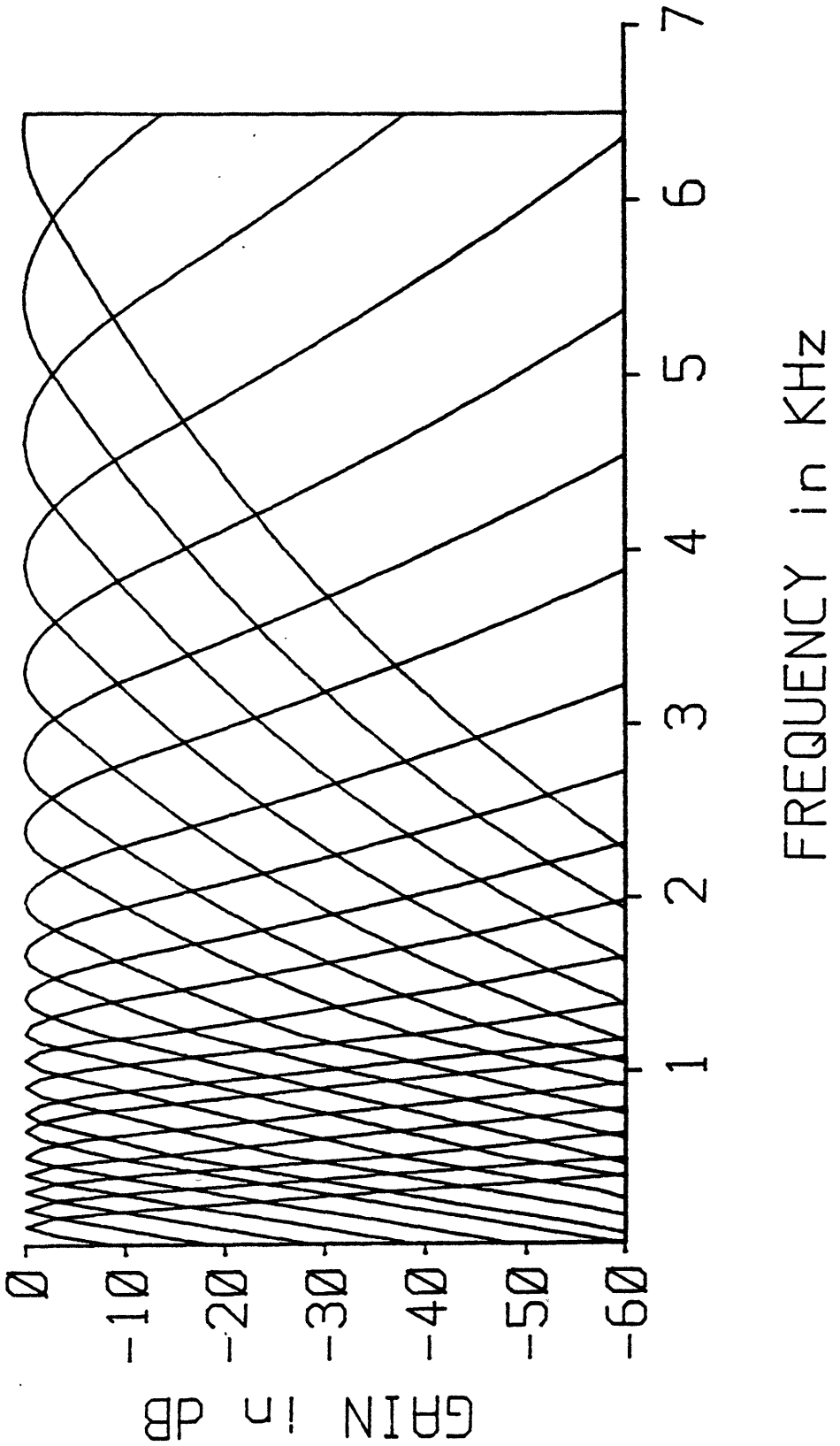


FIGURE 2.8

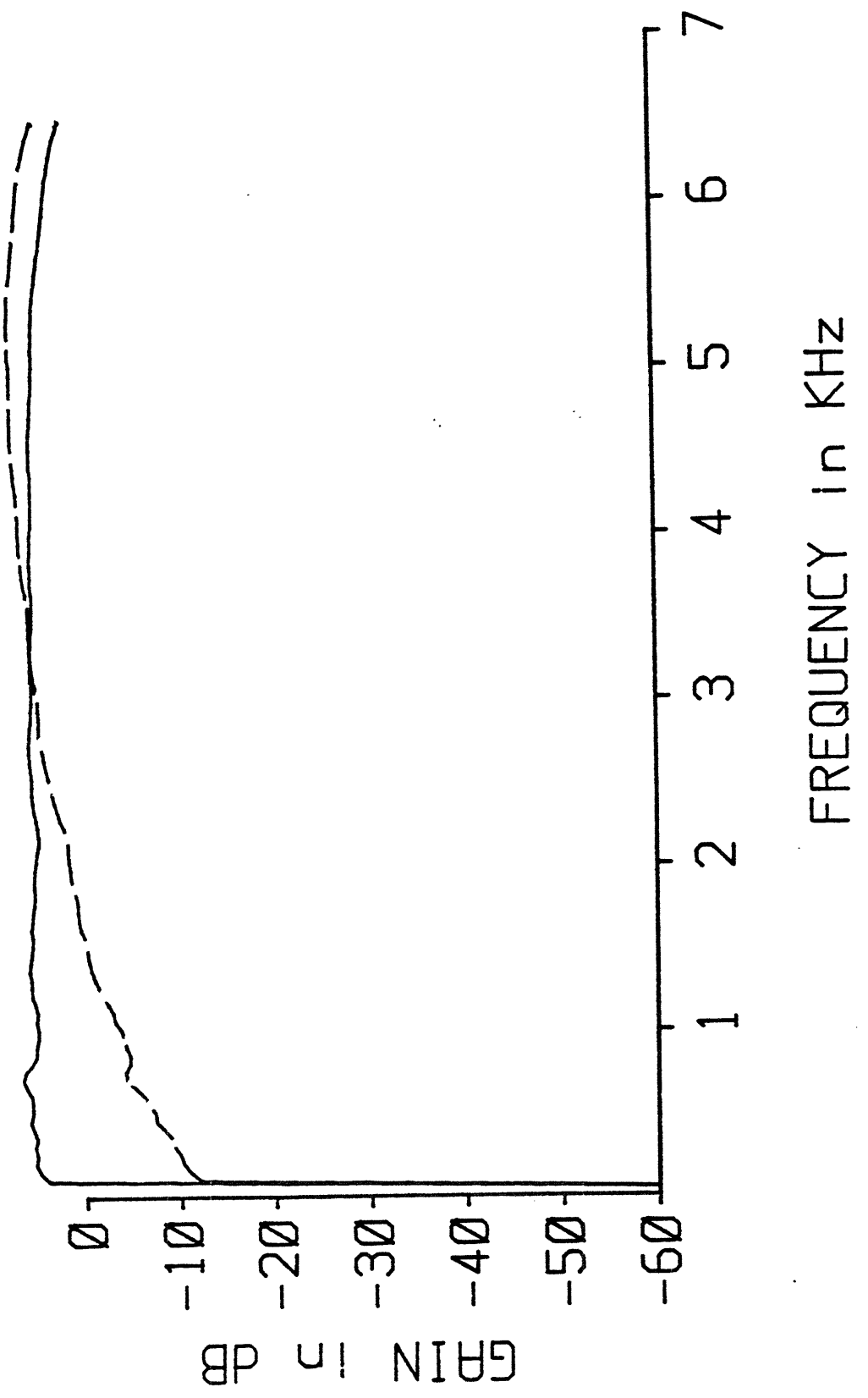


FIGURE 2.9

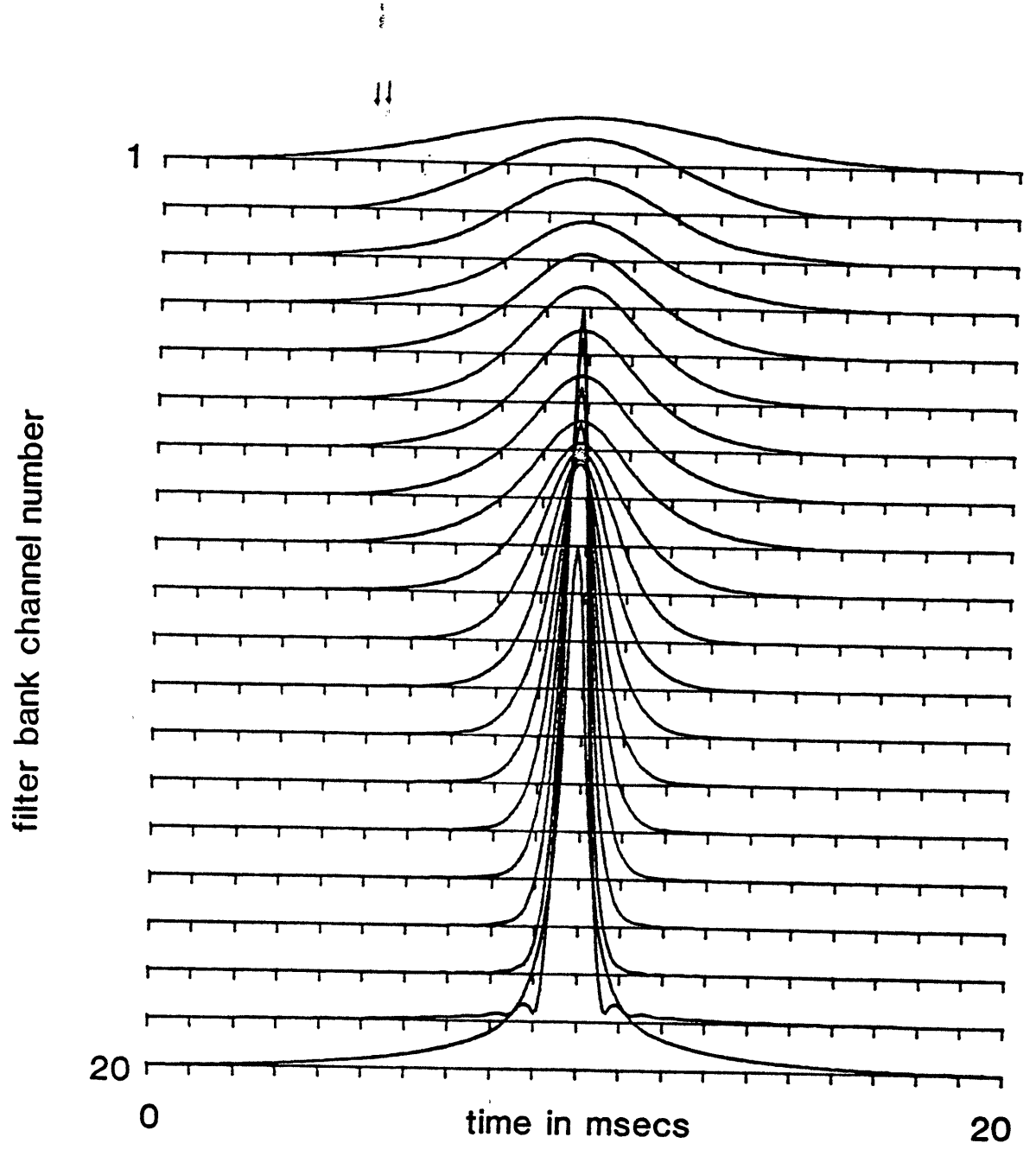


FIGURE 2.10

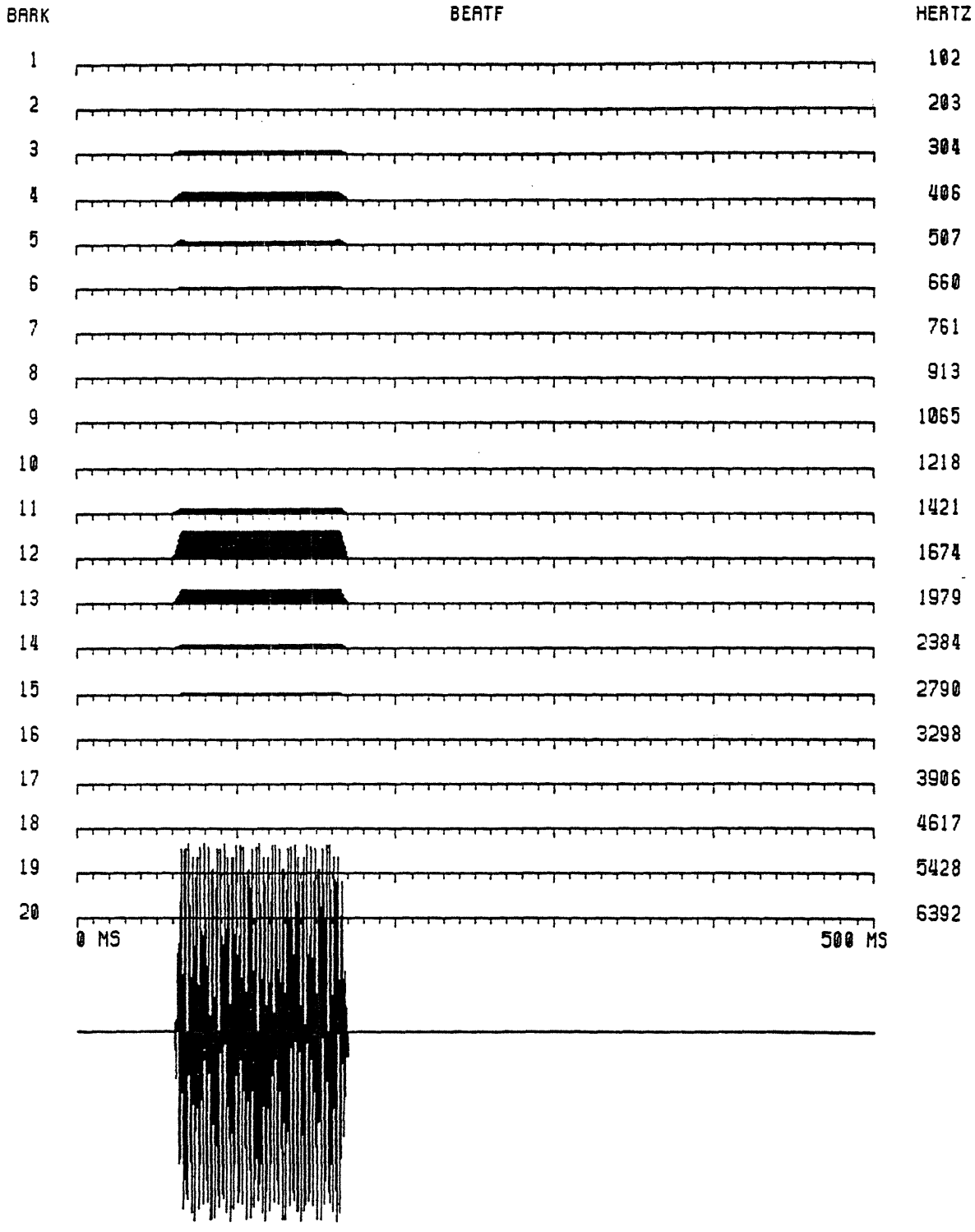
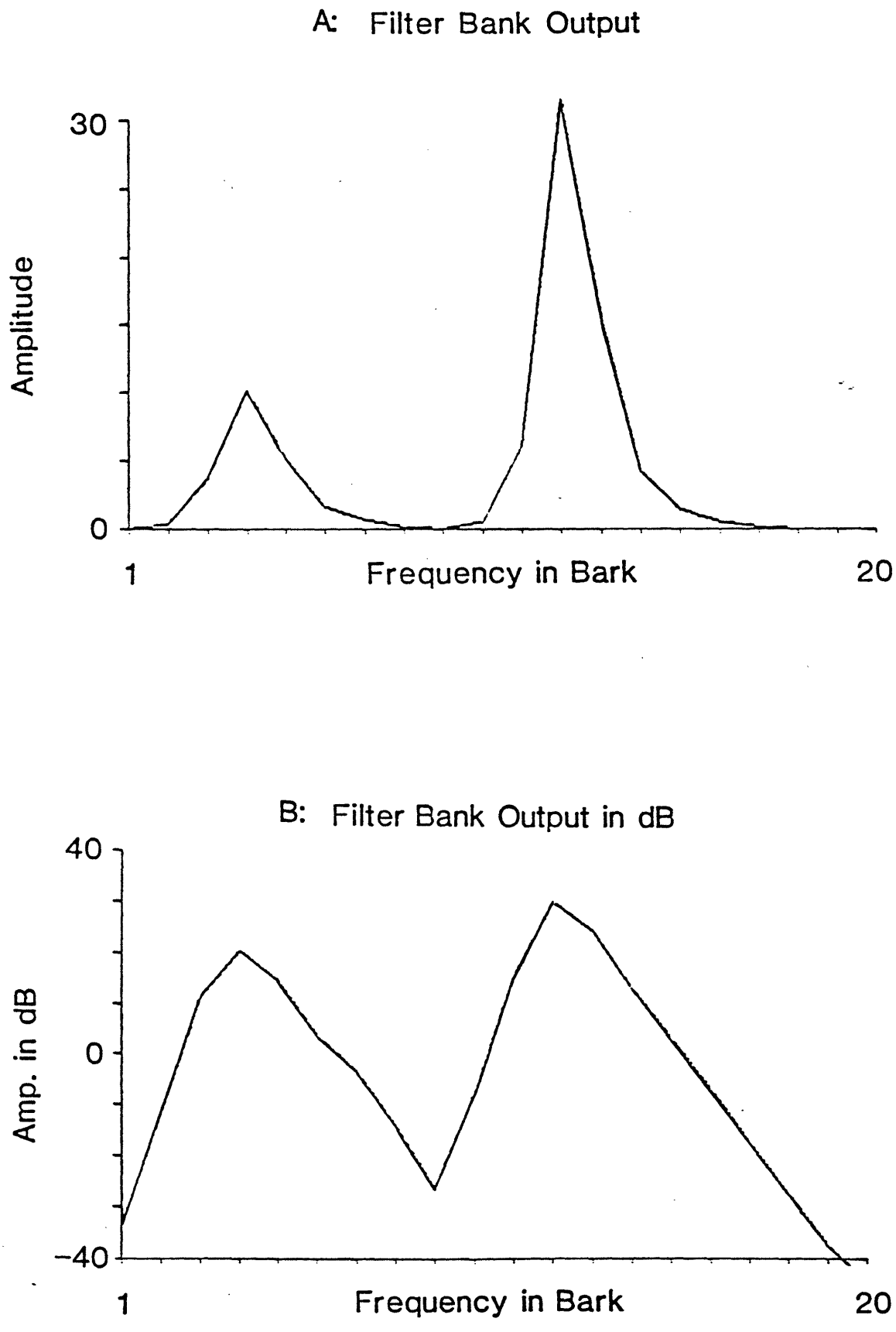
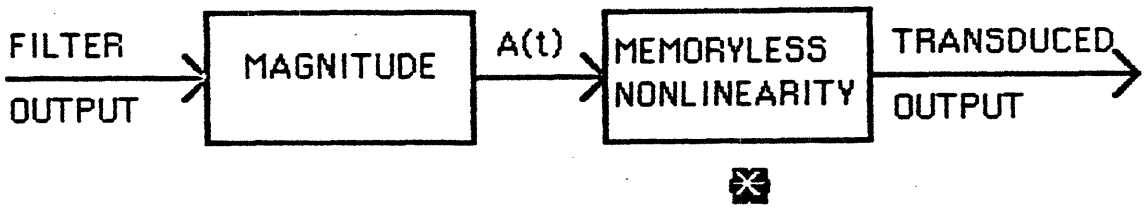

BARK
BEATF
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

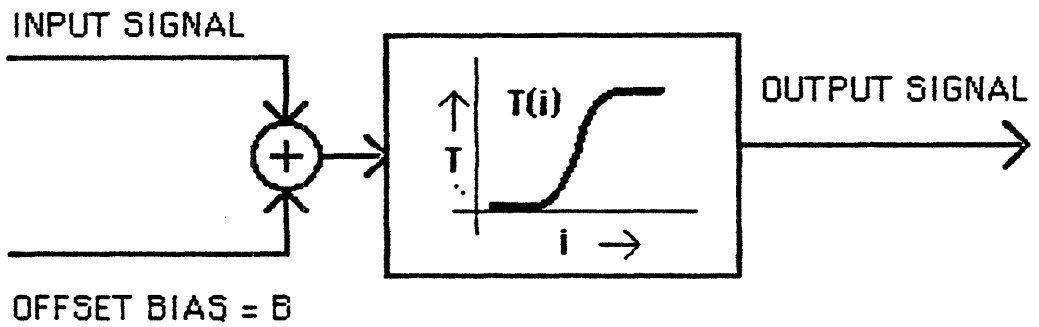
FIGURE 2.11


A: Filter Bank Output
30
0
Amplitude
1
Frequency in Bark
20
B: Filter Bank Output in dB
40
0
-40
Amp. in dB
1
Frequency in Bark
20

A: PAM TRANSDUCTION STAGE



B: IDEALIZED TRANSDUCTION MODEL



C: MEASURING DC TRANSFER FUNCTION

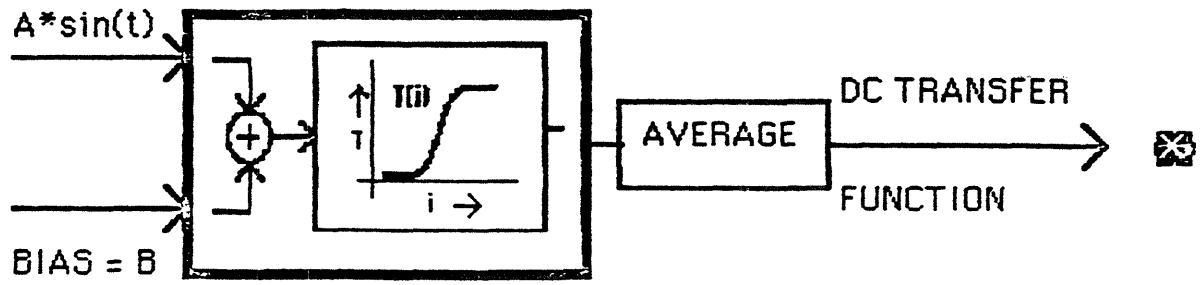


FIGURE 2.13

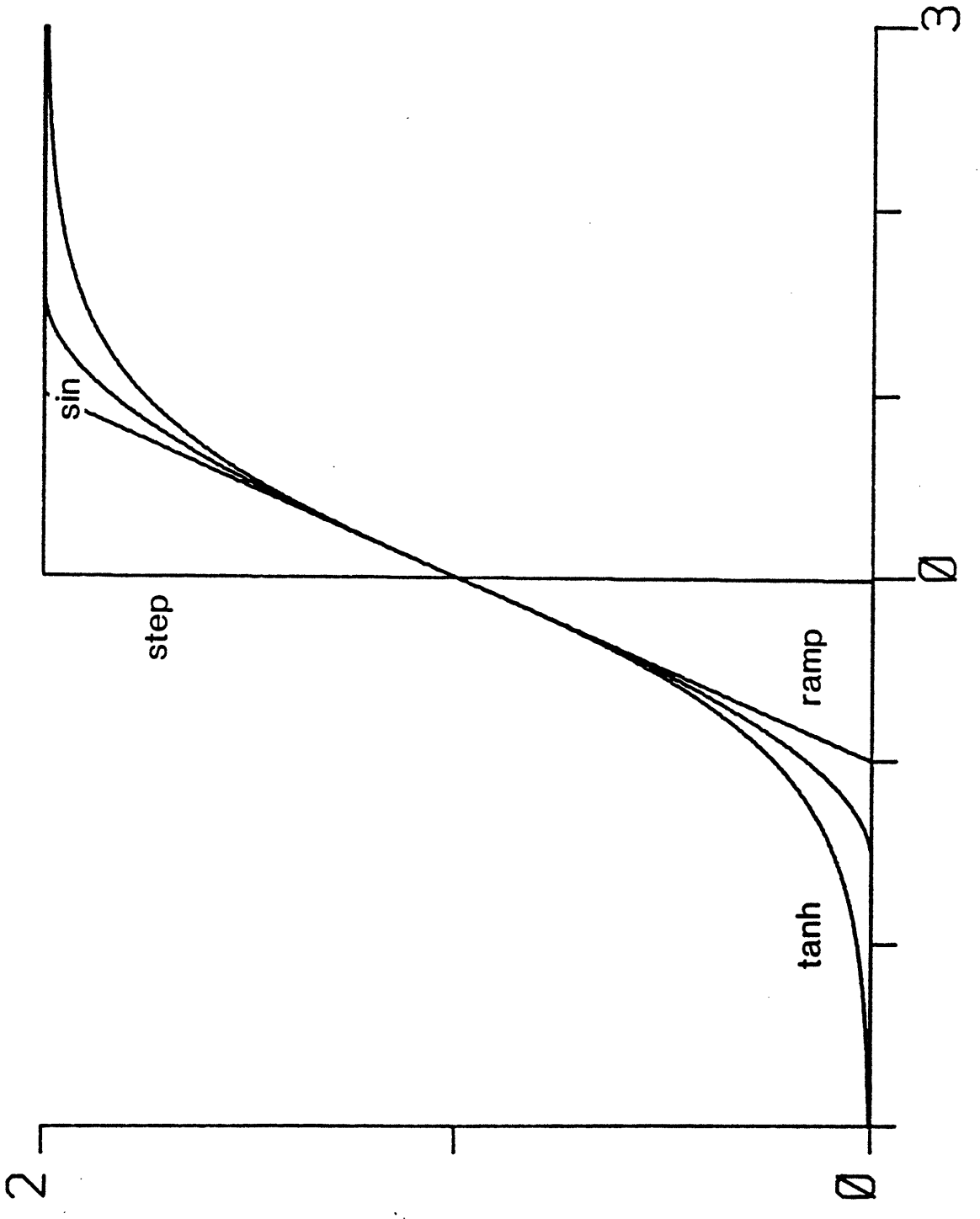


FIGURE 2.14

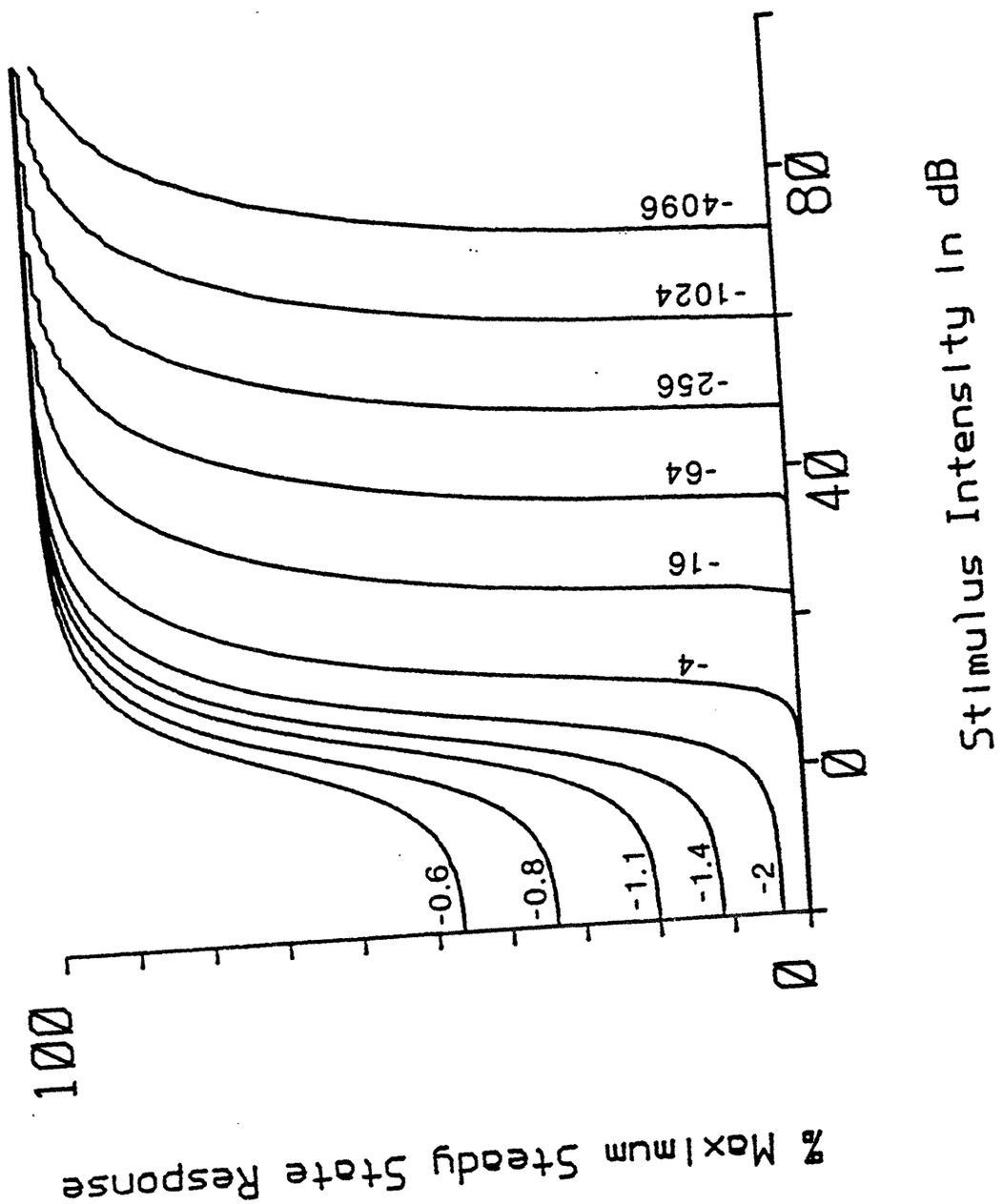

% Maximum Steady State Response
100
0
-0.6
-0.8
-1.1
-1.4
-2
-4
-16
-64
-256
-1024
-4096
0
40
80
Stimulus Intensity in dB

FIGURE 2.15

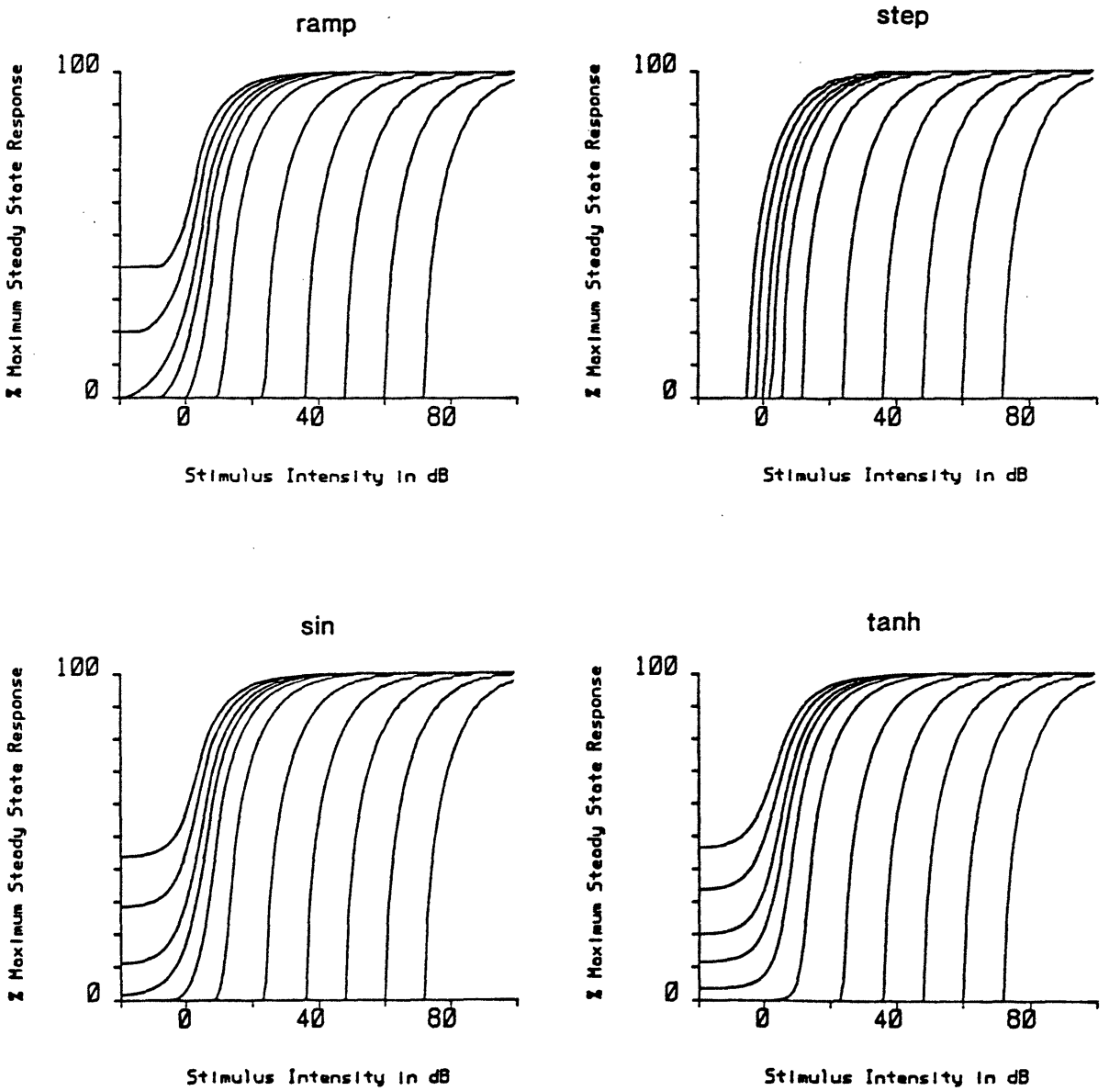
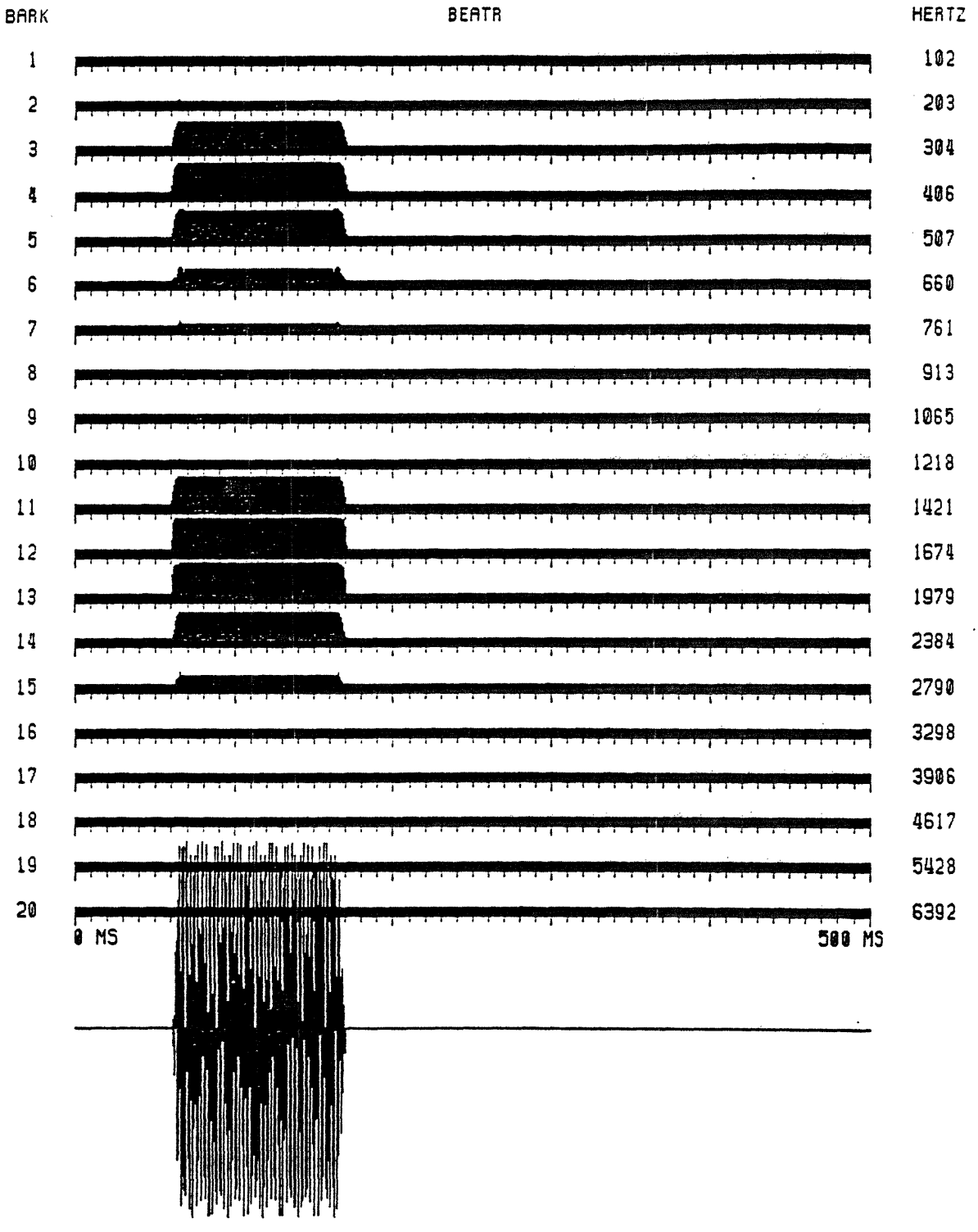
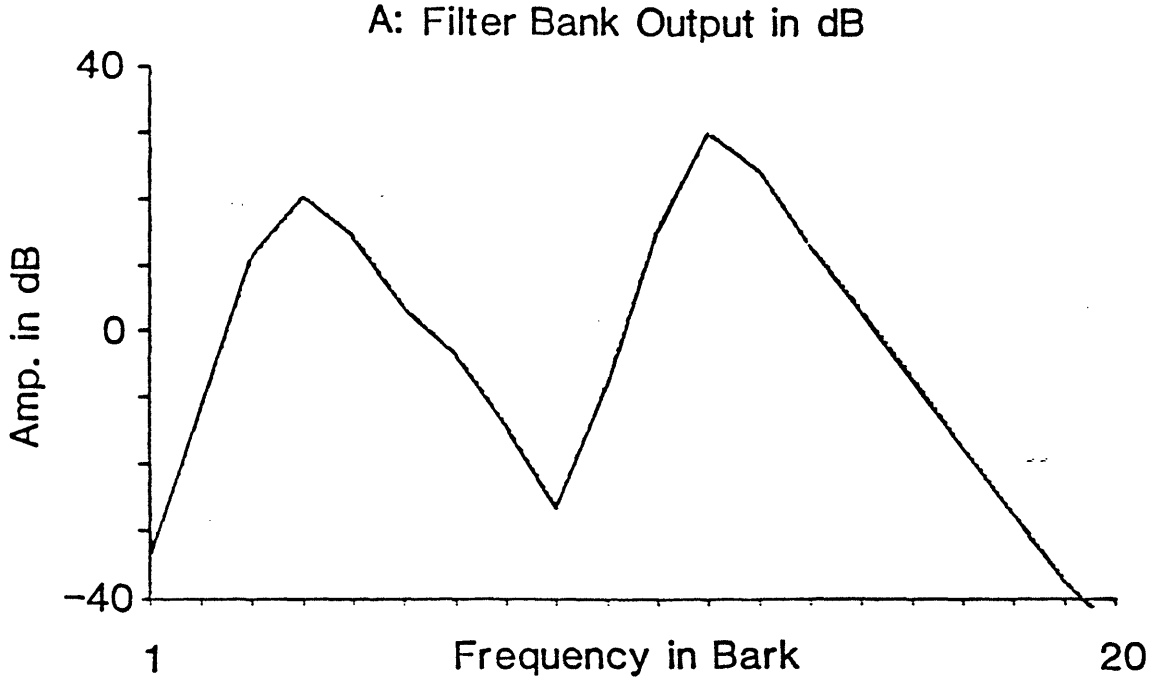
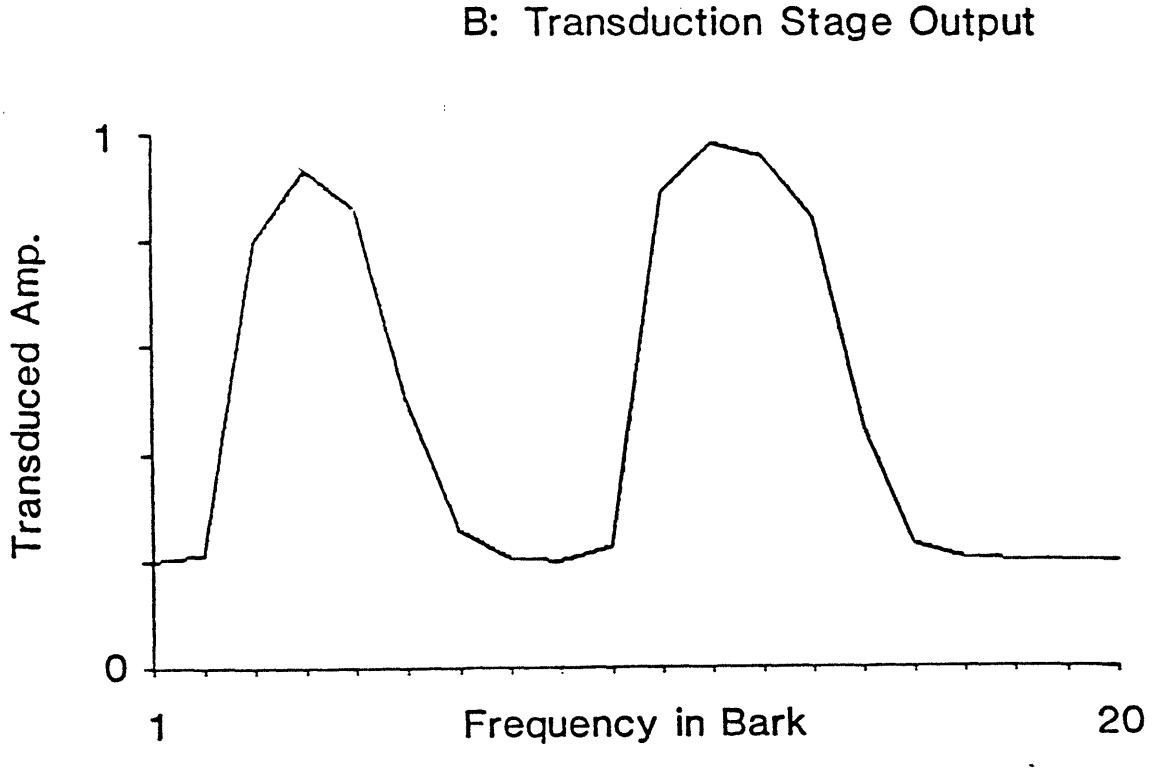
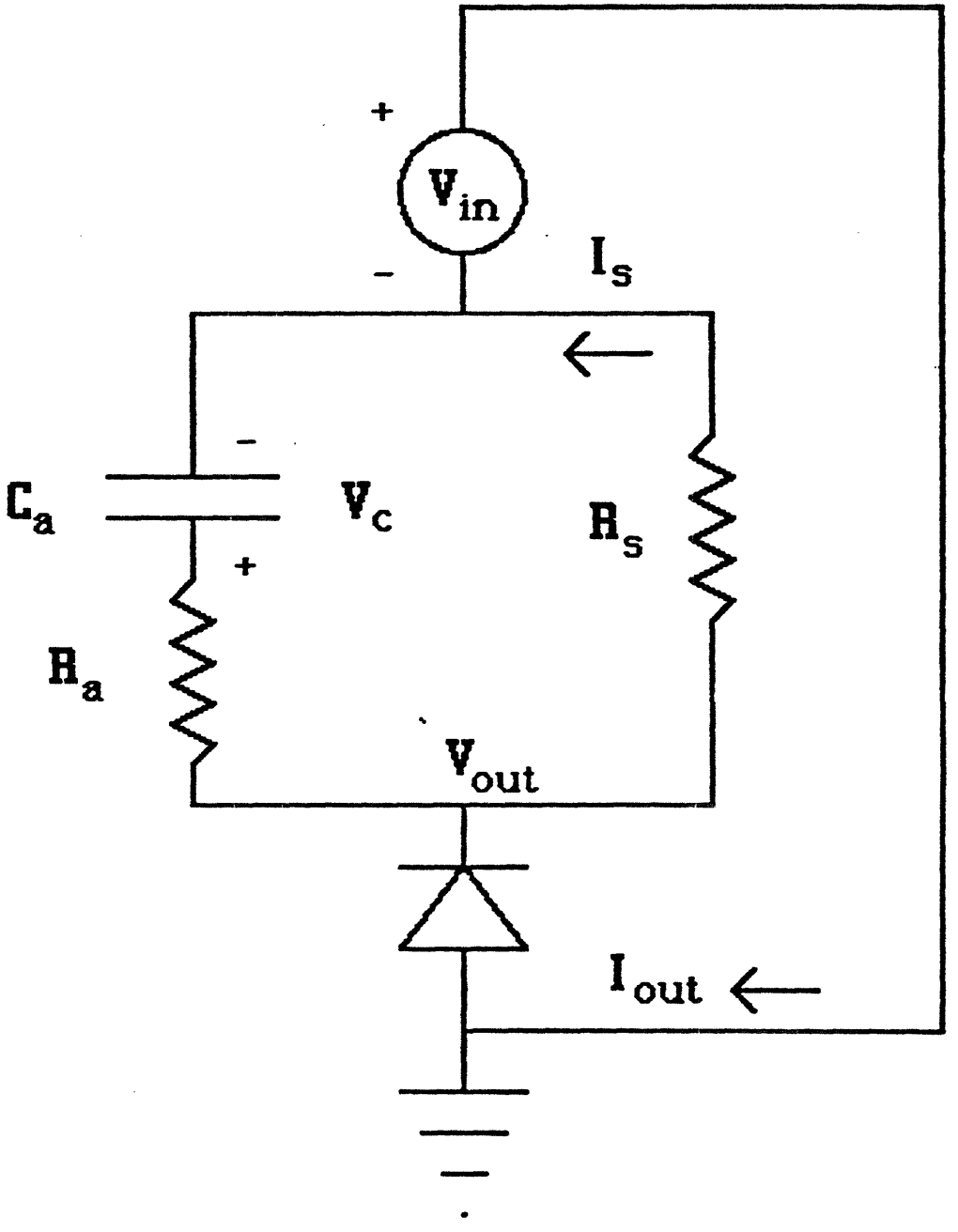

ramp
step
sin
tanh
% Maximum Steady State Response
100
0
0
40
80
Stimulus Intensity in dB

FIGURE 2.16


BARK
BEATR
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS


A: Filter Bank Output in dB
40
0
-40
Amp. in dB
1
Frequency in Bark
20


B: Transduction Stage Output
1
0
Transduced Amp.
1
Frequency in Bark
20



Single Time Constant Adaptation Model

FIGURE 2.19

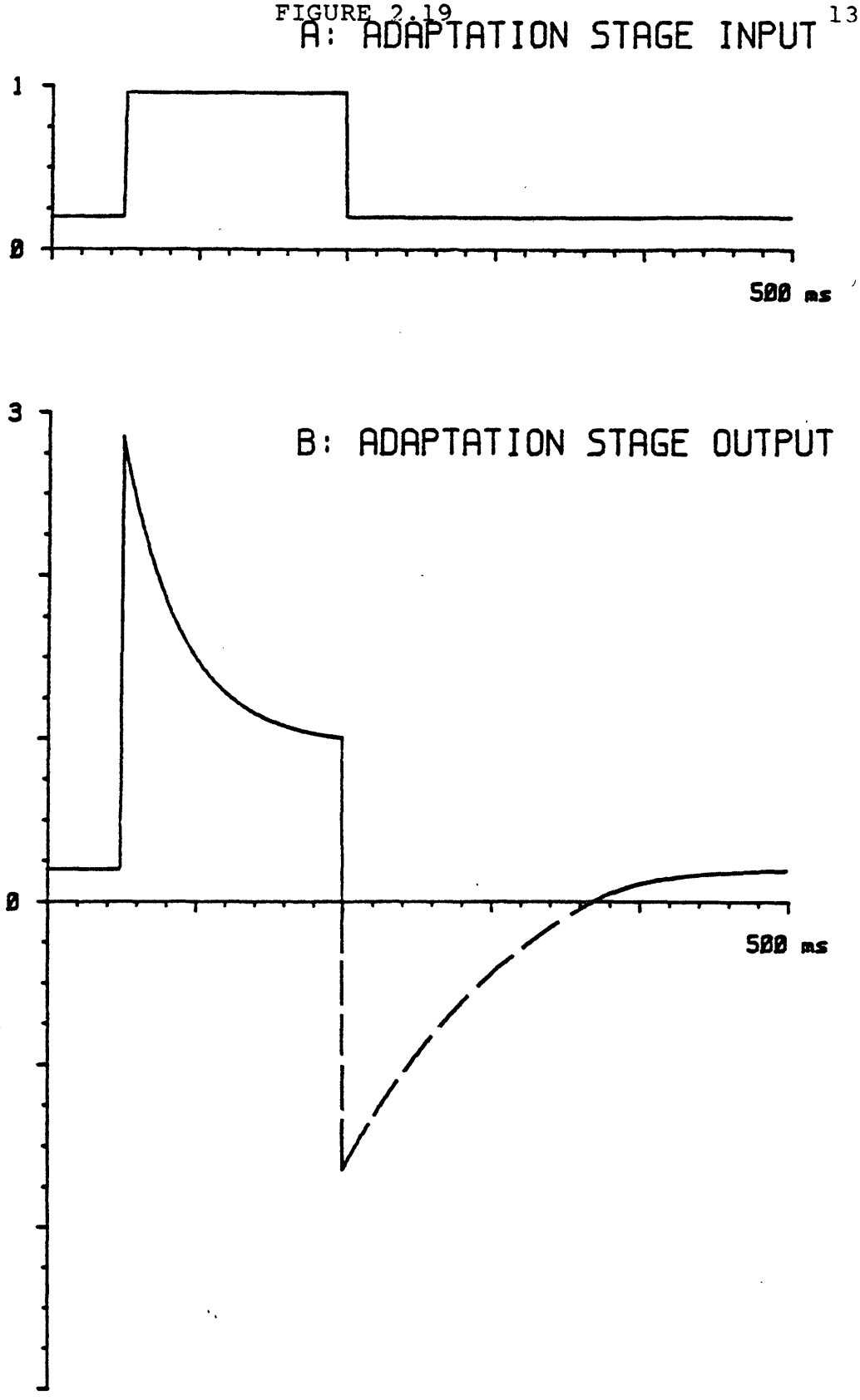

A: ADAPTATION STAGE INPUT
1
0
500 ms
3
B: ADAPTATION STAGE OUTPUT
0
500 ms

FIGURE 2.20

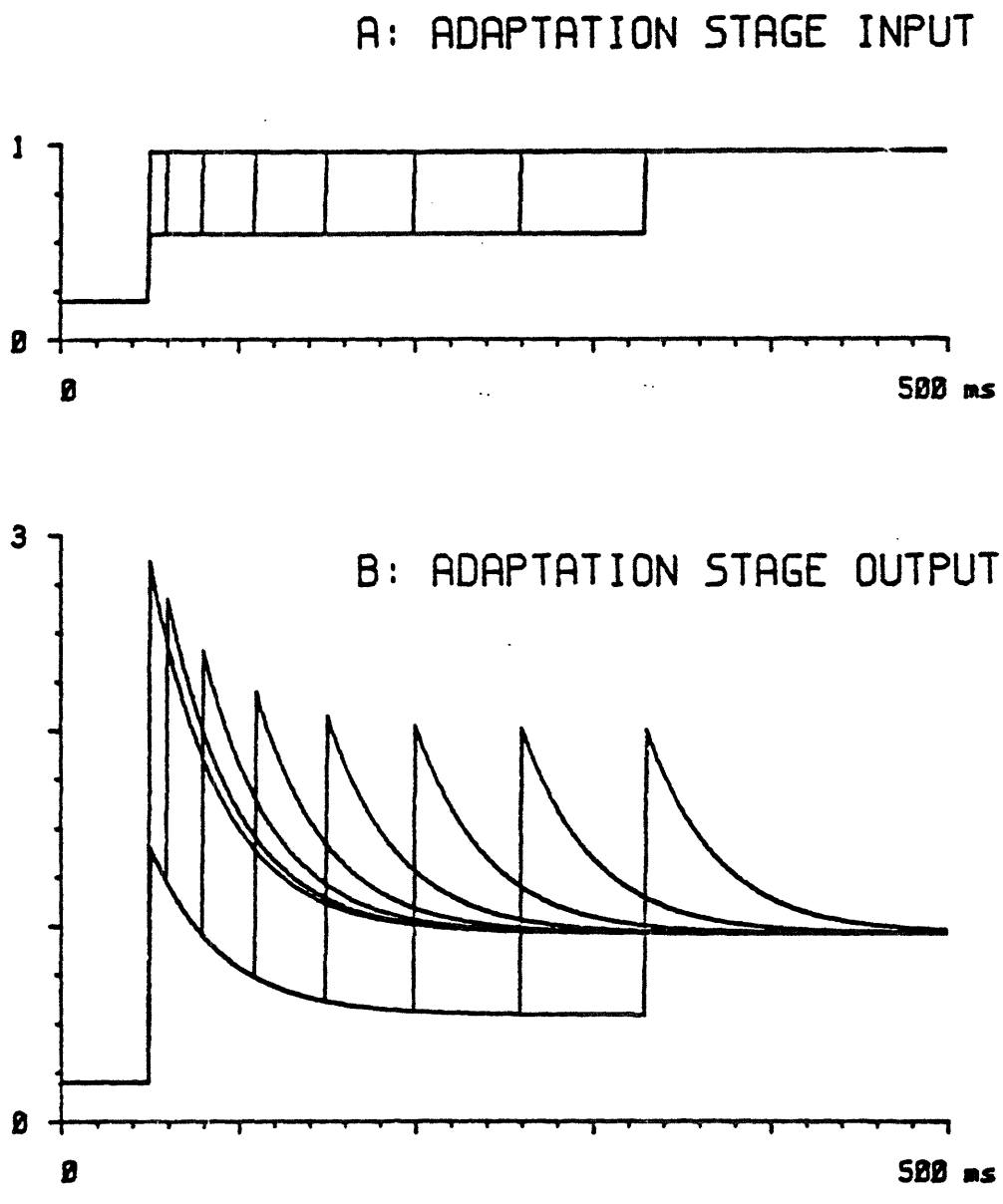
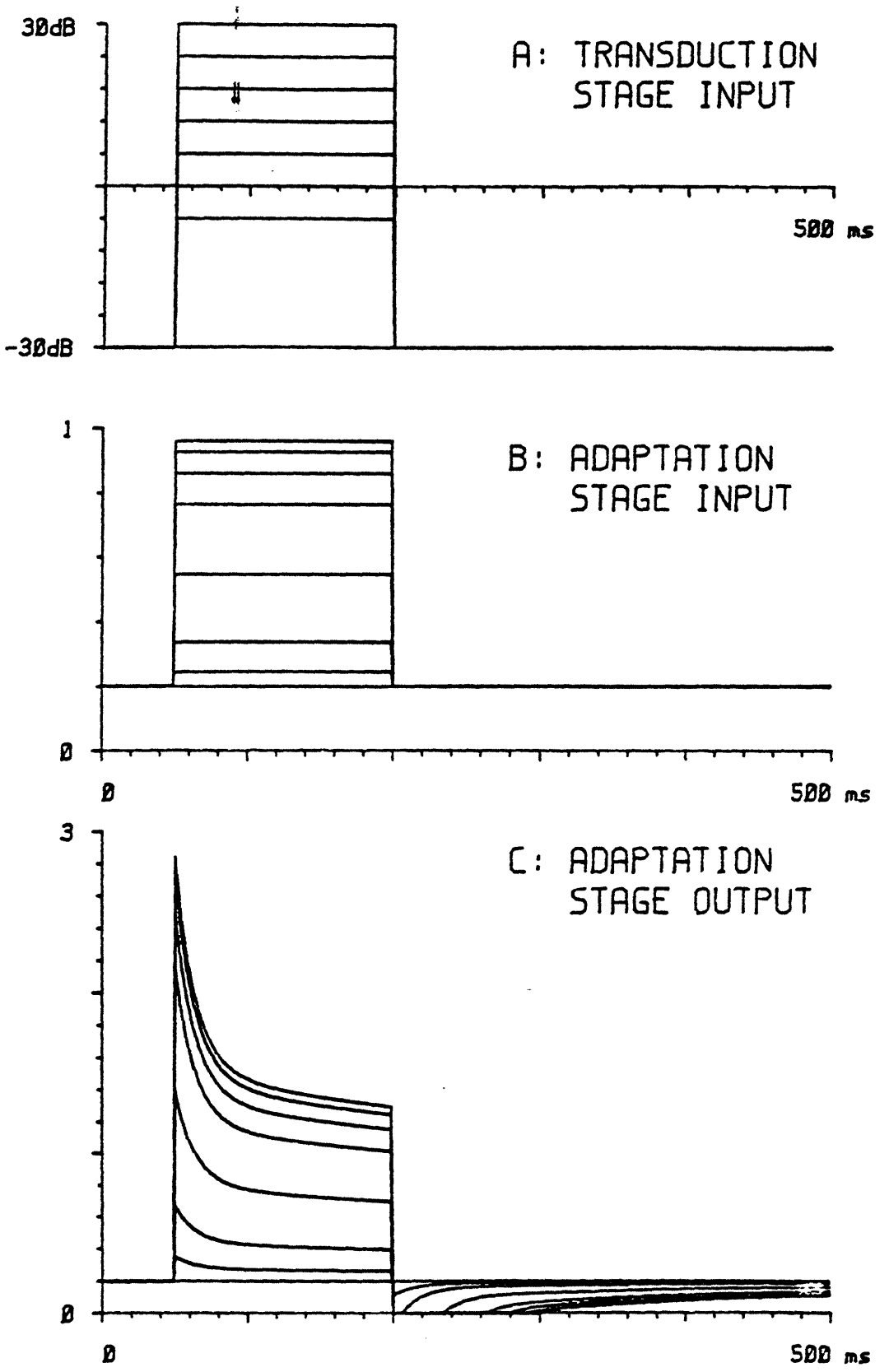
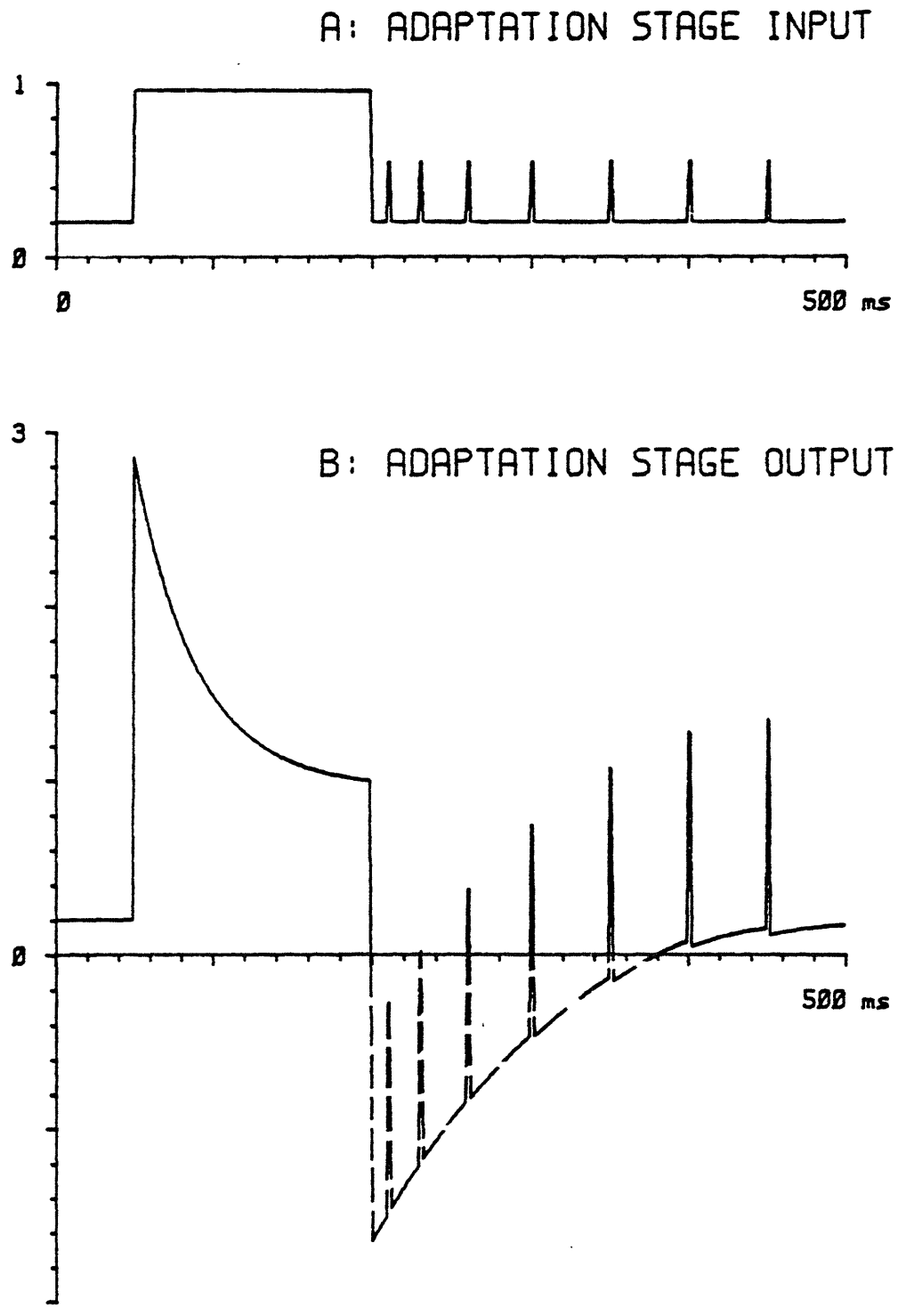
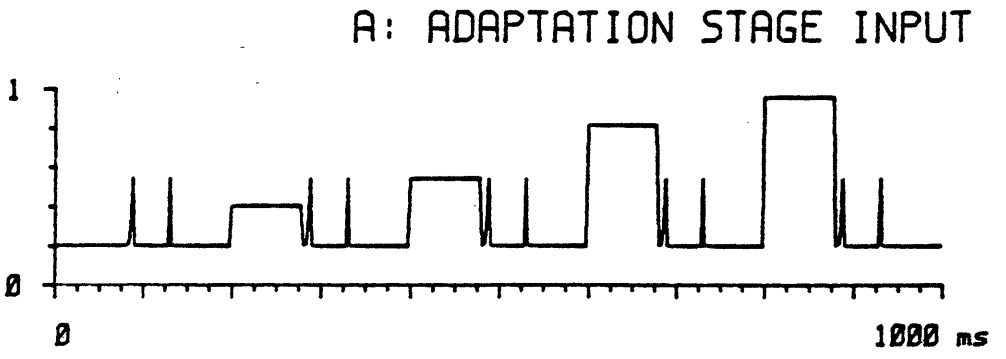
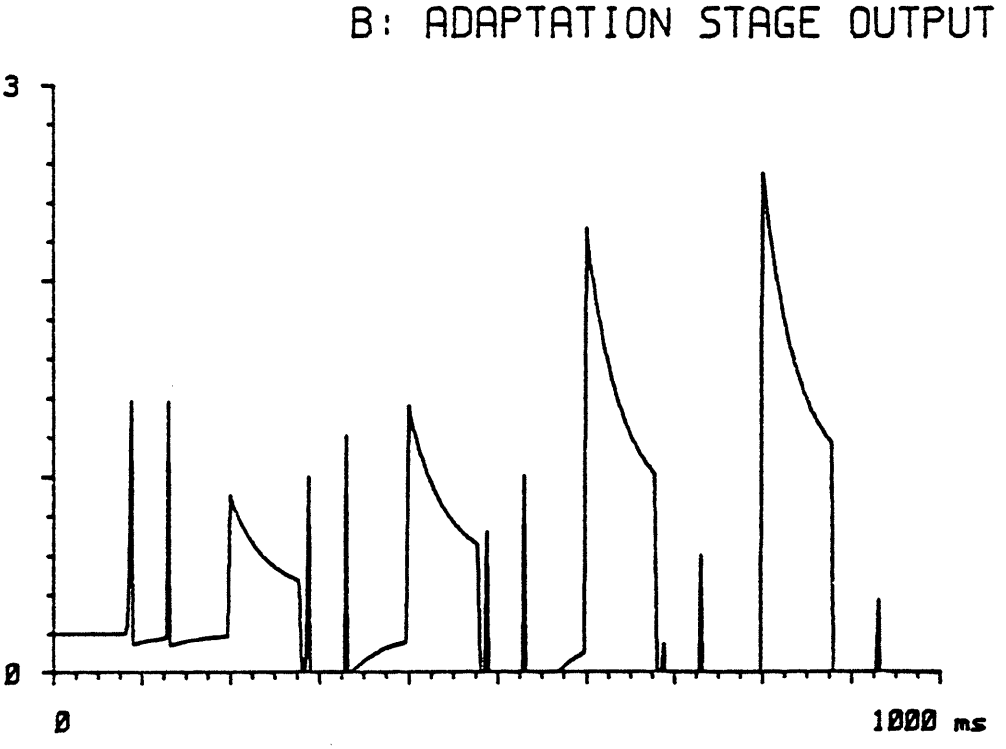

A: ADAPTATION STAGE INPUT
1
0
0
500 ms
3
B: ADAPTATION STAGE OUTPUT
0
0
500 ms

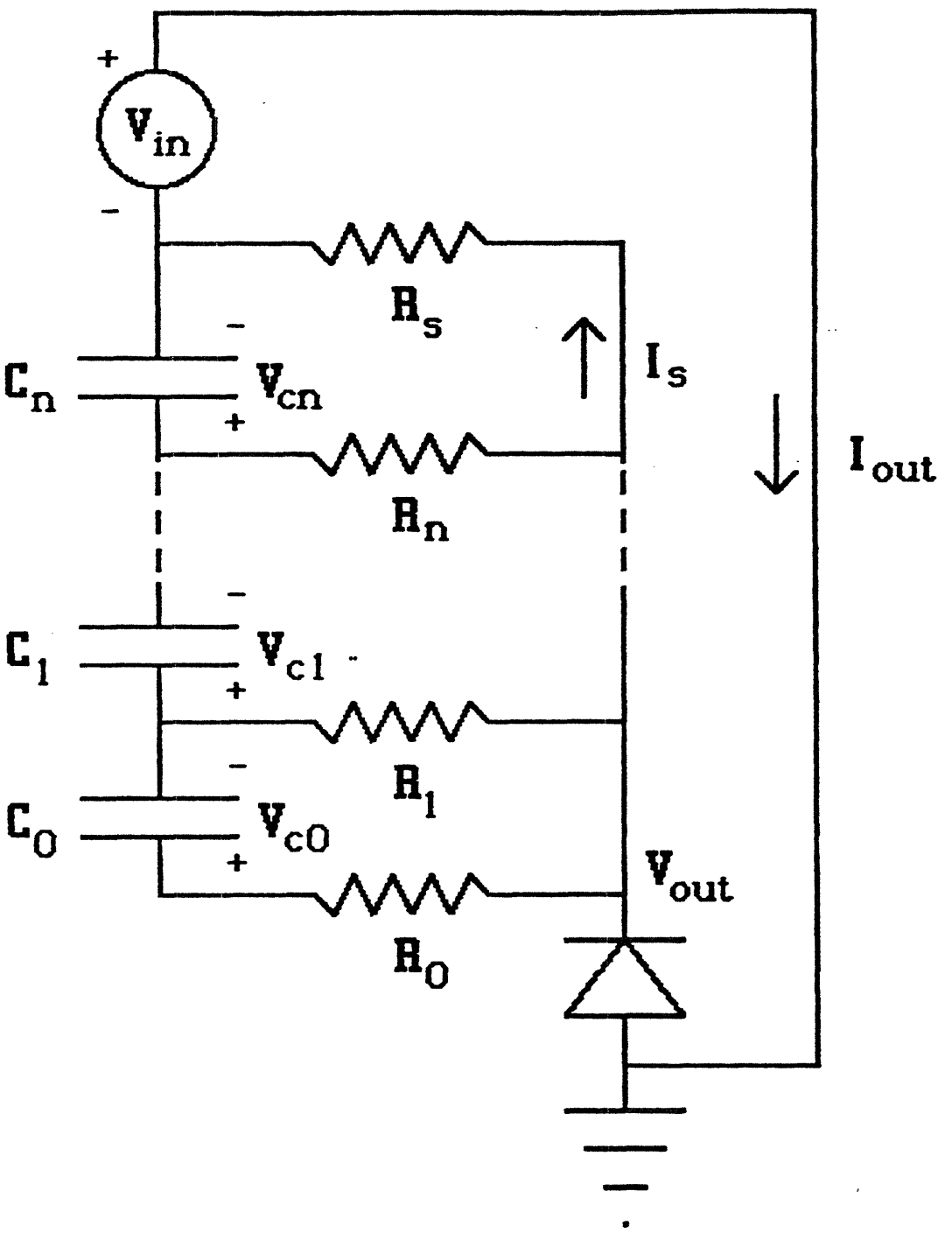
FIGURE 2.21


30dB
-30dB
500 ms
A: TRANSDUCTION
STAGE INPUT
1
0
0
500 ms
B: ADAPTATION
STAGE INPUT
3
0
0
500 ms
C: ADAPTATION
STAGE OUTPUT

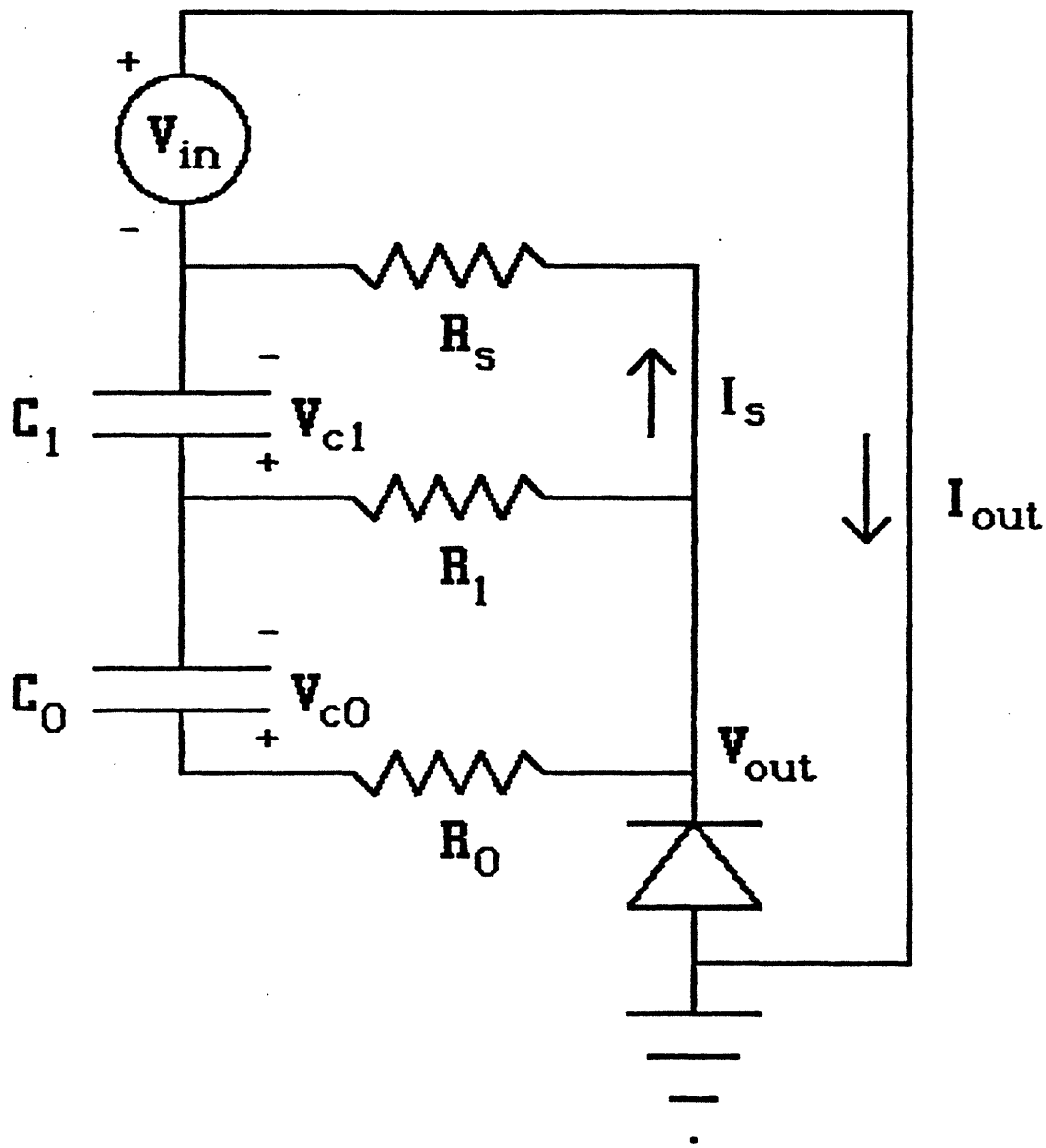

A: ADAPTATION STAGE INPUT
1
Ø
Ø
500 ms
B: ADAPTATION STAGE OUTPUT
3
Ø
500 ms


A: ADAPTATION STAGE INPUT
1
0
0
1000 ms


B: ADAPTATION STAGE OUTPUT
3
0
0
1000 ms



Multiple Time Constant Adaptation Model



Double Time Constant Adaptation Model

FIGURE 2.26

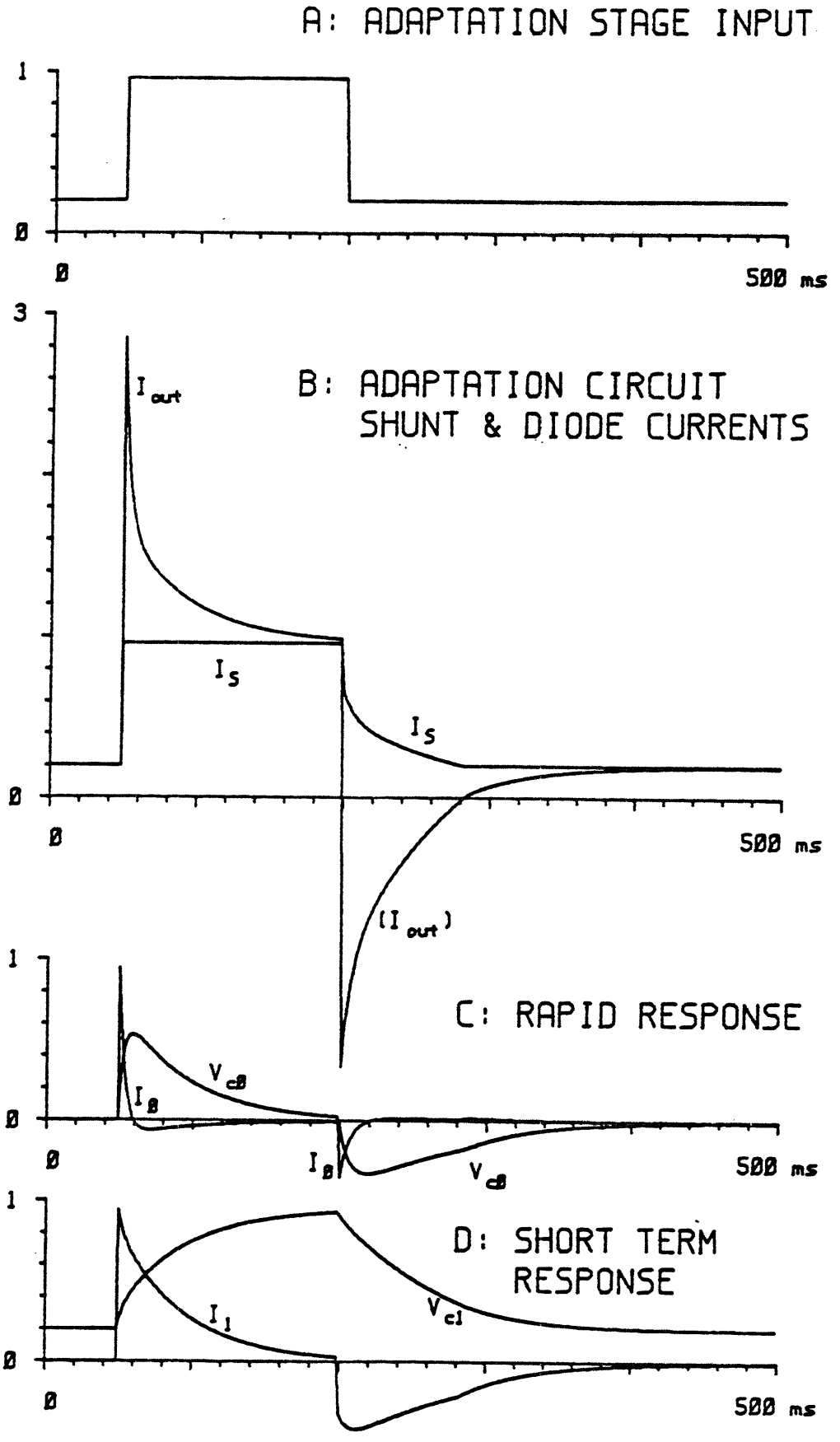
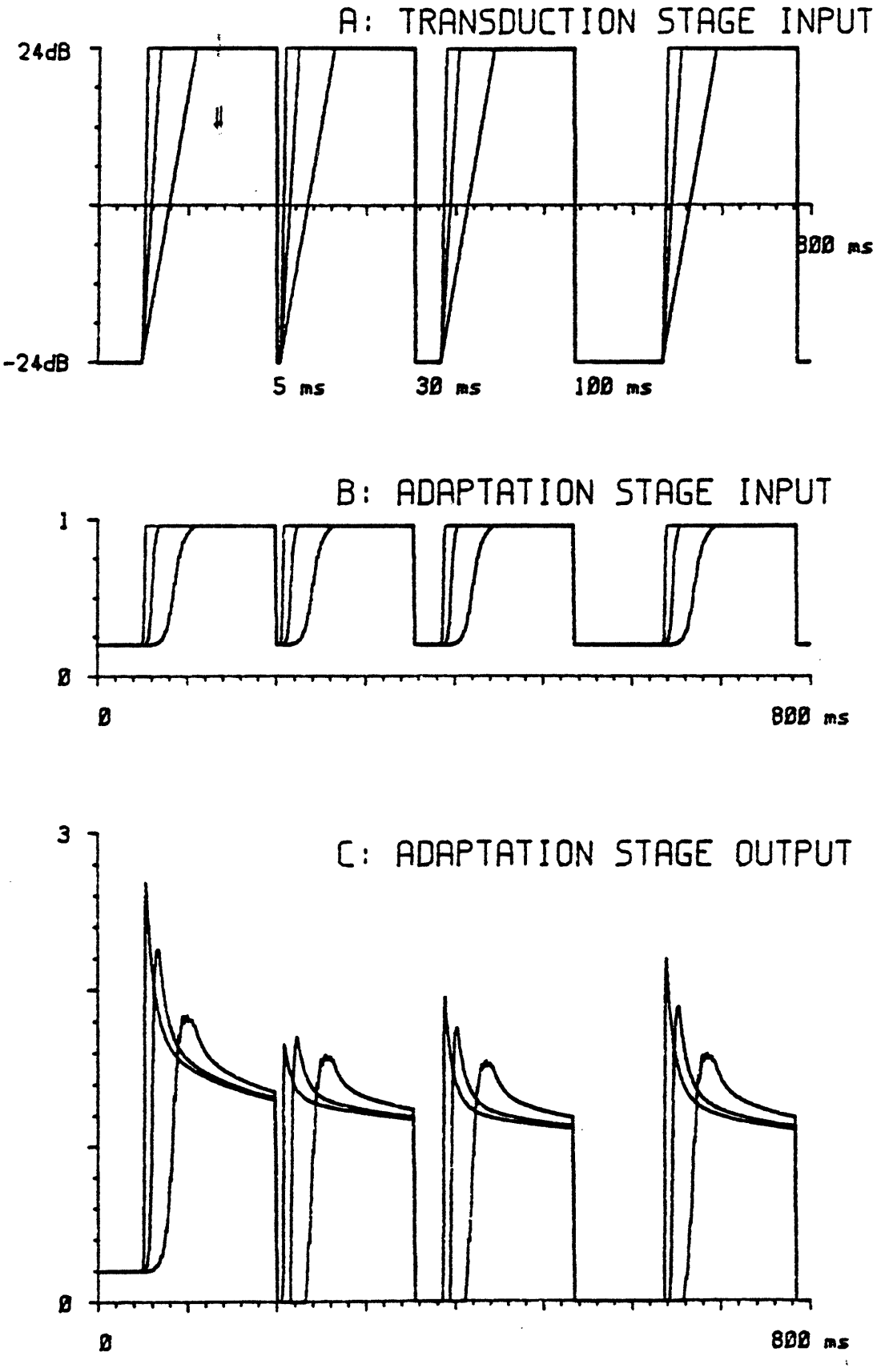

A: ADAPTATION STAGE INPUT
1
Ø
Ø
500 ms
3
B: ADAPTATION CIRCUIT
SHUNT & DIODE CURRENTS
I_{out}
I_S
I_S
Ø
Ø
500 ms
(I_{out})
1
C: RAPID RESPONSE
$V_{cØ}$
$I_Ø$
Ø
Ø
$I_Ø$
$V_{cØ}$
500 ms
1
D: SHORT TERM
RESPONSE
I_1
V_{c1}
Ø
Ø
500 ms

FIGURE 2.27


A: TRANSDUCTION STAGE INPUT
24dB
-24dB
5 ms
30 ms
100 ms
800 ms
B: ADAPTATION STAGE INPUT
1
0
0
800 ms
3
C: ADAPTATION STAGE OUTPUT
0
0
800 ms

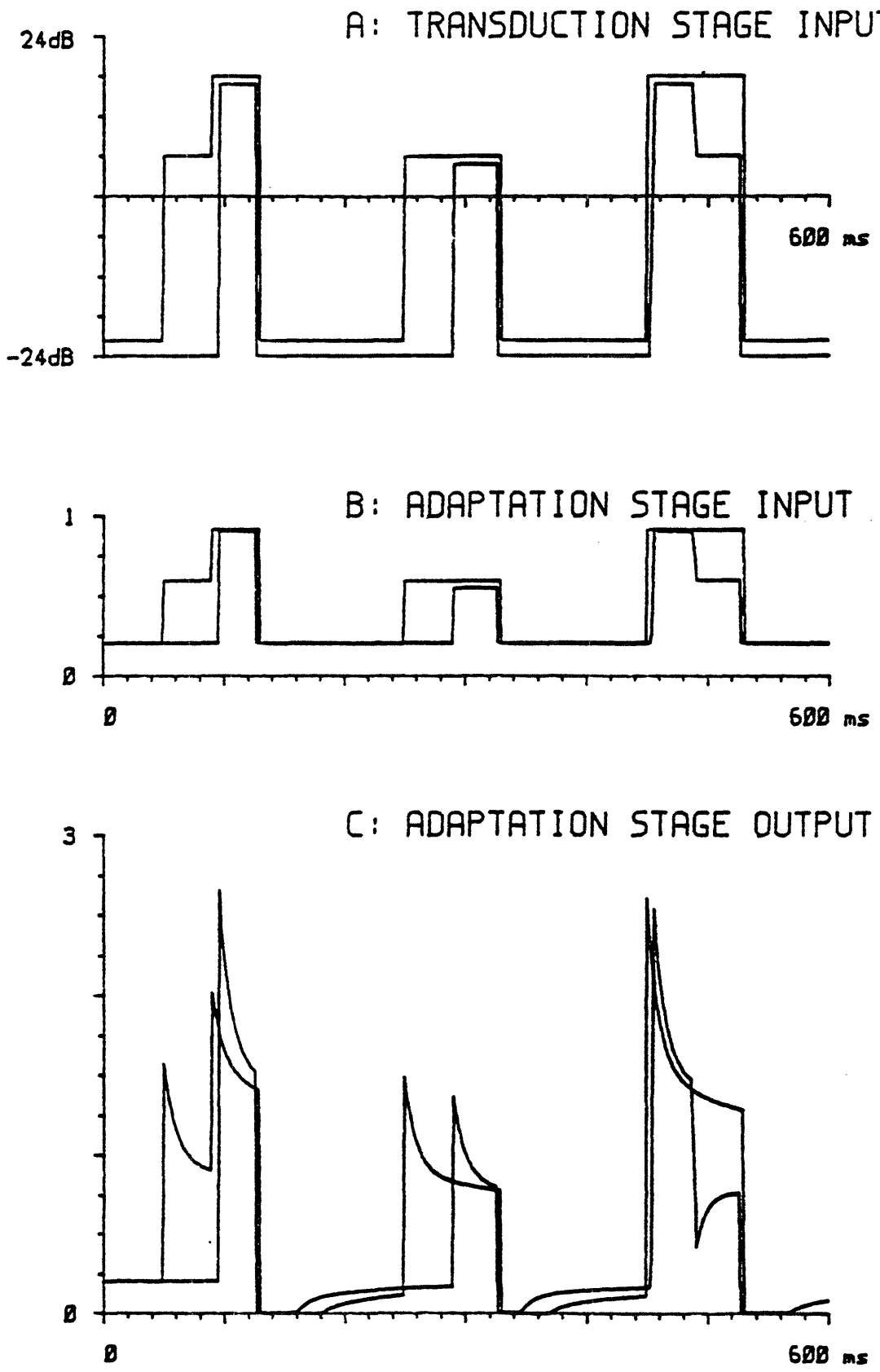

A: TRANSDUCTION STAGE INPUT
24dB
-24dB
600 ms
B: ADAPTATION STAGE INPUT
1
0
0
600 ms
C: ADAPTATION STAGE OUTPUT
3
0
0
600 ms

FIGURE 2.29

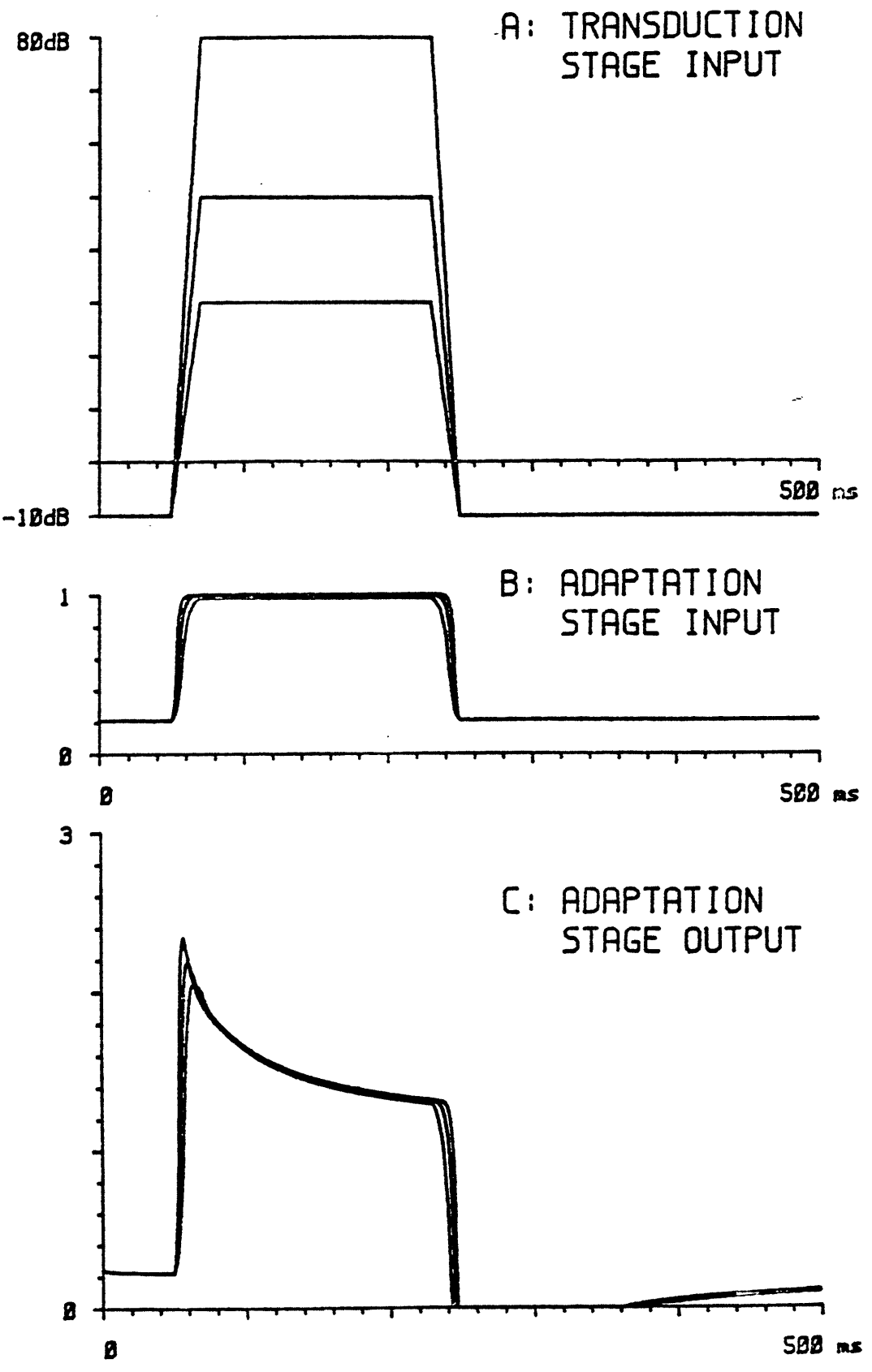
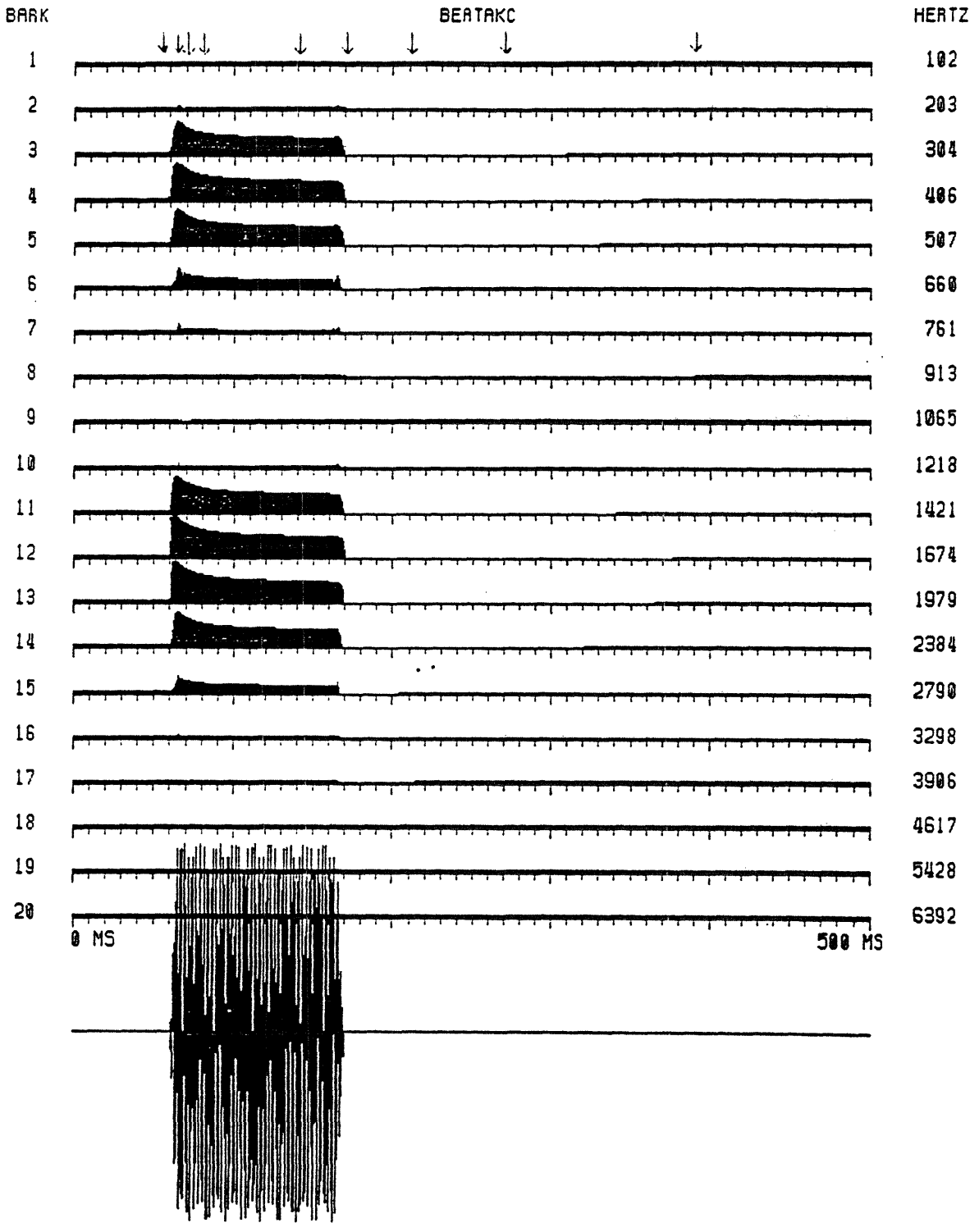
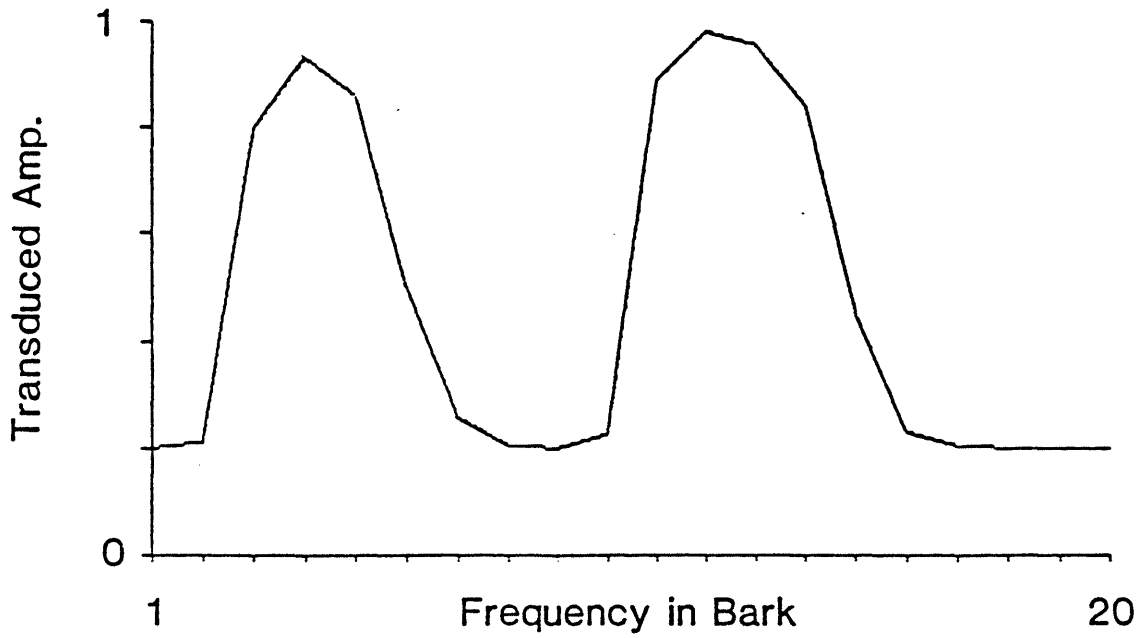

A: TRANSDUCTION STAGE INPUT
80dB
-10dB
500 ms
B: ADAPTATION STAGE INPUT
1
0
0
500 ms
C: ADAPTATION STAGE OUTPUT
3
0
0
500 ms

FIGURE 2.30


BARK
BEATARC
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

A: Transduction Stage Output



B: Adaptation Stage Output

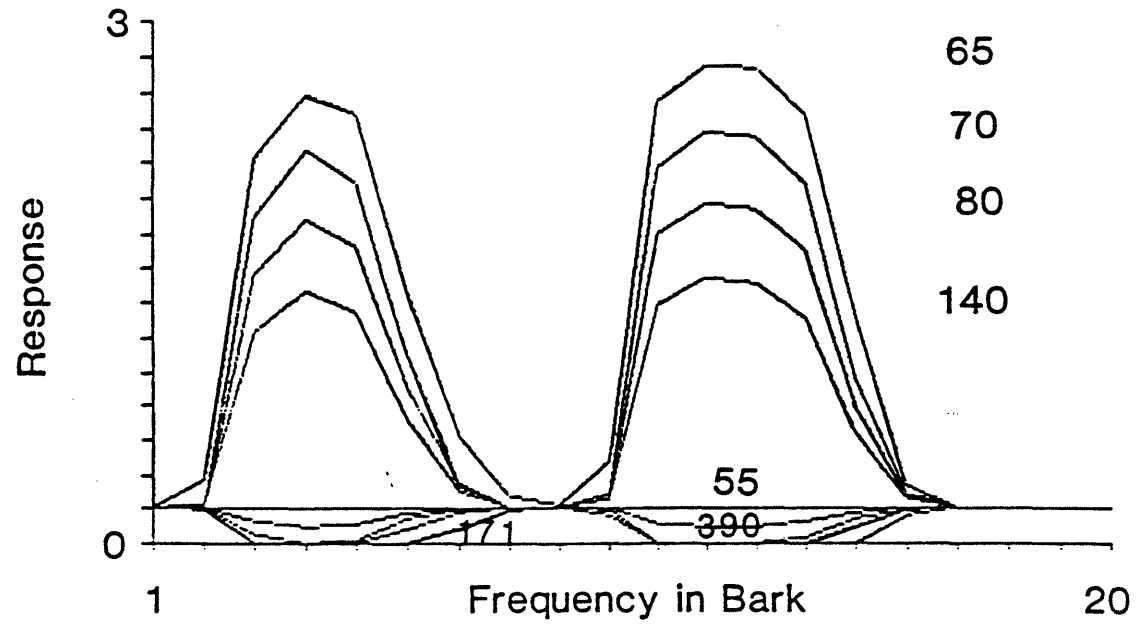
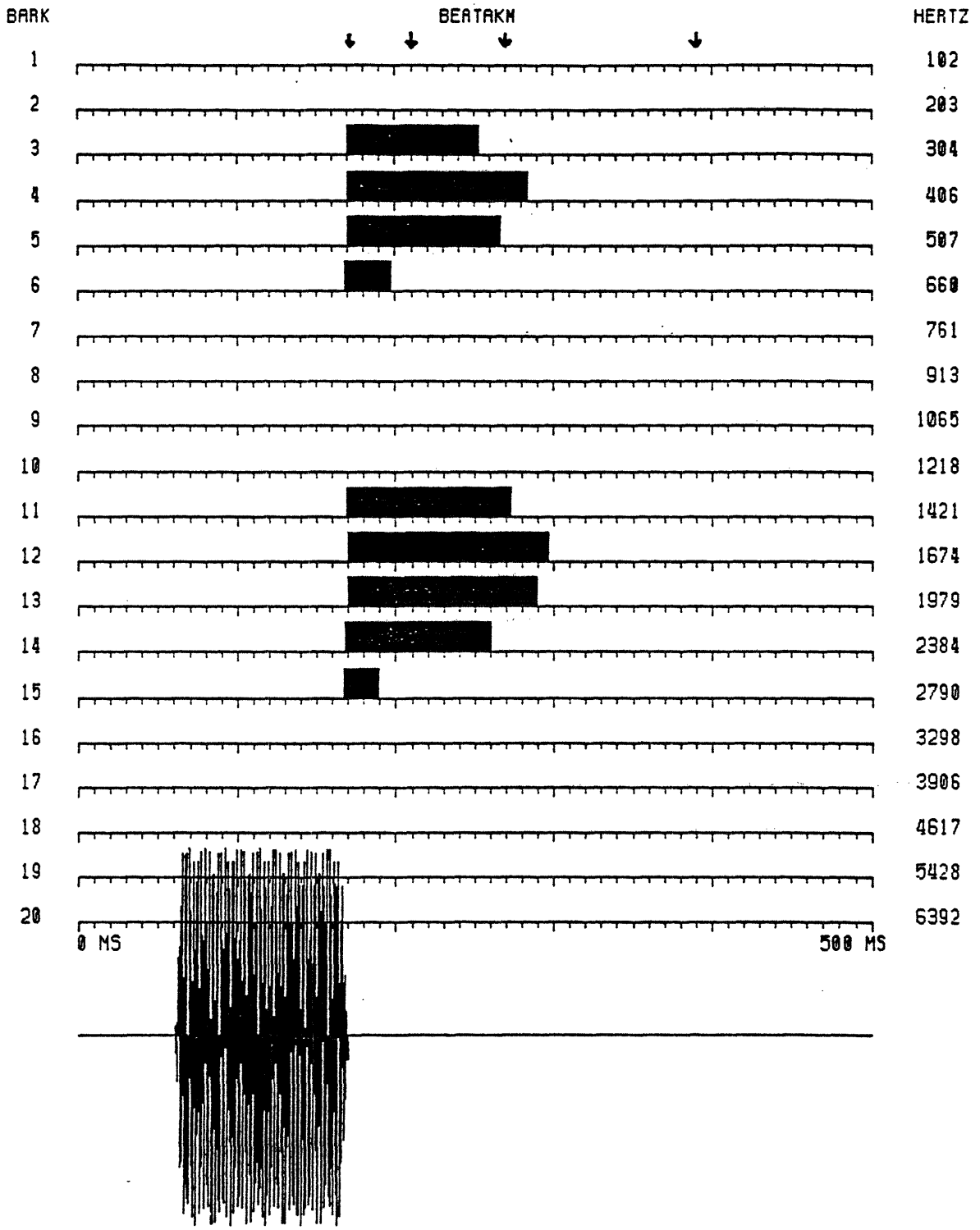


FIGURE 2.32


BARK
BEATAKM
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

CHAPTER 3

PAM RESPONSE PATTERNS AND PHONETIC DISTINCTIONS

3.1 INTRODUCTION: USING THE PAM TO STUDY PERCEPTUAL RESPONSES TO SPEECH

This chapter examines in detail the response patterns of the Peripheral Auditory Model to two sets of speech signals. A simple property of the PAM response to each set is shown to be correlated with listeners' perception of a phonetic contrast exhibited between elements within the set. The phonetic contrasts in question are, in the first set of stimuli, the distinction between a word-initial /b/ and /w/ (a STOP/GLIDE distinction), and in the second set of stimuli, the distinction between an intervocalic /b/ and /p/ (a VOICED/UNVOICED distinction). Although only a single phonetic contrast is perceptible within each set, the stimuli in the sets vary acoustically in two or more ways. The fact that multiple acoustic variations such as these may produce a single phonetic contrast has been interpreted as evidence that the relationship between acoustic properties and phonetic features is context dependent. Many such acoustic "trading relationships"--multiple acoustic properties, combinations of which can be varied to produce an identical phonetic contrast--have been found. Their variety and frequency provide convincing evidence that no simple one-to-one relation-

ship exists between such acoustic properties of speech waveforms as formants, stop gap durations, fundamental frequency, etc., and the distinctive features which efficiently characterize speech at the phonological and perceptual level.

If the correspondence theory described in the Introduction is correct, we might expect that in the peripheral auditory system some of these complexes of acoustic properties are merged into a single auditory property that bears a simpler relationship to its phonetic counterpart. This would constitute a "decoding" of the speech signal by the peripheral auditory system. Another form of decoding would be the creation of a measurable property in the PAS response that represents an abstract property of the acoustic response. For instance, in both of the sets of stimuli examined in this chapter, temporal properties of the acoustic signal, such as closure duration and speaking rate, are important elements in an acoustic phonetic account of listeners' perceptions. The PAS can be said to participate in the decoding of speech if its response patterns contain a measurable auditory property, such as average firing rate, which reflects these abstract temporal properties. This is especially so if it can be shown that the auditory transformation of the property is particularly appropriate, or apt: for instance, if the PAS converts a nonlinear relationship between an acoustic property and a phonetic feature into a linear relationship between an audi-

tory property and the same phonetic feature.

In using the PAM to model peripheral auditory response to speech, and using various postprocessors to measure PAM response properties, values must be assigned to a number of parameters. These parameters determine filter bank characteristics; quiesent (spontaneous) response rate; rapid, short term, and long term adaptation times; the level of firing below which channels are considered to be in a "masked" state; etc. It is reasonable to wonder about the extent to which the results presented in this chapter depend on the exact parameter values chosen. No formal study was made of the sensitivity of the PAM response to various parameter values. However, a few informal observations can be made.

First, a number of parameters are interrelated. Consider, for example, the duration of time during which a channel is in a masked state following the offset of stimulation. This duration is dependent on the selected quiescent response rate, the adaptation time constants chosen, and the masking threshold level used, among other parameters. For a particular stimulus, there are an infinite number of possible sets of parameter values which will result in the same masking duration. Therefore the validity of the results to be presented do not depend on all of the chosen parameter values being "correct": just that their combination is representative of the auditory system.

Secondly, many of the values can be established within reasonable limits from psychoacoustic and physiological data. The quiescent response rate and adaptation time constants are examples of parameters for which such data is available.

Thirdly, we are restricting our analysis goals to demonstrating that there is a correlation between PAM response properties and phonetic features. This is a less rigorous goal than trying to directly calculate perceptual responses from PAM output signals. If a correlation is found, its existance is likely to be relatively insensitive to the exact parameter values chosen.

Fourthly, although no formal study was made of the effect of varying parameter values, a number of different parameter values were used in informal pilot runs. No extraordinary changes in PAM response properties were observed over the range of parameter values used.

Nevertheless, a systematic investigation of the effect of changing PAM parameter values would be useful, as would further attempts to determine optimal values from psychoacoustic and physiological data.

3.2 STOP/GLIDE EXPERIMENT

This experiment involves a set of 55 synthesized CV syllables perceived by listeners as either "ba" or "wa". The tokens were created and analyzed by Landahl and Maxwell (1983). The original experiment and its results will be described first, and then the response of the PAM to the stimuli. A simple auditory analysis will be proposed, and a statistical comparison will be made of the traditional acoustic phonetic analysis and the alternative auditory phonetic analysis suggested by the PAM response patterns.

3.2.1 The original experiment

3.2.1.1 Background

The Landahl and Maxwell study was a replication of a study by Miller and Liberman (1979), in which it was found that overall syllable duration affected the boundary duration of initial formant transition across which listeners' perceptions changed from hearing a stop (/ba/) to hearing a glide (/wa/). To quote Miller and Liberman:

> The combined data ... clearly indicate that, as syllable duration increases, there is a perfectly regular change in the way our subjects identified the [synthetic speech stimuli] as [ba] and [wa]: the longer the duration of the syllable, the longer the duration of transition needed in order to hear [wa].

The Landahl and Maxwell replication used synthetic stimuli that the authors believed would make the tokens more realistic than those used by Miller and Liberman.

3.2.1.2 Experimental Procedure

The stimuli were created using the Klatt (1980) cascade/parallel formant synthesizer. Figure 3.1 shows the way in which the important synthesis parameters were varied to create a continuum of tokens between a distinct initial labial stop and a distinct initial labiovelar glide. Each token has an initial prevoicing and transition region followed by a steady state region. During the initial region both the frequency of the first and second formant, and the amplitude of voicing, vary. For all tokens the initial value of F1 is 230 Hz, and the final value is 770 Hz. F2 is set at 620 Hz initially and rises to 1230 Hz. F3 is held constant at 2860 Hz for all tokens, and the fundamental frequency is held constant at 114 Hz for all tokens. The amplitude of voicing rises during the initial interval from a level 10 dB below its steady state value, and falls again by 15 dB during the final 20 msecs before the end of the syllable.

The stop/glide continuum was created by varying the stimuli in two ways. First, the duration of the initial formant transitions varies in 5 msec intervals between 15 msecs for the most /b/-like tokens and 65 msecs for the most /w/-

like tokens. (The amplitude of voicing has the same rise time, but begins and ends 15 msecs earlier than the formant transitions.) Second, the duration of the steadystate portion of the syllable was varied independently of the onset times. Syllables having a total duration of 87, 123, 158, 228, and 299 msecs were created. For each syllable duration, tokens were created with each of the 11 rise times, resulting in a total of 55 stimuli.

The bandwidths of the formants, higher formant frequencies, and other synthesis parameters were adjusted to be appropriate for a male speaker saying the vowel /a/. The biggest difference between the Landahl and Maxwell stimuli and the Liberman and Miller stimuli was created by adjusting the parameter which controls the low-pass cut-off frequency of the voice source during the initial prevoicing interval. The initial value of this parameter was varied in logarithmic increments between 562 Hz for the most /b/-like tokens and 5000 Hz for the most /w/-like tokens. For all tokens, this parameter rose to 5000 Hz after the end of the prevoicing. According to Landahl and Maxwell,

> [This adjustment] interacts with amplitude of voicing such that for the /w/ endpoint the effect was one of strong prevoicing and a gradual increase in amplitude over the formant transitions, while for the /b/ endpoint the prevoicing was much weaker, as is appropriate for a stop consonant, and amplitude of voicing increased very abruptly, also appropriate for the release of a stop.

Figure 3.2 shows the waveforms of the tokens with the 15 msec rise time and the 60 msec rise time. Each of the five syllable durations are shown for both the fast and slow rise time tokens. (The tokens with the 65 msec rise times have an unexplained ripple in their amplitude envelope. Perceptually, they appear to be good exemplars of /wa/, and are included in all the statistical calculations. However, because of their visual anomaly, tokens with 60 msec rise times will be used in figures to represent the slow-onset endpoint.)

Three test tapes were constructed, each containing eight repetitions of each stimulus, for a total of 440 stimuli per tape. The tapes differed only in their random ordering of the tokens.

Ten paid subjects listened to the tapes. Eight of the subjects were male, two female. All of the subjects were college students, and none reported any history of speech or hearing disorders. Each subject listened to each tape. The subjects were told to judge each stimulus and respond with either /b/ or /w/: no other responses were allowed.

3.2.1.3 Results and Analysis

The average percent of stop responses for all subjects, for each of the 55 tokens, is shown in Table 3.1. The same information is shown in graphical form in Figure 3.3. It can

be seen that for abrupt transitions (rise time less than or equal to 20 msecs) syllables of any duration are interpreted as beginning with a stop. Similarly, syllables with rise times of 55 msecs or more are consistantly interpreted as beginning with a glide. However, in the intermediate region of 25 to 50 msec rise times, the percentage of stop judgements depend on the syllable duration: the longer the syllable, the more likely it is that a particular rise time will be perceived as stop-like.

The rise time associated with the phonetic boundary between STOP and GLIDE, as a function of syllable duration, can be calculated by converting the percent stop judgements to normalized z scores and fitting a linear regression line to those scores. (This conversion assumes a psychophysical categorical decision model (e.g., Durlach, 1968) in which the percentage of stop responses is assumed to depend on the difference between the value of an internal perceptual variable, which is affected by Gaussian noise, and a categorical criterion or boundary value for the same variable.) The interpolated rise time duration which results in a z score of zero is then taken to be the boundary duration. In order to avoid unduly weighting the regression line by results far removed from the boundary duration, z scores of less than -2.0 or more than 2.0 (percentages above 97 or below 3 percent) are not included in the linear regression calculation.

The resulting rise time boundaries are listed in Table 3.1, and plotted as a function of syllable duration in Figure 3.4. The boundary rise time appears to be a nonlinear function of syllable duration, with a lower limit of 27.5 msecs for 87 msecs syllables and an upper limit of 37.6 msecs for 299 msecs syllables. The corresponding data from Miller and Liberman show a greater syllable duration effect: 80 msec syllables had a boundary duration of 31.9 msecs, and 296 msec syllables had a boundary duration of 46.6 msecs. In both experiments the boundary duration is sensitive to syllable duration for short syllables, but lengthening the syllable beyond 200 msecs appears to have little effect.

Landahl and Maxwell's acoustic phonetic analysis of these tokens suggests that the phonetic feature [-continuant] is associated with an abrupt change in amplitude. Since /b/ has the feature [-continuant], and /w/ has the feature [+continuant], rise times which are fast enough to be considered abrupt signal the stop consonant. To the extent that acoustic properties which cue other phonetic differences between /b/ and /w/ are ambiguous, perceptual judgements appear to depend on the auditory system's interpretation of the abruptness of the onset. Presumably the duration of the steady state portion of the syllable serves to calibrate the onset time: the longer the syllabic nucleus, the slower the speaking rate, and the more abrupt a particular onset time will appear. If the

onset is very quick or very slow, no such normalization is necessary. But if the onset time is in the intermediate range between 25 and 50 msecs, the perceived rate can influence the perceptual judgement.

This acoustic phonetic analysis is straightforward and consistent, but it does leave some questions unanswered:

* Why are the "quick" and "slow" endpoint transitions around 25 and 50 msecs rather than some other values?
* Why is the relationship between boundary duration and syllable duration a nonlinear one?
* Why does increasing the syllable duration beyond 200 msecs have little effect?

3.2.2 The PAM Response

In this section we will examine the response of the PAM to the stop/glide stimuli. We will attempt first to find measurable properties of the PAM response pattern which reflect the abstract acoustic properties of rise time and syllable duration, and then study the correlation between these properties and listeners' perceptual judgements. Again, rise time and syllable duration are considered abstract because they are not values of the acoustic signal, or any equivalent representation, but rather are dependent on a higher level waveform segmentation process.

Much of our attention in this section will be concentrated on the onset and offset characteristics of the PAM in response to these stimuli. Let us define the onset response to be the response of the PAM in the first 20 msecs following any significant increase above the spontaneous firing rate as a stimulus begins. This length of time is long enough to encompass the first two glottal pulses of our stimuli that are strong enough to be well within the dynamic range of the model. Let us define the offset response to be the response of the PAM to the last full glottal pulse before the 20 offset of a stimulus. Although these are informal definitions, in practice it is always easy to identify the location of these regions in the response pattern of the PAM. The strength of the onset or offset response (within a particular channel) is defined to be the maximum response signal value obtained within the onset or offset interval.

3.2.2.1 PAM Processing Procedure and Parameters

The stop/glide stimuli were digitized at 12,987 samples per second from audio tape (sample period = 77.000 usecs). The audio signal was low-pass filtered at 6,400 Hz before digitizing. The digital signals were preemphasized as described in Chapter Two, and then processed by the spectral analysis stage. The nonlinearity used in the transduction stage was the DC transfer function of a raised hyperbolic

tangent function, with a bias parameter of -1.1. A three time constant adaptation stage was used, with time constants of 15, 40, and 320 msecs, contributing equally to the transient response. The adaptation ratio was 3.5.

3.2.2.2 PAM Response Patterns

Figure 3.5 shows the response of the PAM to the stop/glide stimulus with the fastest onset time (15 msecs) and the longest syllable duration (299 msecs). The periodic pulses in the PAM output are the model's response to the glottal pulses in the input. Because of the rapid increase in amplitude, the PAM's onset response in the F1 and F2 region (channels three through eleven) is stronger than its steady state response in the same region. This is due primarily to adaptation, since the acoustic amplitude of the onset is less than or equal to the amplitude of the vowel. For channels three through five, and eight and nine, the formant transitions add to this onset-peak effect: within the onset region the formants move through and then beyond the center frequencies of these channels. These formant transitions can be seen in the figure. In summary, this token sounds like /ba/ (98 percent stop responses), and has an onset response that is stronger than its offset response.

Figure 3.6 shows the response of the PAM to the stop/glide stimulus with the same onset time as the previous

figure (15 msecs) but the shortest syllable duration (87 msecs). The steady state portion of this syllable is only three glottal pulses long. The onset, however, is still fast enough that the envelopes of the PAM channels in the F1 and F2 regions, where most of the energy is, show an initial peak followed by a steady decline. This token also sounds like /ba/ (100 percent stop responses), and has an onset response that is stronger than its offset response.

Figure 3.7 shows a response pattern that looks somewhat different. The stimulus in this figure is the token with a 60 msec onset and a 299 msec duration. The gradual increase in the amplitude of voicing in the stimulus results in a PAM onset response that is smaller in almost all of the channels than the PAM response at the beginning of the steady state portion of the token. Even 200 msecs after the beginning of the steady state region, when most of the potential adaptation has occurred, the firing rate is still higher in most channels than during the onset. This token sounds like /wa/ (1 percent stop responses), and has an onset response that is weaker than its offset response.

Figure 3.8 shows the PAM response to a token with a 60 msec onset and an 87 msec duration. The combination of a slow onset and a short syllable duration results in an offset response that is much stronger than the onset response in all channels containing significant energy. Once again, this

token is perceived to begin with a clear glide, eliciting no stop responses at all from the experimental subjects.

In the previous four figures, syllable duration appeared to have little effect on either perceptual responses or the relative sizes of onset and offset responses. This is because the onsets in these figures were either very fast or very slow. Figures 3.9 and 3.10 represent the PAM responses to stimuli with intermediate transition and amplitude rise times (30 msecs). For these tokens, syllable duration has much more of an effect. Figure 3.9 shows the PAM response to a token with a 299 msec duration, which is perceived to be strongly stop-like (94 percent stop responses), while Figure 3.10 shows a token with an 87 msec duration, which is perceived to be generally glide-like (17 percent stop responses). In the region of the lower formants for these patterns, the glottal pulse leading to the largest response peak occurs one or two glottal pulses later than for stimuli with the most rapid onset. This tends to place the peak response very close to the offset of the shortest duration syllable, so that the offset response (the pulse near 120 msecs) for channels in the F1 and F2 region is approximately the same size as the onset response. In contrast, when the steady state portion of the vowel is allowed to continue for 200 msecs or more, substantial adaptation occurs and the relative size of the last full pulse (near 330 msecs) is relatively weak.

3.2.2.3 A Possible Auditory Analysis

These informal observations regarding the relative sizes of onset and offset response peaks suggest a possible measure of "onset strength" as a property of the peripheral auditory response to speech. An abrupt onset of acoustic energy should result in an auditory response that is relatively strong compared with the response to the stimulus following the onset. A simple postprocessor was designed to provide a quantitative measure of onset strength for the current set of stimuli. The first step in the postprocessing was to produce a smoothed composite single-channel waveform from the multi-channel PAM response. This was done by using a 20 msec half hamming window to smooth each channel of the PAM response. The left half of the hamming window was used, so that the non-zero portion of the smoothing window included the current response and the response during the previous 20 msecs, with most of the weight given to responses within the last 10 msecs. At each point in time, the smoothed values were averaged over all twenty PAM channels, so that the output value represents the expected whole-nerve firing rate, averaged across all frequencies, and smoothed with a 20 msec window.

The second step in the postprocessing was to select particular glottal pulses to represent the onset and offset response. The glottal pulses chosen were defined as the second significant glottal pulse in the onset response, and

the last full glottal pulse in the offset response. The locations of these pulses, which always occur at the same time in stimuli of the same length, are listed in Table 3.2. A computer program was written to extract the largest value attained by the smoothed composite PAM response signal within the temporal limits established by Table 3.2. In all cases a glottal pulse was well within the limits, and corresponded to a subjective choice of the correct onset or offset pulse.

Thus for each stimulus we have derived a quantitative measure of the strength of the onset and offset of the peripheral auditory response. Notice that we do not have to measure time--either rise time or syllable duration--in any direct way. Instead the PAM itself provides a measure of time, in that the size of its response is dependent on adaptation behavior which is in turn dependent on time in a particular way.

Figure 3.11 shows a set of five stop/glide signals. The acoustic waveform is shown on the left, and the smoothed composite PAM response on the right. The top signal is perceived to be strongly /ba/-like for all syllable durations, while the bottom signal is perceived to be strongly /wa/-like for all durations. The three middle signals are perceived as being more or less /ba/-like depending on the syllable duration. Note that the glottal pulses are still clearly visible in the composite signal. The vertical lines indicate the position of

the response pulses chosen to represent the strength of the onset and offset responses, for each of the five possible syllable durations. The horizontal lines indicate the strength of the onset glottal pulse for each signal.

The percentages above each response pattern are the percent of stop responses to the shortest syllables (left value) and longest syllables (right value) with the indicated rise time. This figure clearly shows the relationship between the relative heights of the onset and offset pulses, and the perceptual response of stop or glide. When the onset is fast enough that it is significantly larger than any of the offset pulses (top signal), syllables of any duration are perceived as beginning with a stop. When the onset is slow enough that it is significantly smaller than any of the offset pulses (bottom signal), syllables of any duration are perceived as beginning with a glide. For intermediate onset times, the relative sizes of the onset and offset pulses will vary with the duration of the syllable, and the perceptual response varies in the same way.

3.2.2.4 Statistical Analysis

In this section we will attempt to answer three questions.

* How well do the acoustic properties of rise time and syllable duration explain the perceptual data?

* How well do the auditory properties of onset and offset rate explain the perceptual data?

* How linear is the relationship between offset firing rate and the onset rate representing the phonetic boundary between stop and glide perception?

The first two questions can be answered using a multiple regression analysis. In both the acoustic and auditory analyses, the dependent variable is the percent of stop judgements by the listeners, converted to a normalized z score. In the acoustic phonetic analysis, the two independent variables are rise time and syllable duration. In the auditory analysis, the two independent variables are onset firing rate and offset firing rate.

The results of the multiple regression analysis are shown in Table 3.3. It can be seen that both the acoustic analysis and the auditory analysis provide a good explanation for the variability in the data: the acoustic analysis explains 88 percent of the variability, and the auditory analysis explains 94 percent of variability. The auditory measures of onset and offset firing rate are more strongly linearly correlated with the perceptual responses. The difference is statistically significant at the p=.Ø1 level, as measured by a Student's t ratio of the size of the residuals in the z scores after the respective linear regressions.

In both cases, the measure of the onset characteristics of the stimuli accounts for most of the variability in the data. From the discussion above, it should be clear that the measures associated with syllable duration only affect the results when the onset measures are equivocal. Since much of the current data contains onset characteristics clearly indicative of a stop or a glide, the onset measures have the larger partial correlation coefficient.

In order to analyze the role of syllable duration and its auditory measure, we can return to Figure 3.4, which graphs the dependence on syllable duration of the phonetic boundary between stop-like onset times and glide-like onset times. It is possible to construct a similar graph showing the dependence on offset firing rate of the phonetic boundary between stop-like onset firing rates and glide-like onset firing rates. Figure 3.12 shows both of these graphs. For the auditory analysis, the stimuli were grouped into clusters with similar offset firing rates. A cluster was formed around each stimulus whose offset firing rate differed by more than two percent from a previous stimulus around which a cluster had been formed. Each stimulus with a z score between +2.Ø and -2.Ø, and whose offset rate was within five percent of the rate of the first cluster member, was added to the cluster. Of the clusters that were formed using this method, eight of them had at least three members, and had both positive and

negative z scores. An interpolated boundary onset firing rate was calculated for each of these eight clusters using a linear least-squares fit to the z scores.

It can be seen that the relationship between the auditory measures of onset and offset firing rate is quite linear. Why is the auditory relationship a straight line while the acoustic relationship is a saturating curve? Presumably because the kind of "syllable duration" that is used perceptually grows in much the same way that firing rate falls off: quickly at first, and then slower and slower. So the relationship between perceptual duration (in this case) and a property of auditory response (offset firing rate) is more direct than the acoustic correlate of syllable duration. The peripheral auditory system appears to provide the central auditory system with a measurable response property that corresponds linearly to the way in which syllable duration is perceived.

3.3 VOICED/UNVOICED EXPERIMENT

In this experiment, natural utterances of the word "rabid" were edited to adjust the duration of the first syllable and the intervocalic stop gap. The resulting set of tokens were perceived by listeners to be either the word

"rabid" or the word "rapid", depending on the edited durations. The tokens were created and analyzed by Price and Simon (1984). The original experiment and its results will be described first, and then the response of the PAM to the stimuli. A simple measure of the degree of forward masking evident in the auditory response will be proposed, and the correlation between that measure and the perception of voicing will be explored.

3.3.1 The original experiment

3.3.1.1 Background

A number of studies have demonstrated that many acoustic properties can cue the distinction between voiced and unvoiced stop consonants. (Edwards, 1981; Klatt, 1975; Lisker, 1957; Lisker, 1978; Stevens and Klatt, 1974; Zue, 1976). For intervocalic stops, one of these properties is the length of the closure duration. Although Edwards found that closure duration plays only a weak role in voicing identification in natural speech, and does not exhibit a bimodal distribution of values, he found that voice onset time and consonant duration play very important roles. In edited speech tokens, where the closure duration has been substantially decreased or increased, the variation in closure duration can have an overwhelming effect on the voiced-unvoiced distinction, perhaps because in these tokens it elicits the same response

that changing VOT or consonant duration does in natural speech. Another acoustic property that can affect voicing identification in English is the duration of the preceding vowel. Edwards found that in normal speech this is the weakest of the 13 acoustic cues to voicing that he investigated, but as we shall see it can significantly affect voicing perception in properly controlled circumstances.

The Price and Simon experiment examined the effect of listener age and stimulus presentation level on the way in which closure and vowel durations signal the linguistic category of voicing in intervocalic stop consonants. Older and younger subjects with approximately equal audiometric hearing levels were used in order to avoid confounding the effects of age and acuity. These two groups of subjects listened to edited "rabid" syllables played at 60 and 80 dB HL. Thus four effects were investigated: closure duration, vowel duration, subject age, and presentation level.

3.3.1.2 Experimental Procedure

Four tokens of the word "rabid", spoken by an adult male native speaker of English, were used as a source of stimuli. Spectrograms of these tokens are shown in Figure 3.13. The tokens were low-pass filtered at 3.5 KHz. The resulting signals had amplitudes at and above 4 KHz that were at least 70 dB below the spectral peak in the middle of the stressed

initial vowel. The authors state that

> The words used ... had stop release bursts with predominant energy at relatively low frequencies (i.e., below 4 KHz). Thus it is likely that little, if any, information was destroyed in the filtering process.

After filtering, a waveform editor was used to replace the original 90 to 95 msec closure durations with one of four intervals of silence: 35, 65, 95, or 125 msecs. In a similar manner, the original preclosure voicing, which had a duration of 200 to 230 msecs, was adjusted by removing or duplicating glottal pulses in the steady-state portion of the vowel to create one of four "vowel" durations: 160, 180, 200, or 220 msecs. Figure 3.14 shows the waveforms of four stimuli representing the endpoints of both the closure durations and vowel durations.

The resulting set of 64 stimili (four original tokens times four closure durations times four vowel durations) were recorded on tape. Eleven randomized sequences of the complete set were recorded. The first sequence was used as a practice set. In the other sequences, each stimulus was separated from the following stimulus by three seconds of silence. Each subject listened to the tape twice, with a delay of several days between the sessions. During the first session the tape was played at 60 dB HL. During the second session the tape was played at 80 dB HL. During the second session only five of the sequences were played followed the practice sequence.

Ten of the experimental subjects were young (between 16 and 26), and ten of the subjects were older (over 55). An attempt was made to find older subjects with pure-tone audiograms as close as possible to the younger subjects. This was done because the authors wanted to study age effects other than hearing level. Figure 3.15 shows the pure-tone audiograms of both groups of subjects. The mean pure-tone thresholds of the older group are less than 10 dB below the mean for the younger group at all frequencies measured below 4 KHz.

The stimuli were presented to subjects monaurally, and the subjects were asked to respond with "b", "p", or some other medial consonant. (At the shortest closure durations, some subjects reported hearing "t" or "d", presumably reflecting the perception of an intervocalic flap.)

3.3.1.3 Analysis

Figure 3.16 shows the average percent of "p" responses from subjects, as a function of age group, level of presentation, vowel duration, and closure duration. By far the strongest effect seen in this figure is the effect of closure duration on the percent of "p" responses. Closure durations of 35 msecs elicted predominantly "b" responses, with some "t" or "d" responses. No "p" responses occurred. Closure durations of 125 msecs elicted predominantly "p" responses. For

tokens with closure durations intermediate to these extrema, vowel duration, level of presentation, and subject age all have an effect:

* The longer the preceding vowel, the longer the boundary duration;

* The older the subject, the longer the boundary duration;

* The quieter the presentation, the longer the boundary duration.

These effects are summarized in Figure 3.17, where the interpolated closure boundary durations are shown as a function of the other independent variables. The interpolation was done using normalized z scores in a manner similar to that described in the previous experiment.

The effect of closure duration on voicing perception is typical of acoustic properties which are perceptually interpreted in a categorical manner. The identification of the stop consonant as voiced or unvoiced changes abruptly across some boundary value of closure duration. The auditory system acts as if some perceptual anchor exists at the boundary location, although no such anchor is visible in the acoustic waveform.

The effect of vowel duration is another example of a trading relation between multiple acoustic cues to a single

phonetic distinction. One possible explanation for the effect of vowel duration follows along much the same line as the "context dependent" explanation for the effect of syllable duration in the stop/glide experiment: short closures indicate a voiced stop, and long closures an unvoiced stop, but what constitutes "short" and "long" depends on speaking rate. An intermediate closure duration would be considered longer if it appeared in the context of a rapidly articulated sentence, while the same closure, occurring in the midst of a slowly articulated sentence, would constitute a shorter closure. Articulation rate, in turn, can be estimated from vowel duration. Therefore a longer vowel duration signals a slower articulation rate, and results in more "short closure" judgements and hence more voiced responses.

Price and Simon present a tentative alternative explanation, based on assumptions regarding the peripheral auditory response to their stimuli. They assume that "the listener is sensitive to the amplitude of the release burst in deciding whether a /p/ or /b/ was heard." They then argue that changing vowel duration, level of presentation, and subject age will affect the level of auditory response to the release burst. This explanation is problematic for several reasons. First, neither Edwards nor Zue found any systematic differences in the burst amplitude of voiced and unvoiced stops. Second, our model shows relatively small differences in the size of the

release burst as a function of the parameters mentioned above. Hence we shall not pursue this proposal in detail.

Continuing their analysis, Price and Simon cite a number of studies which suggest that there are age-related effects in the processing of temporal information in speech (Beasley and Maki, 1976), and that temporal resolution deteriorates with age (Ludlow, Cudahy and Bassich, 1982). They speculate that the age related differences in boundary duration that they found "are correlated with the neural response of the auditory system to the release burst of the consonant, and that this correlation may be due to a lengthening with age of the recovery time of neural fibers."

3.3.2 The PAM Response

In this section we shall examine the response of the PAM to the voiced/unvoiced stimuli. We will first attempt to find some measurable property of the PAM response pattern which corresponds to the abstract acoustic phonetic property of closure boundary duration. Then we shall investigate the way in which vowel duration and neural recovery time--as a hypothesized correlate of age--affect this property. We shall have little to say about the effect of presentation level on the closure boundary, because the dynamic range of the PAM is too limited to allow us to test stimuli that differ in intensity by 20 or even 10 dB.

3.3.2.1 PAM Processing Procedure and Parameters

The stimuli were digitized at 12,987 samples per second from audio tape. The audio signal was low-pass filtered at 6,400 Hz before digitizing. The digital signals were preemphasized as described in Chapter Two, and then processed by the spectral analysis stage. The nonlinearity used in the transduction stage was the DC transfer function of a raised hyperbolic tangent function, with a bias parameter of -1.1. To model the response of the younger subjects, a three time constant adaptation stage was used, with time constants of 15, 40, and 320 msecs, contributing equally to the transient response. The adaptation ratio was 3.5. (These are the same PAM operating parameters used for the stop/glide experiment.) To model the response of the older group of subjects, only the middle adaptation time constant was changed. It was increased from 40 to 50 msecs, in line with Price and Simon's suggestion.

3.3.2.2 PAM Response Patterns

Figure 3.18 shows the response of the peripheral auditory model to a stimulus created from Token 1. The stimulus in this figure corresponds to the shortest of the four vowel durations and the shortest of the four closure durations. Figures 3.19 through 3.21 show the other three endpoint stimuli corresponding to the four possible combinations of

shortest and longest vowel and closure durations. General features of interest that are visible in these response patterns include:

* the steady level of spontaneous firing before the beginning of the first syllable;

* the rising first and second formant during the transition into the steady state portion of the first syllable;

* the clear definition of the glottal pulse in channels that are not saturated;

* the saturation of the channels which include the first and second formants;

* the significant amount of adaptation of response in the formant channels during the course of the initial vowel;

* the almost total lack of spontaneous firing across all channels at the beginning of the intervocalic stop closure;

* the recovery of spontaneous firing in some channels during the course of the longer closures;

* the strong alveolar offset transitions in the first and second formant at the end of the word;

* the recovery of spontaneous firing following the end of the word, with heavily stimulated channels around the formants recovering later than the less stimulated channels at high frequencies and in the valleys between formants.

As mentioned above, Price and Simon suggested that changing the duration of the stop closure and the vowel might change the size of the auditory response to the stop burst.

Figure 3.22 shows this response, displayed as a spectral pattern. It can be seen that shorter closure durations, and longer vowel durations, do result in a somewhat smaller burst response. However, these changes are relatively small, corresponding to changes in stimulus amplitude of less than 5 dB. Because these differences are so small, burst response size did not appear to be a robust auditory property to try to correlate with voicing perception.

If we look at the auditory response within the intervocalic stop gap, and ask ourselves what other property of the response pattern might be associated with a categorical boundary for voicing perception within the range of 50 to 100 msecs, an obvious possibility is the return of spontaneous firing following recovery from the initial syllable. It should be noted that the small vertical scale of Figures 3.18 through 3.21, and the poor resolution of the device on which they were plotted, make it appear that spontaneous firing begins considerably later than it actually does.

The recovery of firing is most easily examined by using the masking postprocessor described in Chapter 2. Figures 3.23 through 3.26 show the output of this postprocessor when applied to the PAM response patterns to the endpoint stimuli. In these figures, a high value in a channel at a given instant means that the expected firing rate in that channel is zero. We will describe such a channel as "masked". As can be seen,

masking begins to appear in heavily stimulated channels during the first syllable, for one or two msecs during the latter half of each glottal pulse. However, sustained masking across almost all channels begins only with onset of the stop closure. When the closure duration is only 35 msecs, masking continues up to the release of the stop, with spontaneous firing reappearing only in the lowest and some of the highest channels, where there is little acoustic energy in the first syllable. In contrast, when the closure duration is 125 msecs there is more opportunity for recovery, and spontaneous firing spreads across all but the most heavily stimulated channels. By comparing Figure 3.24, which shows the masking pattern in response to a stimulus with a "short" vowel duration, and Figure 3.26, which shows the masking pattern in response to a stimulus with a "long" vowel duration, it can be seen that vowel duration affects the recovery period. This is due to the long term adaptation constant, which results in more adaptation, and therefore a longer recovery time, in response to 220 msec vowels than to 160 msec vowels.

3.3.2.3 A Possible Auditory Analysis

Examination of the masking patterns in Figures 3.23 through 3.26 suggests that some composite measure of the amount of masking present in the peripheral auditory response at the end of the closure in an intervocalic stop may be

related to the perception of the closure as "short" or "long". In this section we will develop a quantitative measure of closure length based on this idea, and explore its correlation with the perceptual responses of listeners in this experiment.

A simple quantitative measure of degree of masking at a particular moment can be produced by counting the fraction of channels for which the output of the masking detector is high (equal to one). The resulting measure will have the value one when all channels have an expected firing rate of zero, and will have the value zero when all channels have an expected rate greater than zero. (Although we are not presently concerned with constructing a measure of masking which is psychologically valid, it should be noted that the central nervous system could implement a masking detector similar to the one we have proposed. To do so requires knowing that the spontaneous firing rate of an auditory neuron is greater than zero, detecting the fact that the current rate is zero, and generating a neural response that signals those facts.)

This measure was applied to the response of the PAM to the stimuli in this experiment. As a final step, the measure was smoothed using a 5 msec left half-hamming window. The resulting signal represents a smoothed composite measure of the degree of masking in the peripheral auditory nerve.

Figure 3.27 demonstrates the behavior of this signal and its relationship to the acoustic signal and the composite PAM firing rate signal used in the previous experiment. The acoustic waveform, shown on top, is a stimulus produced from Token 1 and has a 160 msec vowel duration and 220 msec closure duration. The middle signal is the composite PAM response smoothed with a 20 msec half hamming window. Interesting features in this signal include the initial spontaneous firing, the clear response to the onset burst visible in the acoustic waveform, the saturation in firing rate during the stressed vowel (no glottal pulses visible), the steady recovery of spontaneous firing during the stop closure, and the glottal pulses during the second vowel. The bottom signal is the smoothed composite masking signal generated by postprocessing the PAM response to the top signal. The masking signal has an initial value of zero, because all channels are firing at their (nonzero) spontaneous rate before the word begins. As the acoustic energy increases, heavily stimulated channels begin to exhibit masking (complete lack of firing) during the second half of glottal cycles. This appears in the bottom trace as glottal pulse-like spikes, visible for both the first and second vowels. At the beginning of the stop closure the amount of masking shows an abrupt increase, and then decreases in a step-like fashion over the course of the closure. The staircase effect is due to the fact that the PAM has only 20 channels, and only a minimum amount of temporal

smoothing has been done. With a sufficiently large number of channels the masking level would presumably decline in a ramp-like fashion.

Figure 3.28 shows these three stimulus representations for all four of the endpoint stimuli produced from Token 1: the four combinations of the shortest and longest vowel and closure durations. Two things should be noted. First, short closure durations result in high masking levels at the release of the stop. Second, a longer vowel duration causes the masking level to remain higher for a longer period of time.

All of the experimental stimuli can be divided into sets, such that all members of a set were produced from the same original token, and have the same vowel duration. There are 16 such sets: four original tokens times four vowel durations. Analysis of the offset of masking in the stop closure can be simplified by realizing that its time course is the same for all stimuli in each set. For a given set, the only thing that differs from one member of the set to another is the duration of closure, and hence the level to which masking falls before the release of the stop. Therefore if we can characterize the time course of the decline of masking for the longest closure duration in each set, we can use that characterization to analyze all four members of the set.

The decline of masking is sufficiently linear that we can characterize it as a ramp, and fit a straight line to it using a least-squares fit. Table 3.4 shows the result of fitting a straight line to each of the postprocessed PAM responses to the sixteen stimuli with 320 msec closure durations. These results are with the PAM operating using adaptation constants of 5, 40 and 320 msecs, the values used in the previous experiment, and used in this experiment to model the response of the younger subjects. The lines are fitted to masking level values that meet all of the following critia: located within the stop closure; less than 90 percent of the maximum masking level value; more than 10 percent above the minimum value; and between 0.3 and 0.7 (30 to 70 percent of the channels masked). Typically, more than 50 sample values per signal meet these criteria and are included in the least-squares calculation. The goodness-of-fit measure listed in Table 3.4 is the fraction of the variation in masking level values predicted (fitted) by the straight line. It can be seen that the decline of masking can be represented well by a straight line.

To review, we are attempting to construct a peripheral auditory measure of the "length" of a closure duration, and have decided to base this measure on the degree of recovery from adaptation. This, in turn, is being measured by the fraction of channels which have recovered sufficiently to have a nonzero spontaneous rate. We can arbitrarily pick a certain

fraction, for example Ø.5, and say that a closure is long if more than this fraction of the channels have recovered from masking before the closure comes to an end. The rightmost column in Table 3.4 shows the interpolated times at which this masking boundary is reached for each of the sixteen stimulus sets.

What is the relationship between our masking boundaries and listeners' perceptual boundaries? Figure 3.29 provides a graphical answer to this question. Each of the four panels in this figure show the relationship between the masking boundaries and the perceptual responses for one group of subjects and one level of presentation. Each data point corresponds to the results for a single vowel duration, for stimuli from a single original token.

In general, as vowel duration increases, both perceptual and masking boundaries increase. If the perceptual and masking boundaries increase linearly with respect to each other, the connecting line between data points will be straight. If the perceptual and masking boundaries increase the same amount, the straight line will be parallel to the diagonal lines in the panels. If the perceptual and masking boundaries are equal, the corresponding data point will fall directly on the diagonal.

It is clear from Figure 3.29 that a linear relationship exists between the interpolated masking boundaries and listeners' perception of the stimuli as voiced or unvoiced. For some tokens, groups of listeners, and presentation levels, the match appears to be almost perfect (c.f., Token 3, young group, 60 dB; and Tokens 1 and 3, young group, 80 dB). On the other hand, it is also clear that the masking boundary is not perfectly correlated with all the perceptual data. The degree of correlation differs between tokens. Token 4, for instance, shows a much greater change of perceptual boundary than masking boundary as vowel duration increases. Clearly, listeners are sensitive to some properties of these tokens that are not being captured by our masking measure. This is hardly suprising, given the number of acoustic features that have been found to affect voicing perception in intervocalic stops. The important point is that a simple measurable property of the peripheral auditory response shows a strong linear correlation with the phonetic feature of voicing, that the nature of that property makes it reasonable to suppose that it could serve as an anchor for categorical perception, and that the peripheral auditory system combines in a single response property two apparently unrelated acoustic properties of the stimulus, namely vowel duration and closure duration.

What of the other two controlled variables in this experiment? As we mentioned before, the dynamic range of the PAM

is too limited to be able to change the input intensity by 20 or even 10 dB. However, small changes in input intensity show that the current model does not predict the observed results: increasing the input intensity results in longer masking boundaries (because of increased adaptation), while listeners report shorter perceptual boundaries. (This is visible in Figure 3.29 as a general shift downward of the data points in the right panels compared with the left panels.)

The final controlled variable is the age of the subjects. In general, older subjects reported longer perceptual boundaries. This can be seen in Figure 3.29 as a general shift upward of the data points in the bottom panels compared with the top panels. Price and Simon suggest that this might be explained by an increase with age of the recovery time of auditory neurons. In the PAM, this translates into an increase in adaptation time constants. To test this hypothesis, all of the stimuli in this experiment were reprocessed with the PAM operating with the short term adaptation time constant increased from 40 msecs to 50 msecs. No other changes in parameter values were made. Table 3.5 shows the interpolated results for this configuration of the model, and Figure 3.30 shows the resulting relationships between perceptual and masking boundaries. Whereas in Figure 3.29 the data points for the younger subjects are roughly centered around the diagonal lines marking an ideal relationship, and the data

points for the older subjects are mostly above the diagonal, in Figure 3.3Ø the data points for the older subjects are clustered symmetrically about the diagonal, and the data points for the younger subjects tend to fall below the diagonal. This suggests that Price and Simon's hypothesis would be an adequate explanation of the differences in the responses of the younger and older subjects.

3.4 CONCLUSION

In this chapter we have presented some evidence in support of the correspondence theory of the role of the peripheral auditory system in speech perception. This theory proposes that the transformations which speech signals undergo in the peripheral auditory system help to decode the underlying phonetic features upon which the linguistic content of the speech depends. What kinds of transformations have we seen, and how might they constitute a decoding of the speech signal? The evidence of this chapter suggests five types of peripheral auditory transformations that may help decode speech.

From abstract to concrete: The peripheral auditory system can transform an abstract acoustic property, such as syllable duration, onset time, vowel duration, or closure duration,

into a concrete, measurable property of the auditory neural response. Whether or not the central nervous system actually responds to that property is an unanswered question, but Figures 3.12, 3.29, and 3.30 demonstrate that the CNS responds to something highly correlated with the properties we have identified.

From complex to singular: The peripheral auditory system can transform a group of acoustic properties that are known to participate in a phonetic trading relationship into a single auditory property. The clearest example of this is the way in which closure duration and preceding vowel duration combine to influence the level of masking in the auditory response.

From nonlinear to linear: The peripheral auditory system can transform a nonlinear acoustic phonetic relationship into a linear auditory phonetic relationship. This is clearly shown by Figure 3.12, where the nonlinear relationship between syllable duration and onset boundary duration is transformed into a linear relationship between offset firing rate and onset boundary firing rate.

From continuous to segmental: The peripheral auditory system can transform an undifferentiated acoustic dimension, such as time or amplitude, into a more sharply segmental auditory dimension that can form the basis for quantal or categorical perception. This type of transformation is in some sense

the reverse of the previous type, and could be described as from linear to nonlinear. Auditory properties such as spontaneous rate, saturation, and masking serve to warp uniform steps in an acoustic scale into nonuniform steps in an auditory scale. Related to this is the development of concrete response patterns which could serve as an auditory anchor for the categorical perception of a phonetic distinction. Both of the experiments discussed in this chapter provide examples of this pre-categorical transformation. In the first experiment, the offset firing rate appears to provide a normalizing measure which allows a continuous range of onset rates to be divided into the categories of "fast" and "slow". In the second experiment, the recovery of spontaneous firing may constitute an auditory event that allows a continuous range of closure durations to be divided into the categories of "short" and "long".

From context free to context dependent: The peripheral auditory system can transform an acoustic signal whose properties are uncorrelated with its properties 50 or 100 msecs earlier into an auditory response whose properties reflect the time course of the acoustic signal over the previous 100 or 200 msecs. In other words, adaptation of firing rate provides the peripheral auditory system with a simple kind of short term memory. This dependence of the peripheral auditory response on previous context is analogous to some of the

aspects of contextual perception of speech, and we saw clear examples of the correlation between perceptual memory and peripheral auditory memory in the effects of vowel and syllable duration on the phonetic boundaries of continuant-noncontinuant and voiced-unvoiced speech.

FIGURE CAPTIONS FOR CHAPTER 3

FIGURE 3.1

Schematic diagram showing synthesis parameters of stimuli in STOP/GLIDE experiment. At the bottom is shown the onset and offset of amplitude of voicing. The upper section shows the change in F1 and F2. The vertical dashed lines indicate the five syllable durations. (From Landahl and Maxwell, 1983.)

FIGURE 3.2

Waveforms of stimuli with rapid and slow onset times. On the left are shown stimuli with 15 msec onsets, on the right stimuli with 60 msec onsets. The set of five syllable durations are shown for each of the onset times.

TABLE 3.1

Percent STOP judgements for each stimulus. The average percentage of STOP responses from all of the subjects are listed for each token. At the bottom of the table the phonetic rise time boundary between STOP and GLIDE is listed. These boundaries were calculated by fitting a linear regression line to the percentages converted to z-scores.

FIGURE 3.3

Percent stop judgements as a function of onset time. A separate curve is shown for each of the five syllable durations. The longer the syllable the greater the interpolated 50% rise time boundary. (From Landahl and Maxwell, 1983.)

FIGURE 3.4

Boundary duration as a function of syllable duration. The vertical axis is the interpolated boundary transition time for which subjects would respond with an equal number of "b" and "w" judgements.

FIGURE 3.5

PAM response to stimulus with 15 msec onset and 299 msec duration. The nominal center frequencies of each channel in Hertz and Bark are listed along the sides. The vertical dimension is expected firing rate. The vertical distance between the zero lines of adjacent channels corresponds to an expected firing rate 3.Ø times the maximum steady state firing rate. The stimulus waveform is shown at the bottom, delayed by 9.9 msecs, the group delay of the PAM filters, so that glottal pulses in the waveform and in the PAM output channels are aligned vertically.

FIGURE 3.6

PAM response to stimulus with 15 msec onset and 87 msec duration. See the legend for Figure 3.5 for further information.

FIGURE 3.7

PAM response to stimulus with 6Ø msec onset and 299 msec duration. See the legend for Figure 3.5 for further information.

FIGURE 3.8

PAM response to stimulus with 6Ø msec onset and 87 msec duration. See the legend for Figure 3.5 for further information.

FIGURE 3.9

PAM response to stimulus with 3Ø msec onset and 299 msec duration. See the legend for Figure 3.5 for further information.

FIGURE 3.1Ø

PAM response to stimulus with 3Ø msec onset and 87 msec duration. See the legend for Figure 3.5 for further information.

TABLE 3.2

Location of onset and offset pulses in smoothed composite response signal. The size of the onset and offset response was taken to be the largest value obtained by the smoothed composite response signal within the temporal limits specified by this table.

FIGURE 3.11

Smoothed composite PAM responses. The acoustic waveforms of five stop/glide stimuli are shown on the left. The smoothed composite PAM responses to those waveforms are shown on the right. The stimuli shown represent the two endpoints of rapid and slow onset times at the top and bottom, and three intermediate times in the middle. The longest duration syllables are shown in each case. The leftmost vertical line indicates the location of the onset pulse, while the five vertical lines to its right indicate the location of the offset pulses for each of the five syllable durations. The number above and to the left of each response pattern is the percentage of stop responses to the shortest syllable with the onset time shown. The number above and to the right of each response pattern is the percentage of stop responses to the longest syllable with the onset time shown.

TABLE 3.3

Summary of multiple regression analysis of stop/glide data.

FIGURE 3.12

Phonetic boundary of the onset measure as a function of the offset measure. Panel A is a redrawing of Figure 3.4, using the acoustic phonetic measures of transition time and syllable duration. Panel B is the corresponding graph for the auditory phonetic measures of onset firing rate and offset firing rate.

FIGURE 3.13

Spectrograms of the four original "rabid" tokens. These original utterances are referred to in the text as Token 1, Token 2, Token 3, and Token 4 (figure courtesy of Price and Simon).

FIGURE 3.14

Waveforms of stimuli with shortest and longest closure and vowel durations. These four waveforms represent the endpoints of both the closure and vowel duration sequences. All of the waveforms shown are edited versions of original Token 1. The closure durations shown are 35 and 125 msecs, and the vowel durations are 160 and 220 msecs.

FIGURE 3.15

Audiograms of the two groups of subjects. The means and standard deviations of the pure-tone hearing thresholds are shown for the younger subjects ("Y"s) and older subjects ("O"s). (From Price and Simon, 1984).

FIGURE 3.16

Percent of "p" responses. The results for each age group and level of presentation are shown in separate panels. The label "YN" indicates the younger group of subjects, while "ON" indicates the older group. The four curves in each panel indicate the results for each of the four vowel duration (1 = 160 msecs, 2 = 180 msecs, 3 = 200 msecs, and 4 = 220 msecs). The dotted horizontal line indicates the 50 percent phonetic boundary for voiced-voiceless responses. The vertical axis in each panel is percent of "p" responses. (From Price and Simon, 1984).

FIGURE 3.17

Interpolated voiced/unvoiced phonetic boundaries. The lines marked with "Y" are for the younger group of subjects; "O" indicates results for the older group. The dotted lines

indicate the results for the 80 dB HL presentation level, while the solid lines are for the 60 dB HL. The independent variable in this figure is vowel duration. The arrows below the horizontal axis indicate that the shorter the boundary duration, the more "p" responses a subject would make, whereas a longer boundary duration would result in more "b" responses. (From Price and Simon, 1984).

FIGURE 3.18

PAM response to stimulus with 160 msec vowel and 35 msec closure. This stimulus was created by editing the original Token 1. See the legend for Figure 3.5 for further information.

FIGURE 3.19

PAM response to stimulus with 160 msec vowel and 125 msec closure. This stimulus was created by editing the original Token 1. See the legend for Figure 3.5 for further information.

FIGURE 3.20

PAM response to stimulus with 220 msec vowel and 35 msec closure. This stimulus was created by editing the original Token 1. See the legend for Figure 3.5 for further information.

FIGURE 3.21

PAM response to stimulus with 220 msec vowel and 125 msec closure. This stimulus was created by editing the original Token 1. See the legend for Figure 3.5 for further information.

FIGURE 3.22

PAM response spectra of stop bursts. This figure shows the response of the PAM to the release of the intervocalic stop. Panel A shows the response for stimuli produced from Token 1, with 160 msec vowel durations, 35, 65, 95, and 125 msec closure durations. Panel B shows the response for stimuli

produced from Token 1, with 125 msec closure durations, and 160, 180, 200, and 220 msec vowel durations.

FIGURE 3.23

Output of masking post-processor for stimulus with 160 msec vowel and 35 msec closure. This stimulus was created by editing the original Token 1. The nominal center frequencies of each channel in Hertz and Bark are listed along the sides. In each channel, a low value indicates the PAM output is greater than zero, while a high value indicates the PAM output is zero. The stimulus waveform is shown at the bottom, delayed by 9.9 msecs, the group delay of the PAM filters, so that a postprocessor output value and the most recent stimulus sample that produced it are aligned vertically.

FIGURE 3.24

Output of masking post-processor for stimulus with 160 msec vowel and 125 msec closure. This stimulus was created by editing the original Token 1. See the legend for Figure 3.23 for further information.

FIGURE 3.25

Output of masking post-processor for stimulus with 220 msec vowel and 35 msec closure. This stimulus was created by editing the original Token 1. See the legend for Figure 3.23 for further information.

FIGURE 3.26

Output of masking post-processor for stimulus with 220 msec vowel and 125 msec closure. This stimulus was created by editing the original Token 1. See the legend for Figure 3.23 for further information.

FIGURE 3.27

Three representations of a stimulus perceived as "rapid". The top signal is the acoustic waveform of a version of Token 1 with a 160 msec vowel duration and a 220 msec closure duration. The middle signal is the smoothed composite PAM response pattern, produced using the postprocessor constructed for the stop/glide experiment. The bottom signal is the composite masking detector output, smoothed using a 5 msec half hamming window, as described in the text.

FIGURE 3.28

Three representations of endpoint stimuli. The signals shown contain the four possible combinations of shortest and longest vowel and closure duration. They were all constructed from Token 1. The signals on the left are the acoustic waveforms, the signals in the middle are the smoothed composite PAM response patterns, produced using the postprocessor constructed for the stop/glide experiment, and the signals on the right are the composite masking detector outputs, smoothed using a 5 msec half hamming window, as described in the text.

TABLE 3.4

Linear fit of decline of masking during stop closure: "Young" analysis. These results are for the PAM operating with a short term adaptation time constant of 40 msec, the value used to model young subjects. Approximately 50 sample values are included in the least squares fit for each line. The initial masking level, slope, and closure duration at which the masking level reaches 50 percent are interpolated values.

FIGURE 3.29

Relationship between perceptual and masking boundaries: "Young" analysis. These results are for the PAM operating with a short term adaptation time constant of 40 msec, the value used to model young subjects. The horizontal axes are interpolated masking boundary durations, and the vertical axes are interpolated perceptual boundary durations. Data points

would fall on the diagonal line if there were a perfect match between the perceptual and masking boundaries.

TABLE 3.5

Linear fit of decline of masking during stop closure: "Old" analysis. These results are for the PAM operating with a short term adaptation time constant of 50 msec, the value used to model old subjects. For more information, refer to the legend for Table 3.4.

FIGURE 3.30

Relationship between perceptual and masking boundaries: "Old" analysis. These results are for the PAM operating with a short term adaptation time constant of 50 msec, the value used to model old subjects. For more information, refer to the legend for Figure 3.29.

FIGURE 3.1

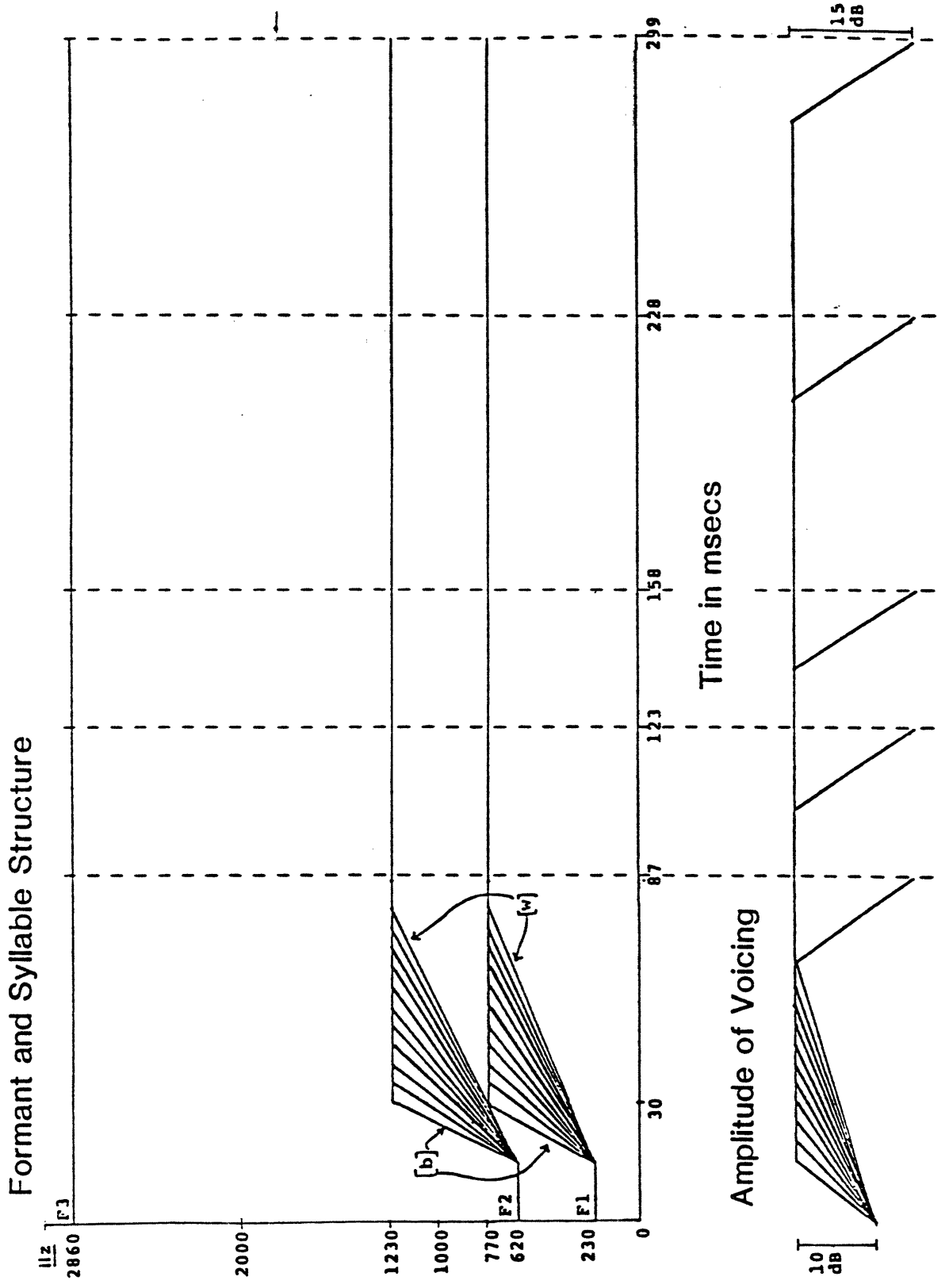
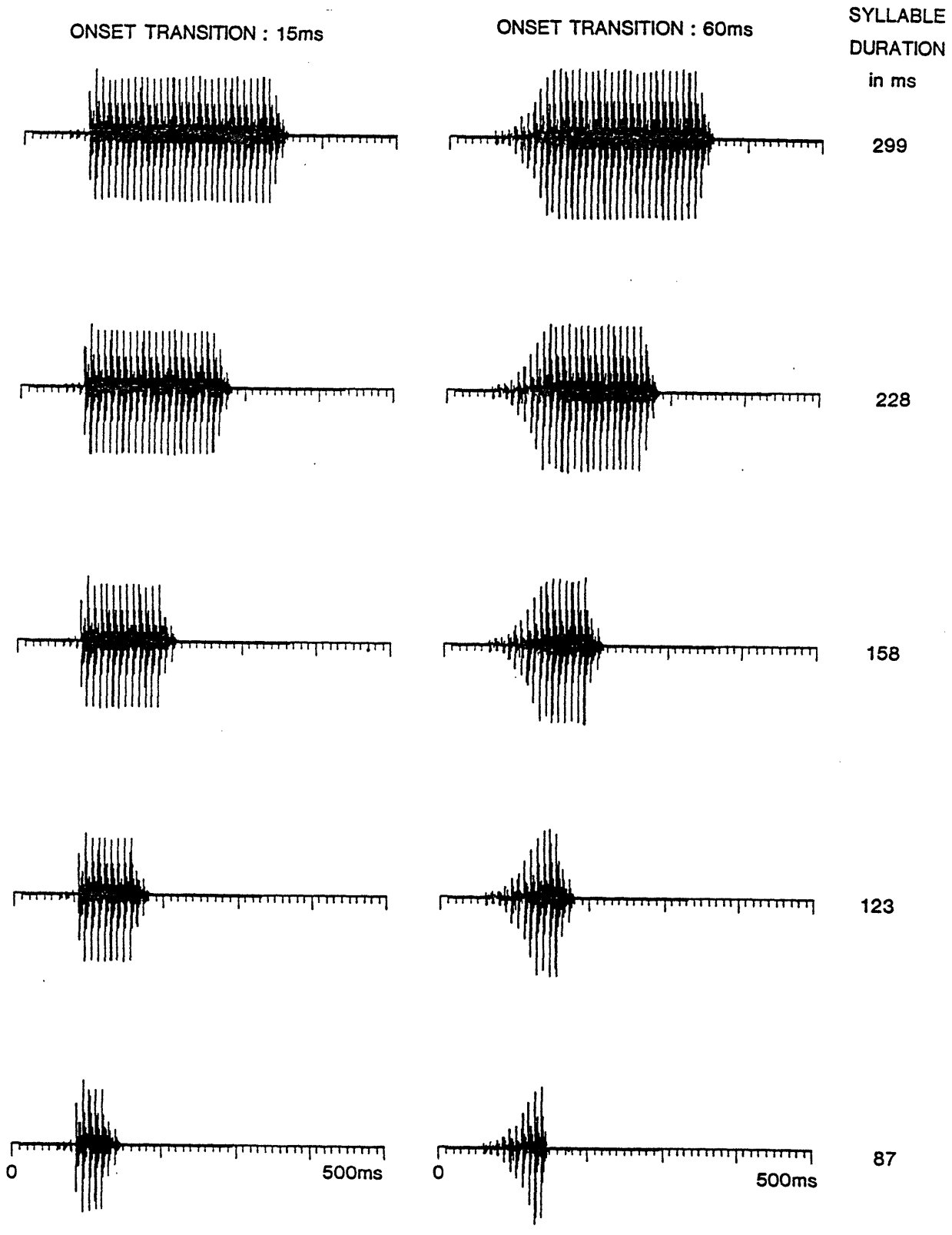

Formant and Syllable Structure
Hz
2860
2000
1230
1000
770
620
230
0
F3
F2
F1
[b]
[w]
0
30
87
123
158
228
299
Time in msecs
Amplitude of Voicing
15 dB
10 dB

FIGURE 3.2


ONSET TRANSITION : 15ms
ONSET TRANSITION : 60ms
SYLLABLE DURATION in ms
299
228
158
123
87
0
500ms
0
500ms

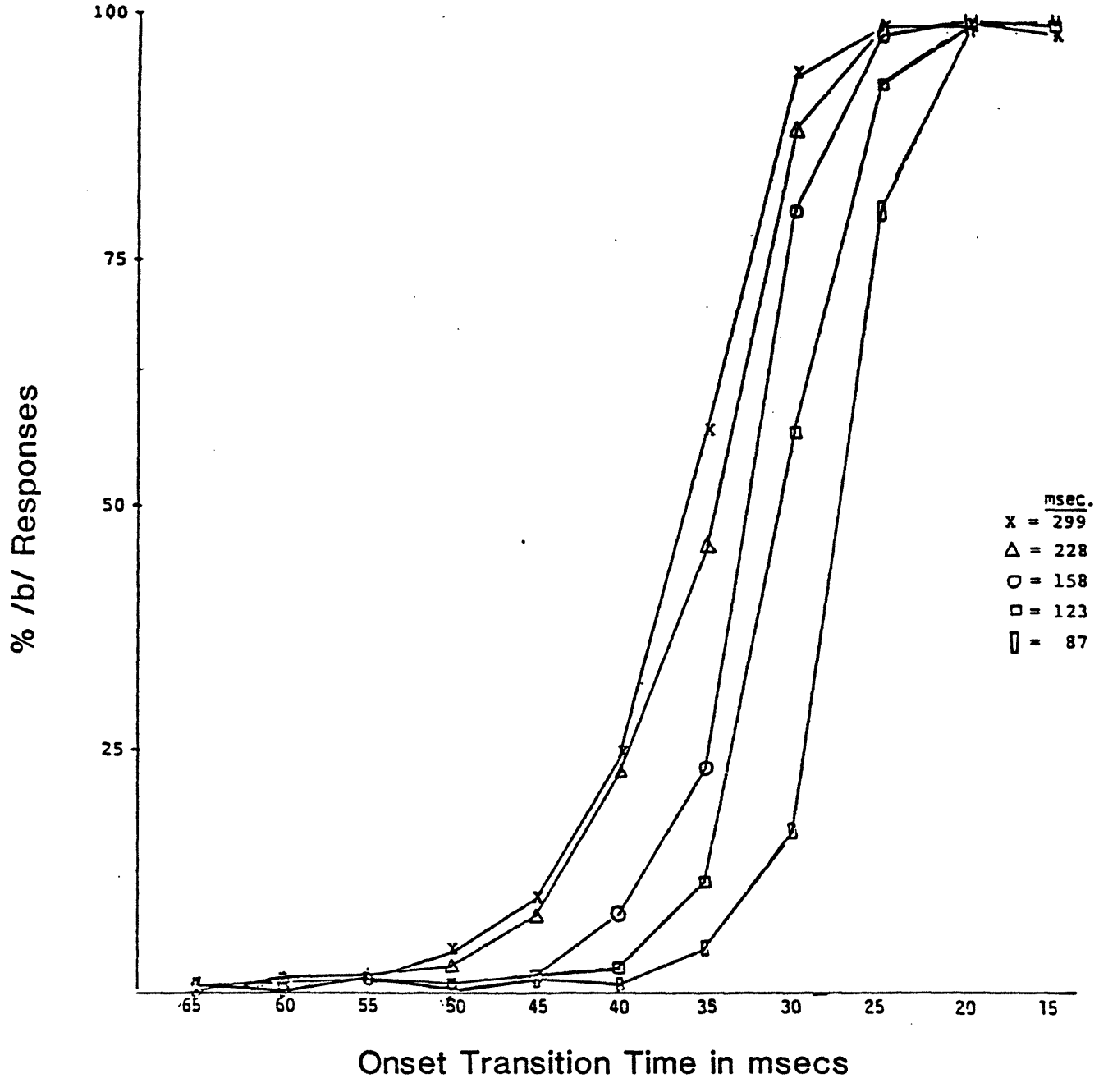
PERCENT STOP JUDGEMENTS

ONSET TIME in ms.	SYLLABLE DURATION in ms. 299	228	158	123	87
15	98	100	100	100	100
20	100	99	100	100	99
25	99	99	98	93	80
30	94	89	80	58	17
35	58	46	23	12	5
40	25	23	8	3	1
45	10	8	2	2	1
50	4	3	1	1	0
55	1	2	1	1	1
60	1	1	0	1	0
65	0	0	0	0	1

STOP/GLIDE BOUNDARY RISE TIMES in ms.

37.6 36.1 33.0 31.0 27.5

Percent /b/ Responses by Syllable Duration



PHONETIC BOUNDARY--TRANSITION TIME
AS A FUNCTION OF SYLLABLE DURATION

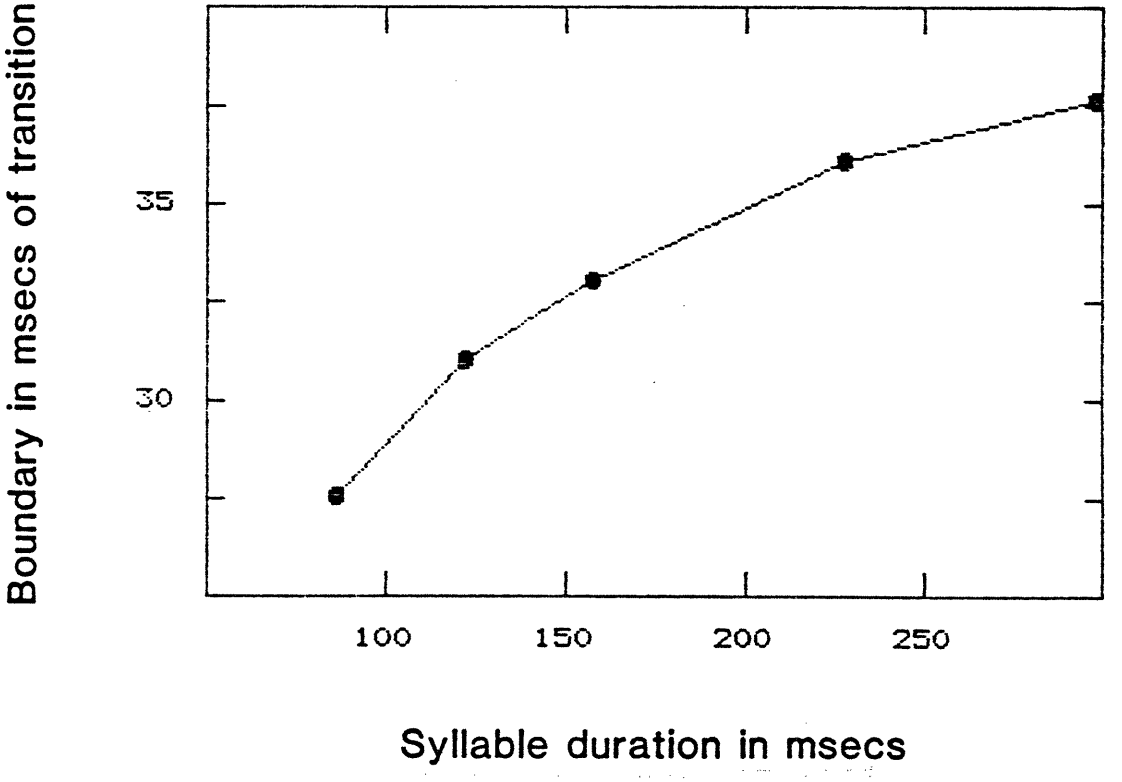


FIGURE 3.5

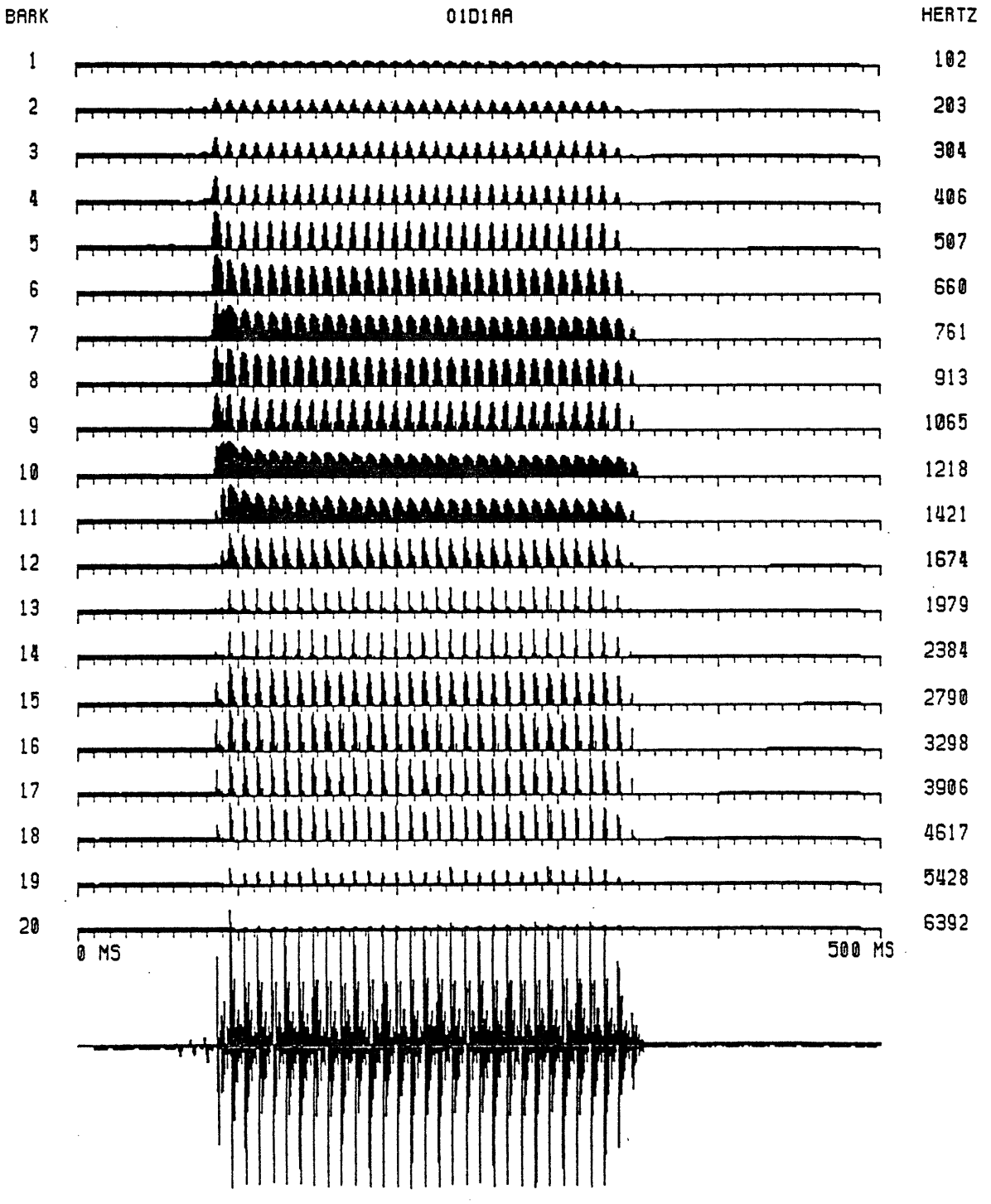

BARK
01D1AA
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.6

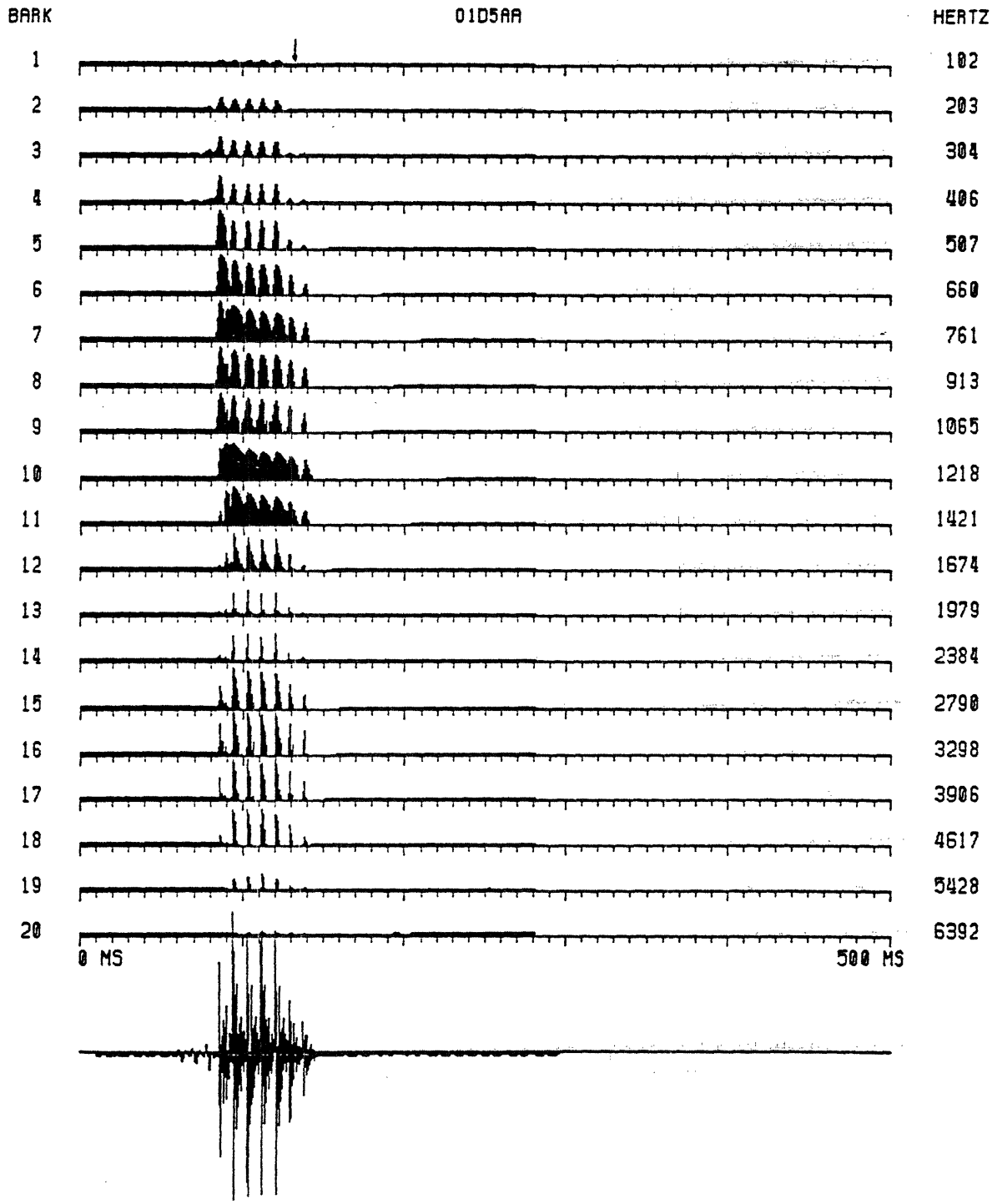

BARK
01D5AA
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.7

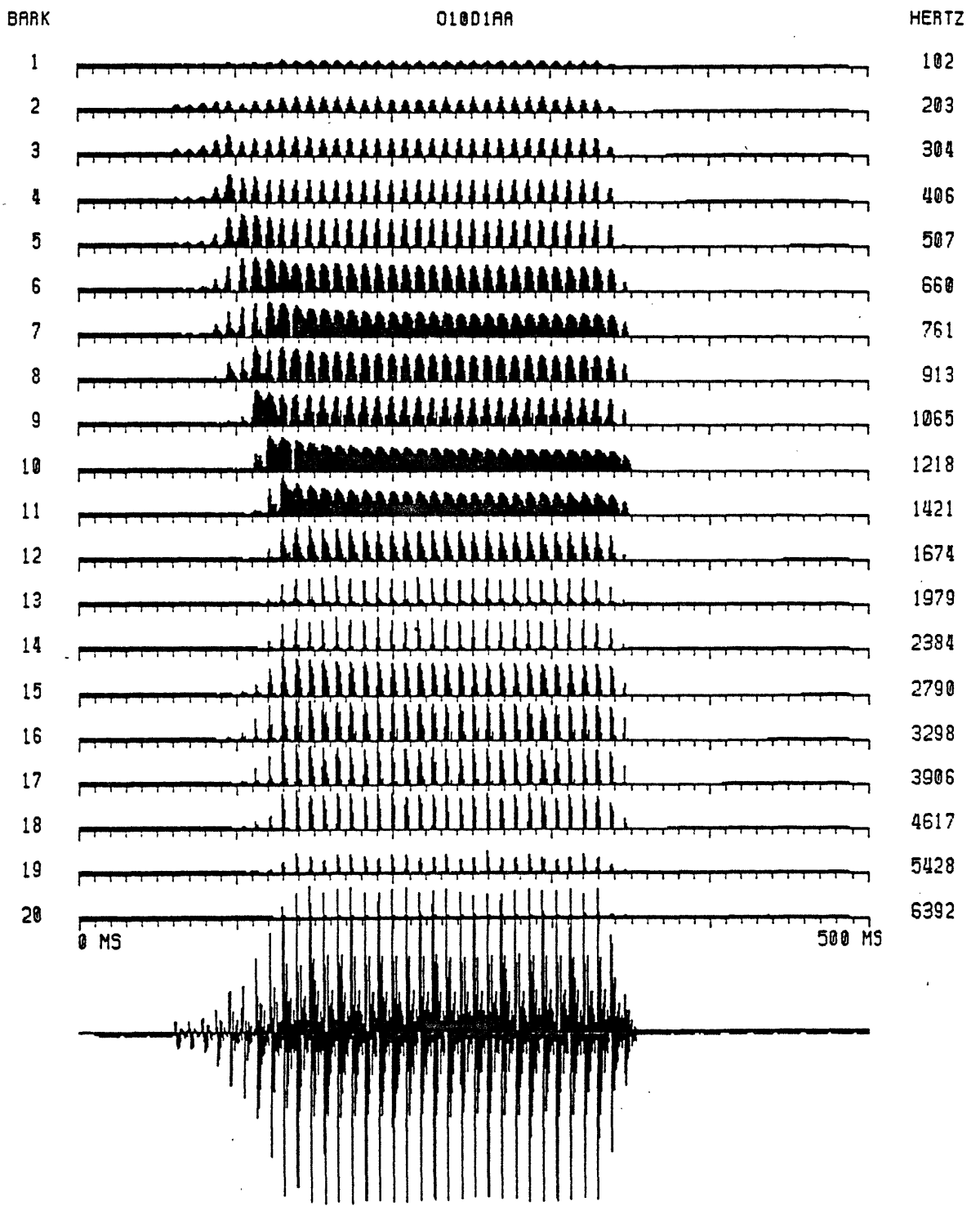

BARK
010D1AA
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.8

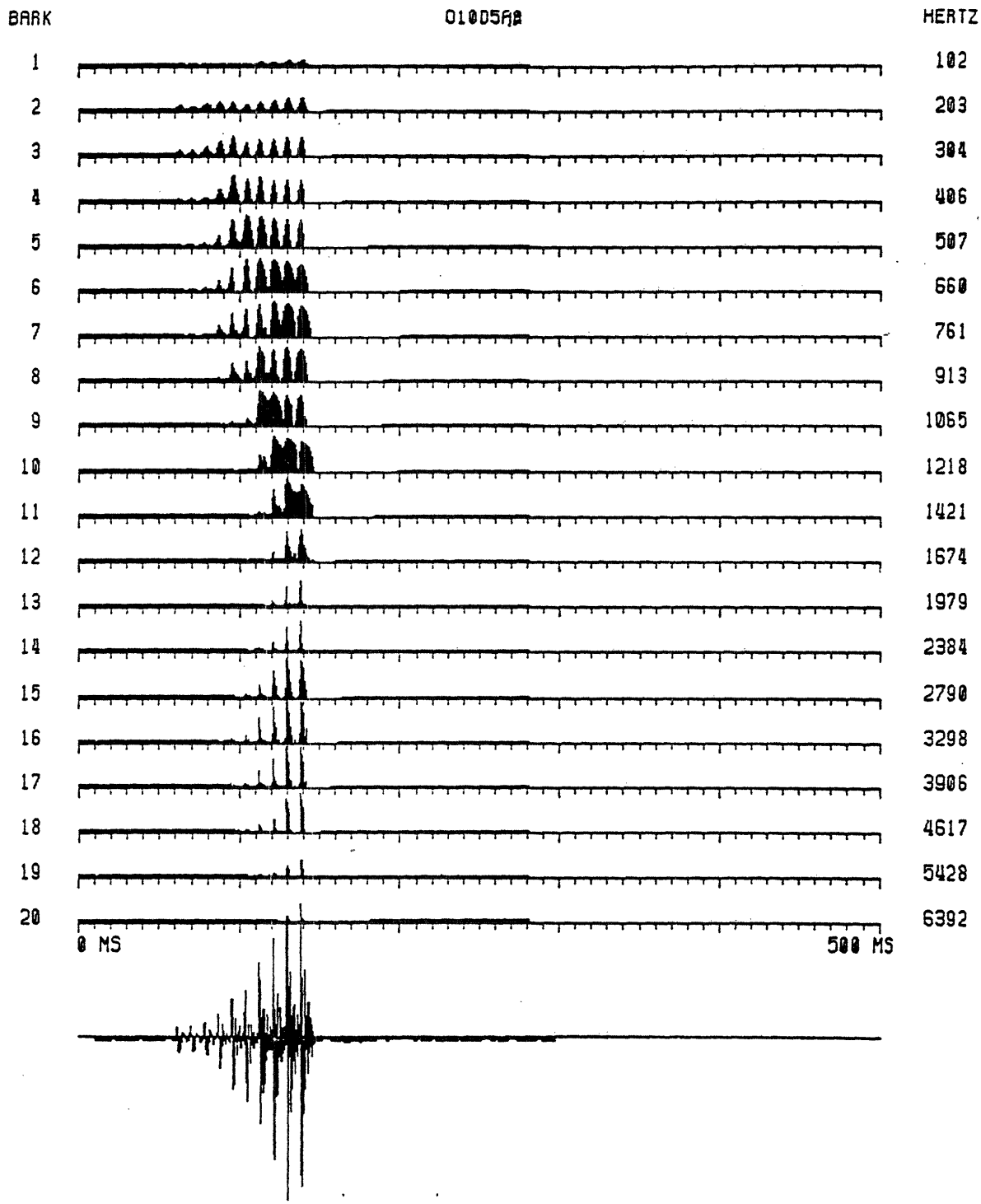

BARK
01005AB
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.9

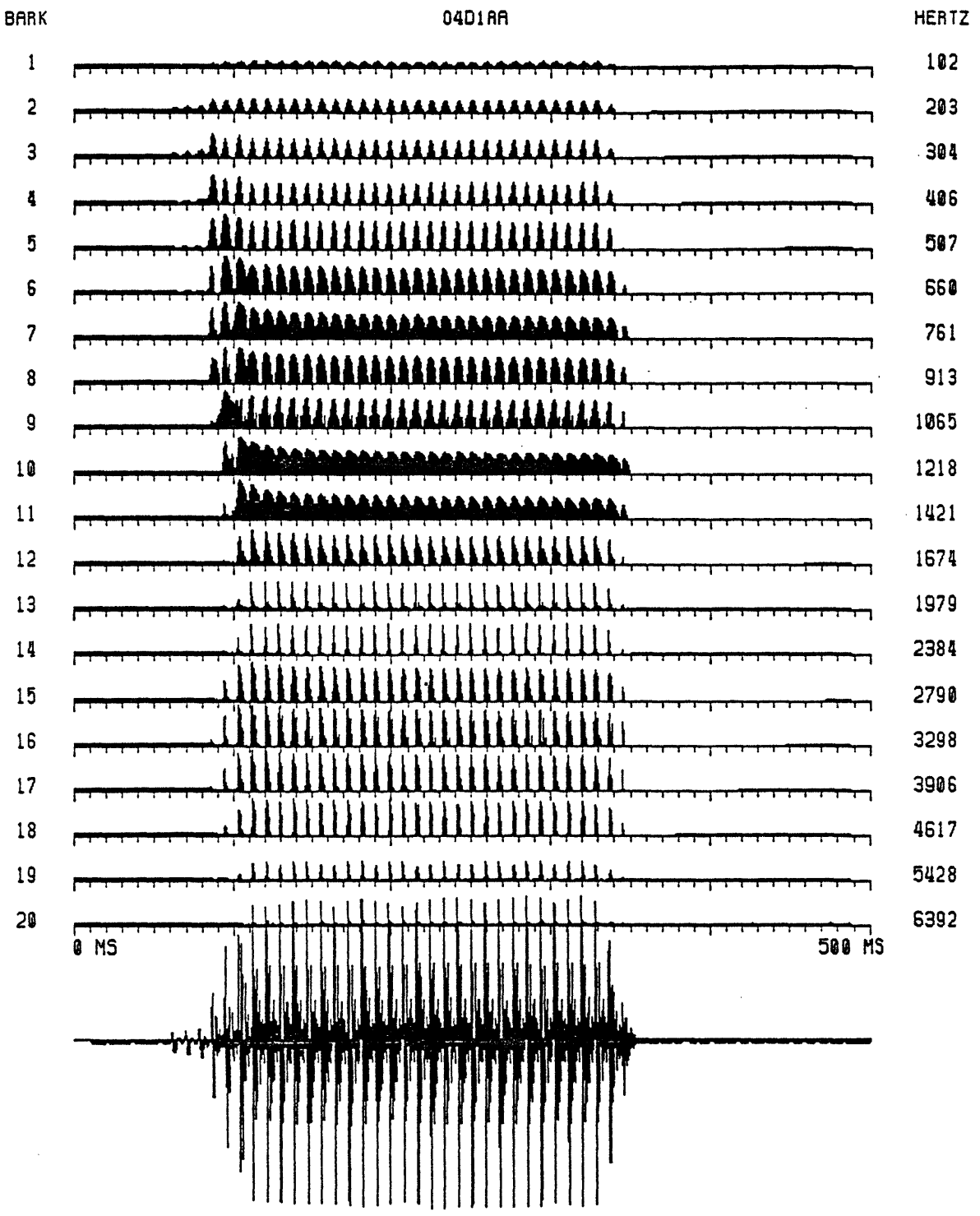
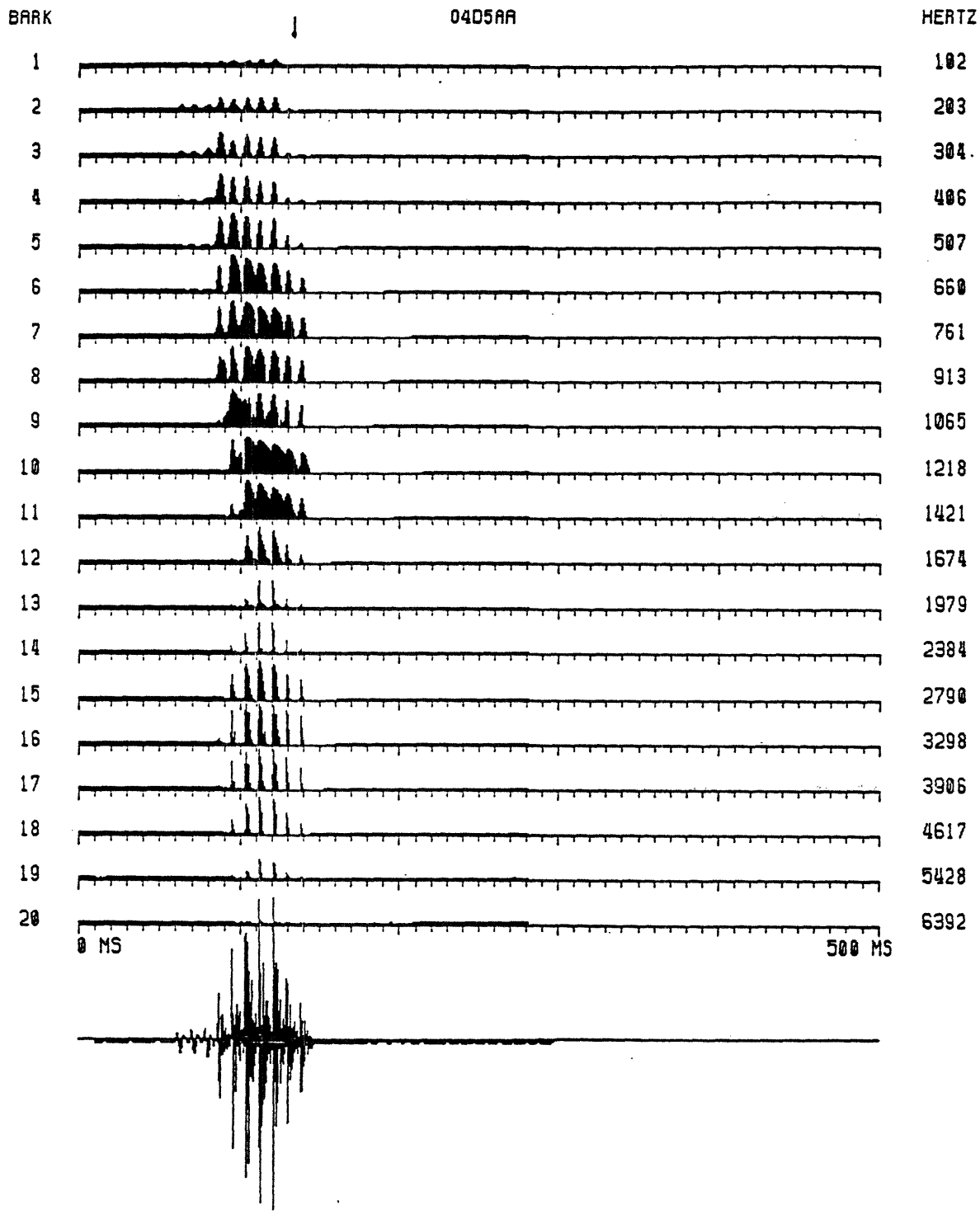

BARK
04D1AA
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.1Ø


BARK
04D5AA
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

LOCATION OF COMPOSITE ONSET AND OFFSET RESPONSE PULSES

PULSE TYPE	APPROXIMATE POSITION (in msec)
onset pulse, all tokens	95 to 100
offset pulse, 87 msec. tokens	120 +- 3
offset pulse, 123 msec. tokens	155 +- 3
offset pulse, 158 msec. tokens	190 +- 3
offset pulse, 228 msec. tokens	260 +- 3
offset pulse, 299 msec. tokens	330 +- 3

FIGURE 3.11

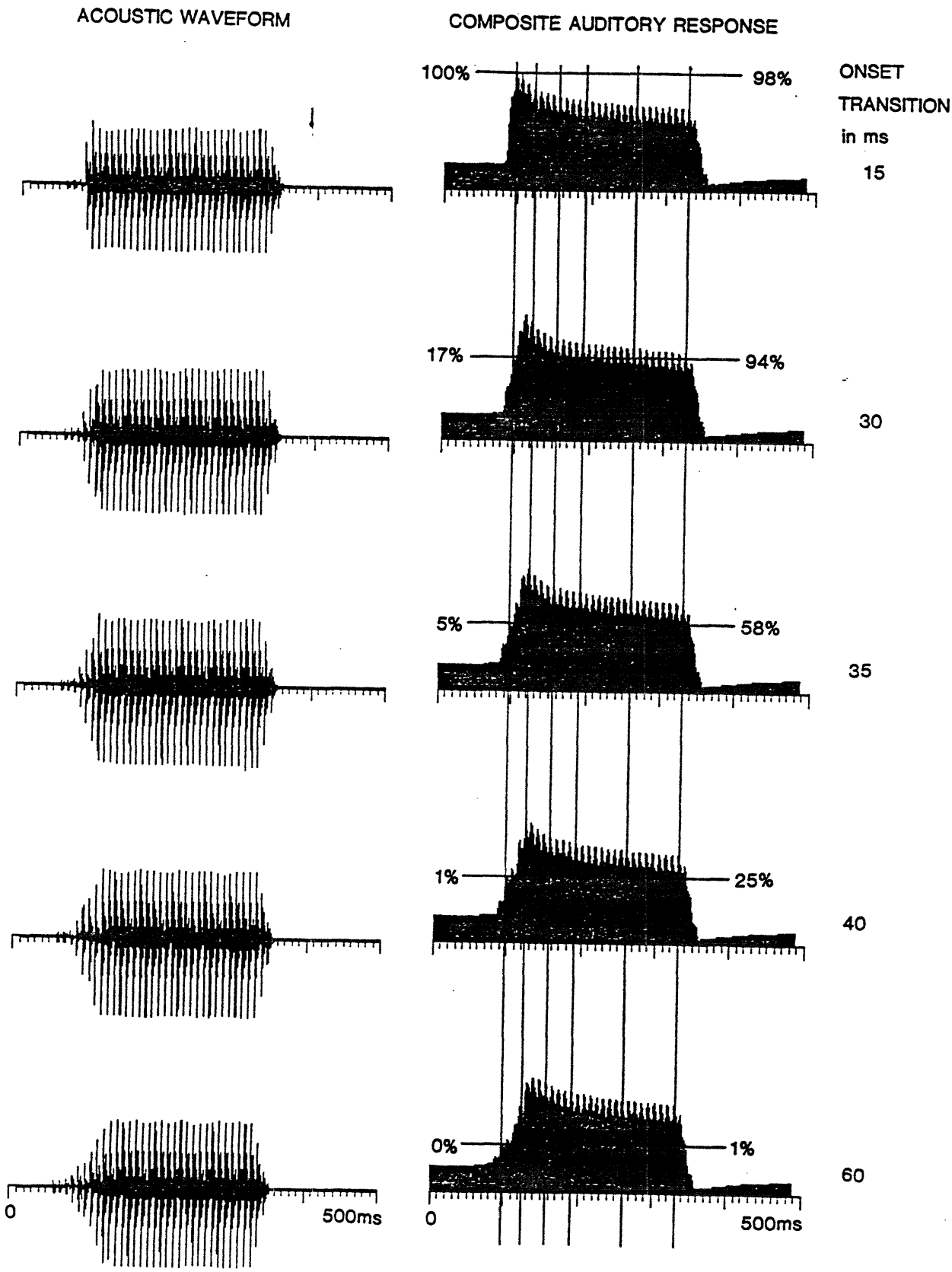

ACOUSTIC WAVEFORM
COMPOSITE AUDITORY RESPONSE
ONSET TRANSITION in ms
100%
98%
15
17%
94%
30
5%
58%
35
1%
25%
40
0%
1%
60
0
500ms
0
500ms

TABLE 3.3

RESULTS OF MULTIPLE REGRESSION ANALYSIS
OF STOP/GLIDE SYLLABLES

Sample Size:	55
Independent Variable (Y):	Z scores of % stop judgements

Acoustic Phonetic Analysis

Dependent Variable 1 (X1):	Onset time in msecs
Dependent Variable 2 (X2):	Syllable duration in msecs

Partial Correlation, Y with X1:	-Ø.92	r2 = Ø.85
Partial Correlation, Y with X2:	Ø.15	r^2 = Ø.Ø2

Multiple R Coefficient for Regression of Y with X1, X2:	Ø.94	R^2 = Ø.88

Auditory Phonetic Analysis

Dependent Variable 1 (X1):	Onset firing rate
Dependent Variable 2 (X2):	Offset firing rate

Partial Correlation, Y with X1:	Ø.93	r^2 = Ø.86
Partial Correlation, Y with X2:	-Ø.18	r2 = Ø.Ø3

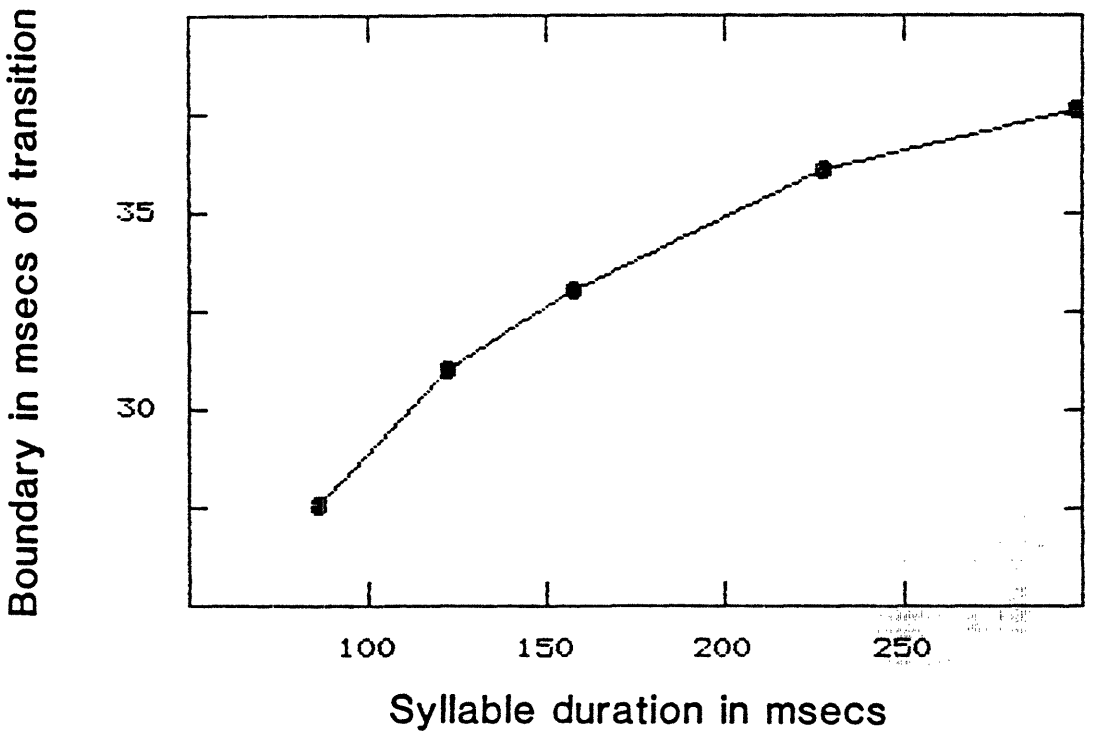
Multiple R Coefficient for Regression of Y with X1, X2:	Ø.97	R2 = Ø.94

Student's t ratio of Y residuals from acoustic and auditory regression:	2.68 with N-3 = 52 df

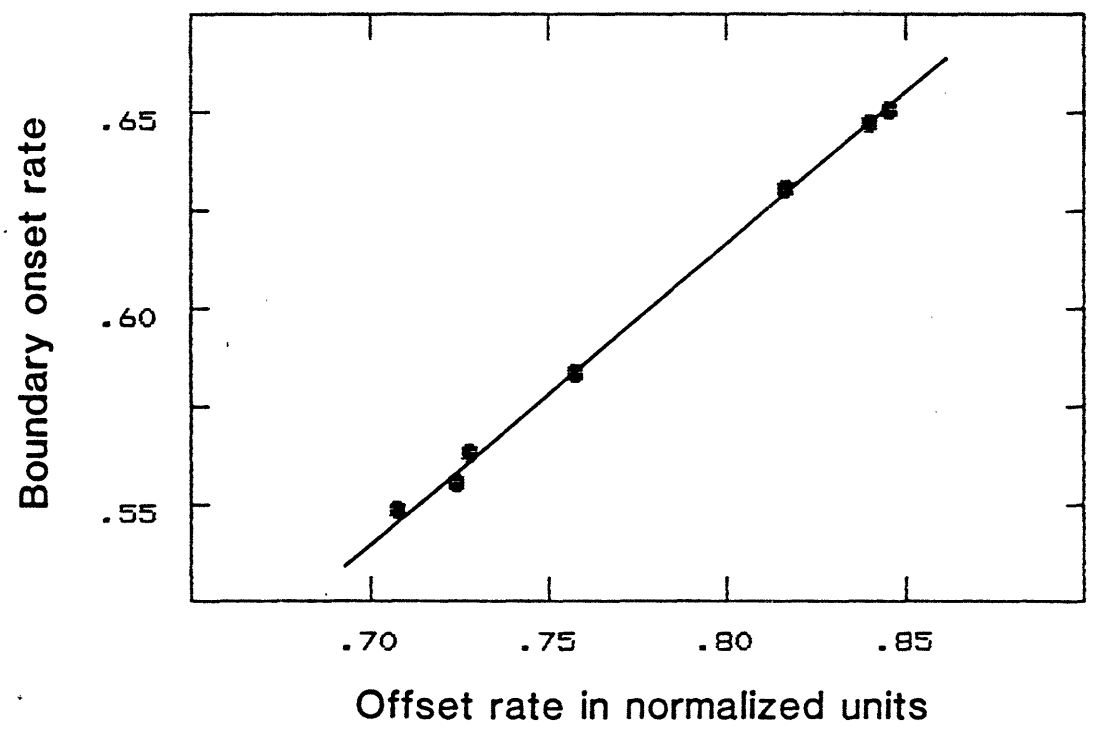
(Rejects alternative hypothesis--that acoustic residuals are less than or equal to auditory residuals--with p < .Ø1)

FIGURE 3.12

A: ACOUSTIC-PHONETIC BOUNDARY FUNCTION



B: AUDITORY-PHONETIC BOUNDARY FUNCTION



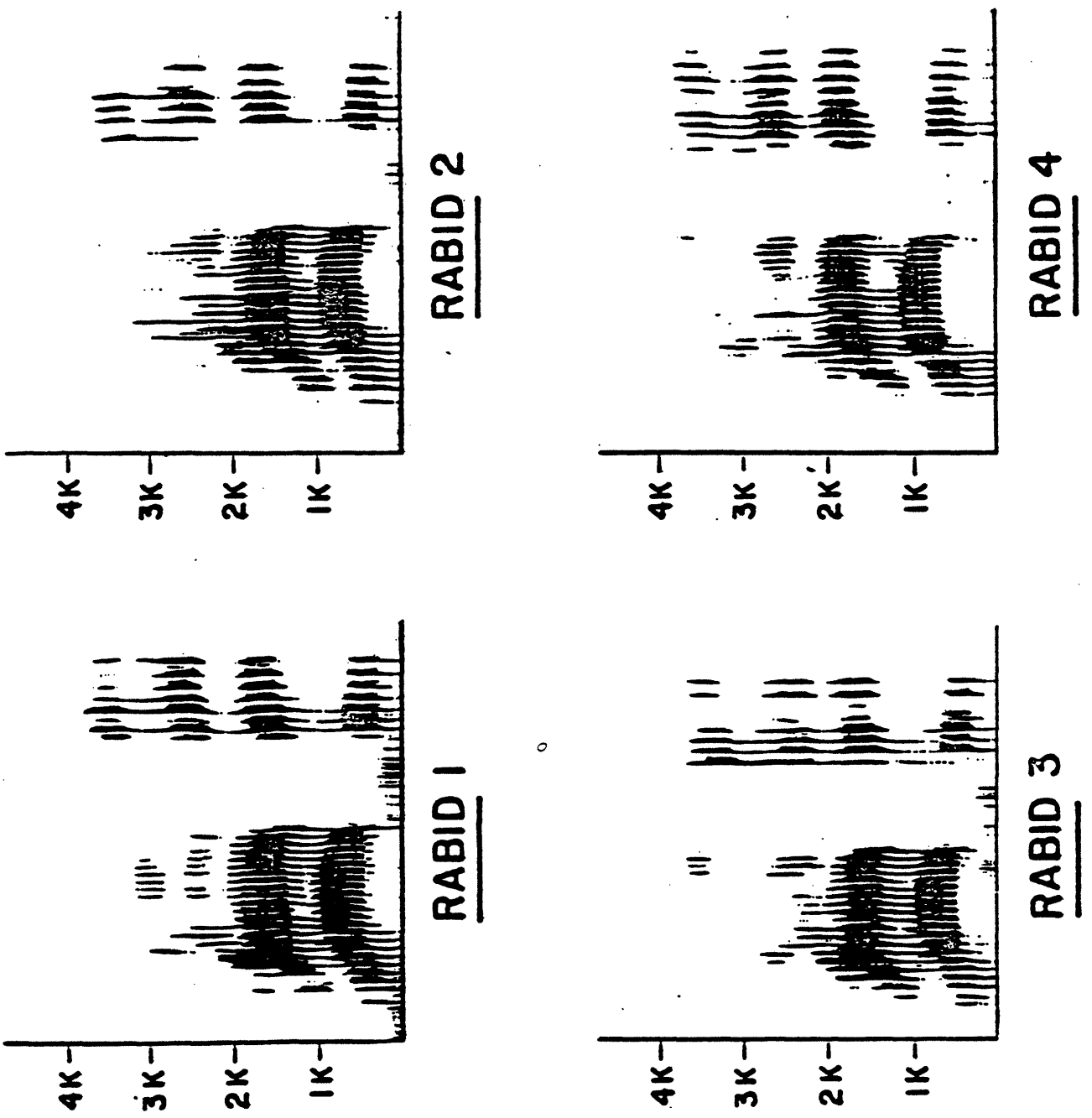
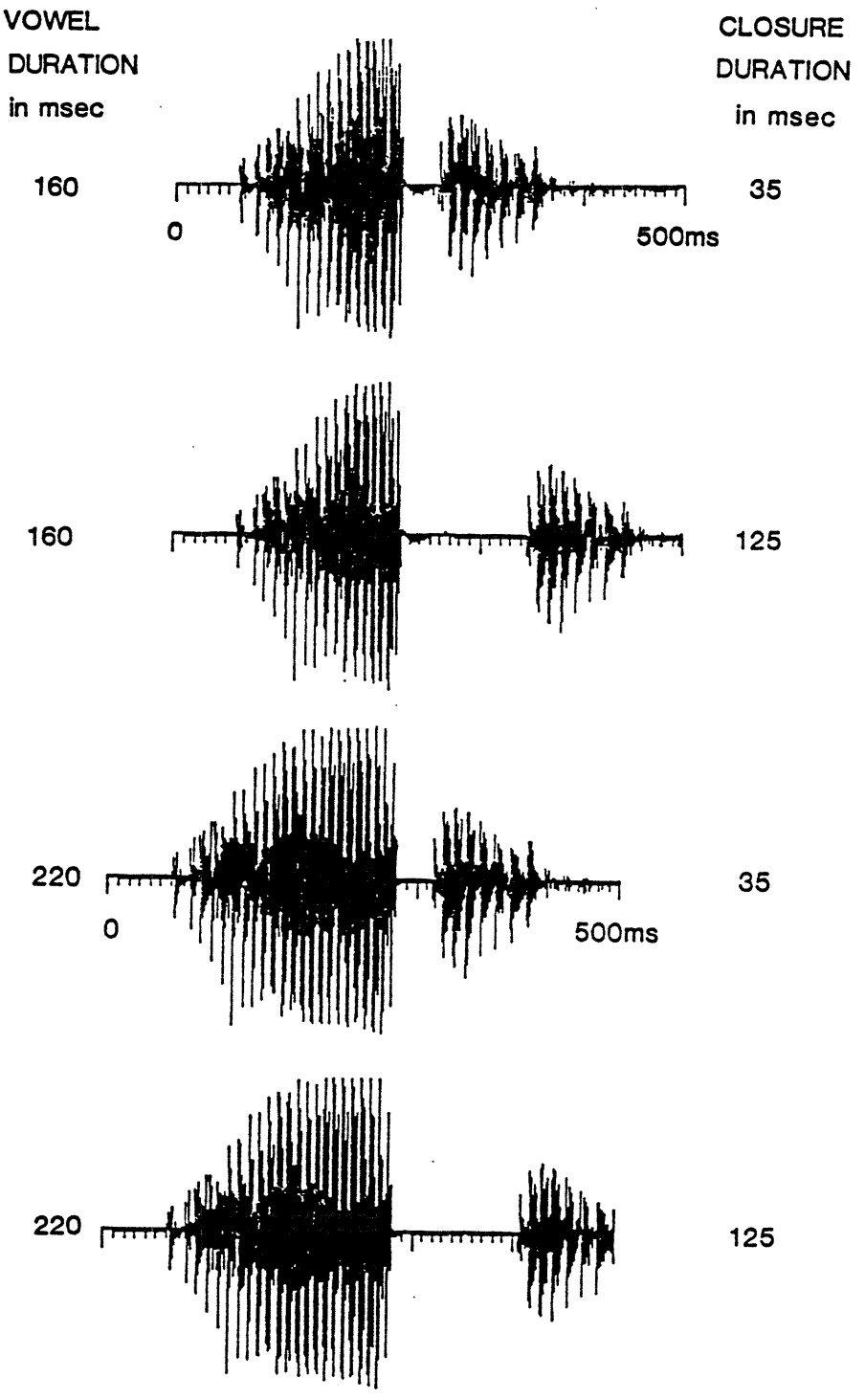

4K
3K
2K
1K
RABID 2
RABID 1
RABID 4
RABID 3

FIGURE 3.14


VOWEL
DURATION
in msec
CLOSURE
DURATION
in msec
160
35
0
500ms
160
125
220
35
0
500ms
220
125

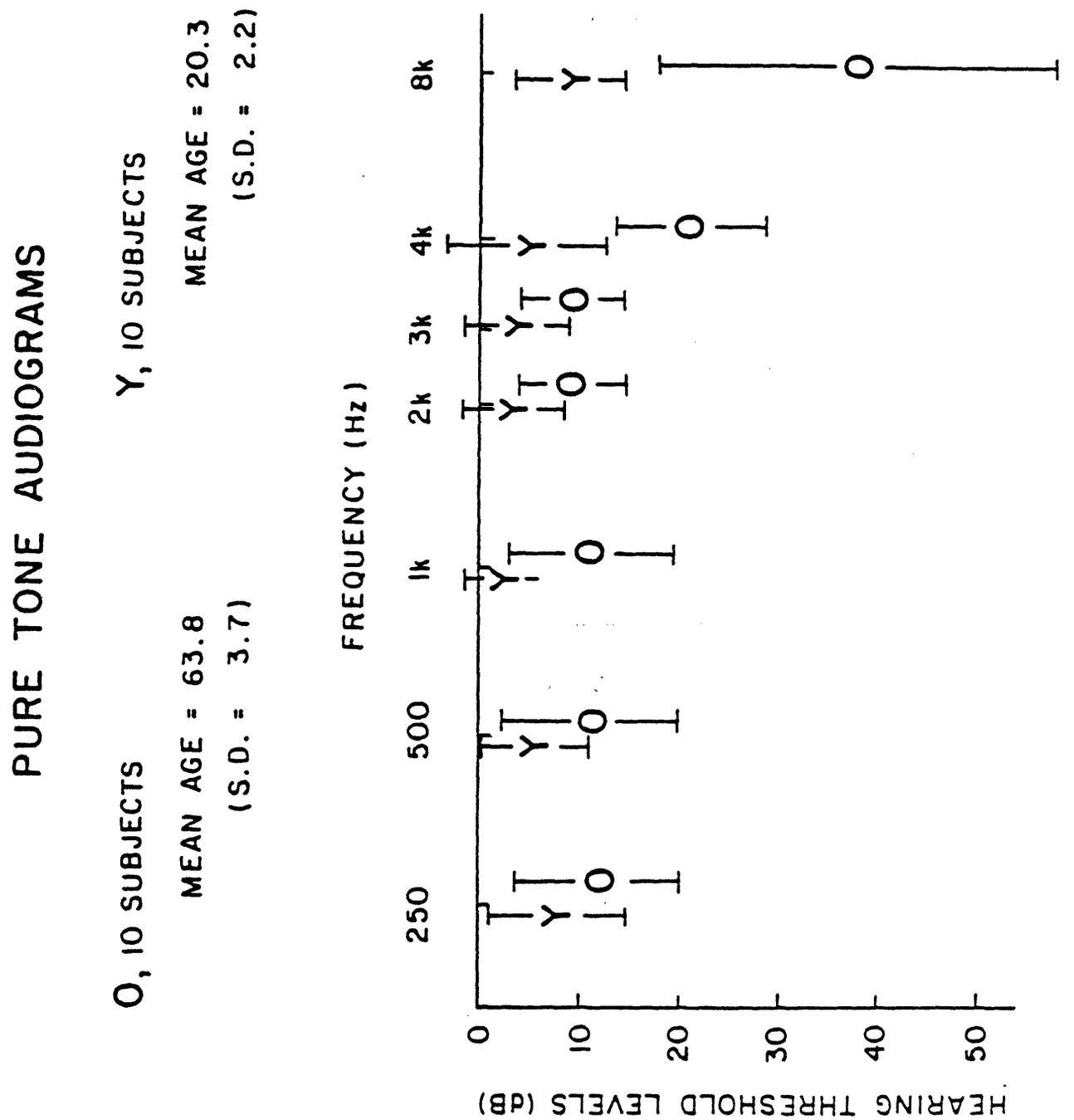

PURE TONE AUDIOGRAMS
O, 10 SUBJECTS
MEAN AGE = 63.8
(S.D. = 3.7)
Y, 10 SUBJECTS
MEAN AGE = 20.3
(S.D. = 2.2)
FREQUENCY (Hz)
250
500
1k
2k
3k
4k
8k
HEARING THRESHOLD LEVELS (dB)
0
10
20
30
40
50

FIGURE 3.16

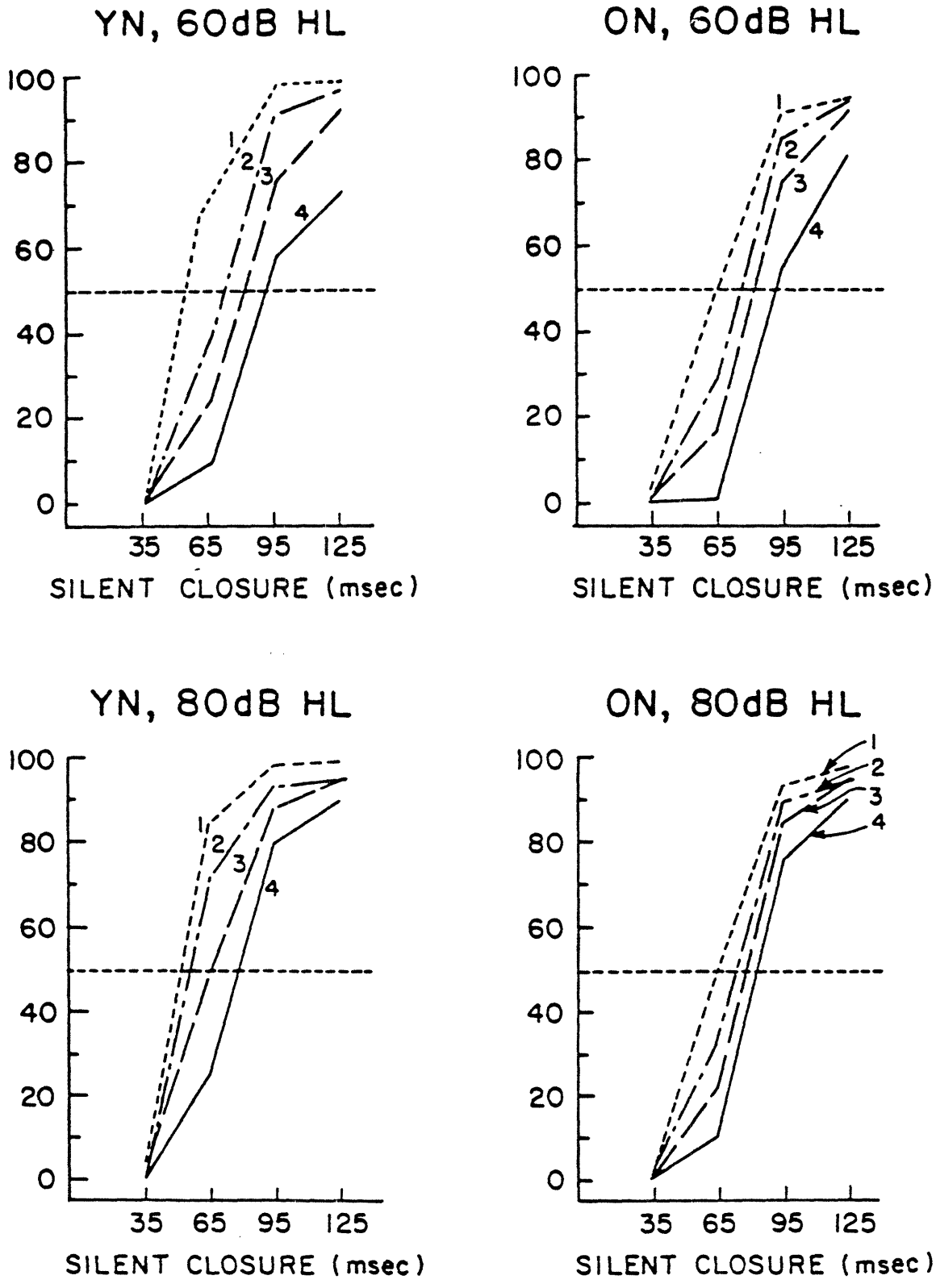

YN, 60dB HL
ON, 60dB HL
YN, 80dB HL
ON, 80dB HL
100
80
60
40
20
0
35 65 95 125
SILENT CLOSURE (msec)
1
2
3
4

FIGURE 3.17

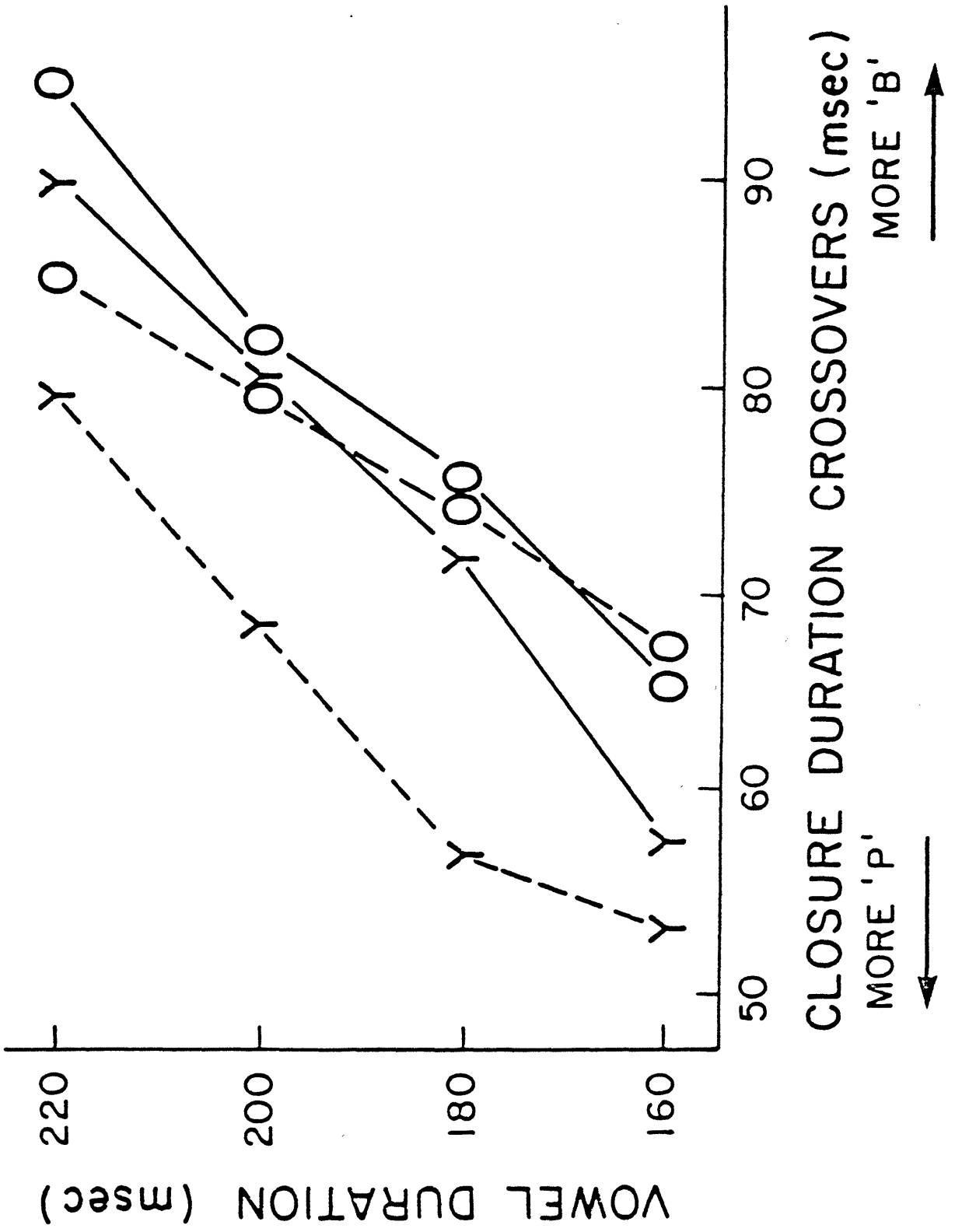

VOWEL DURATION (msec)
220
200
180
160
CLOSURE DURATION CROSSOVERS (msec)
50
60
70
80
90
MORE 'P'
MORE 'B'

FIGURE 3.18

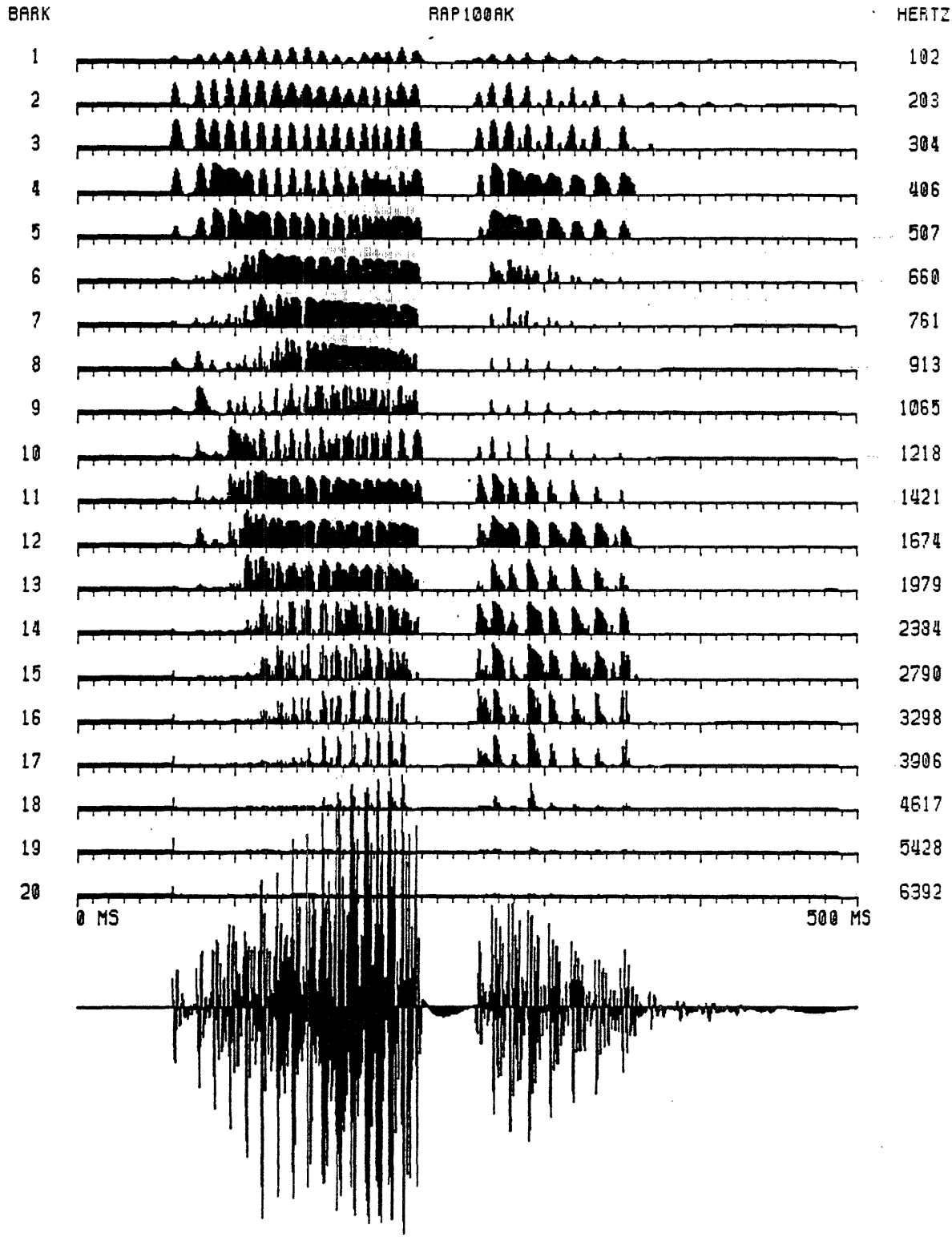

BARK
RAP100AK
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.19

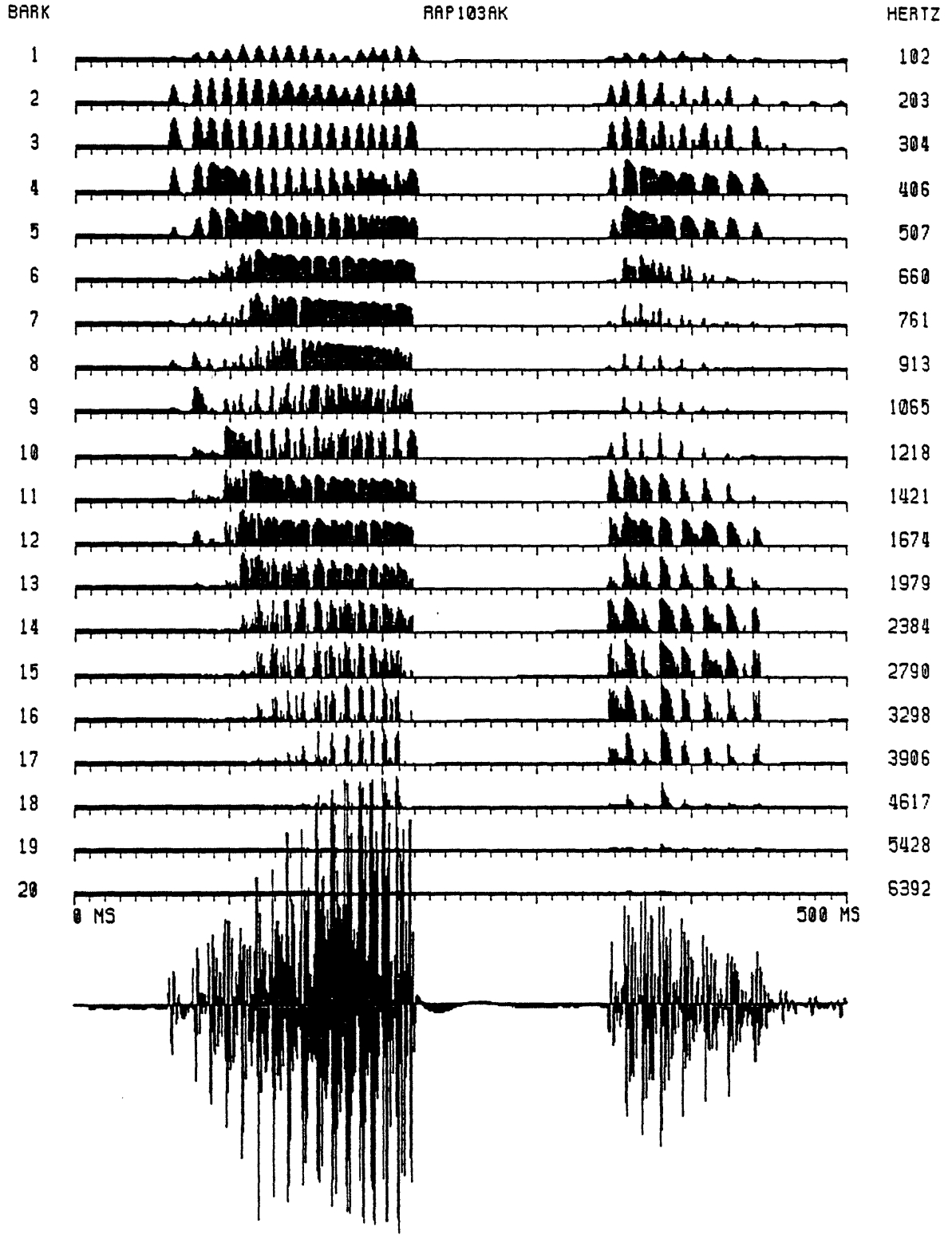

BARK
RAP103AK
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.20

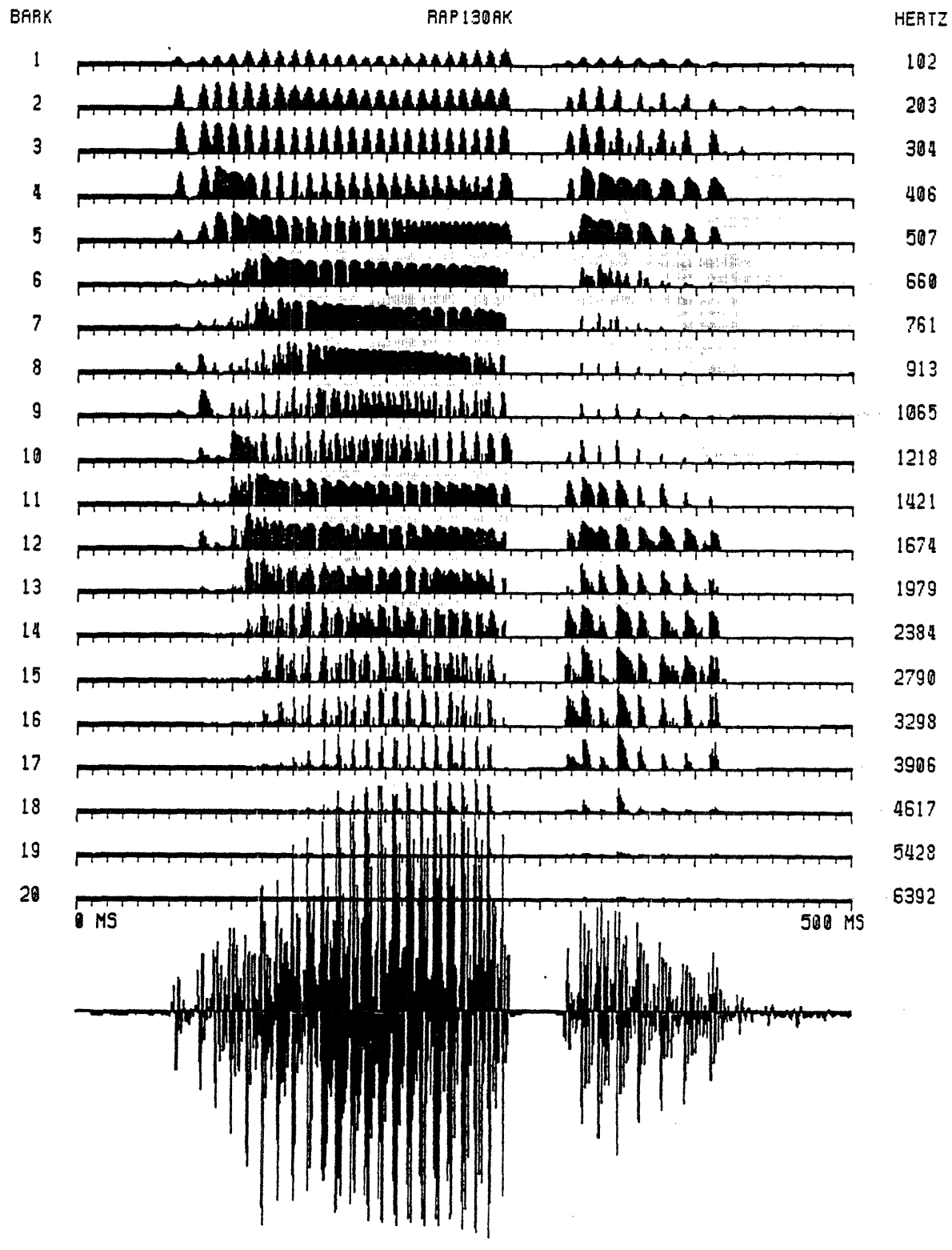

BARK
RAP130AK
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.21

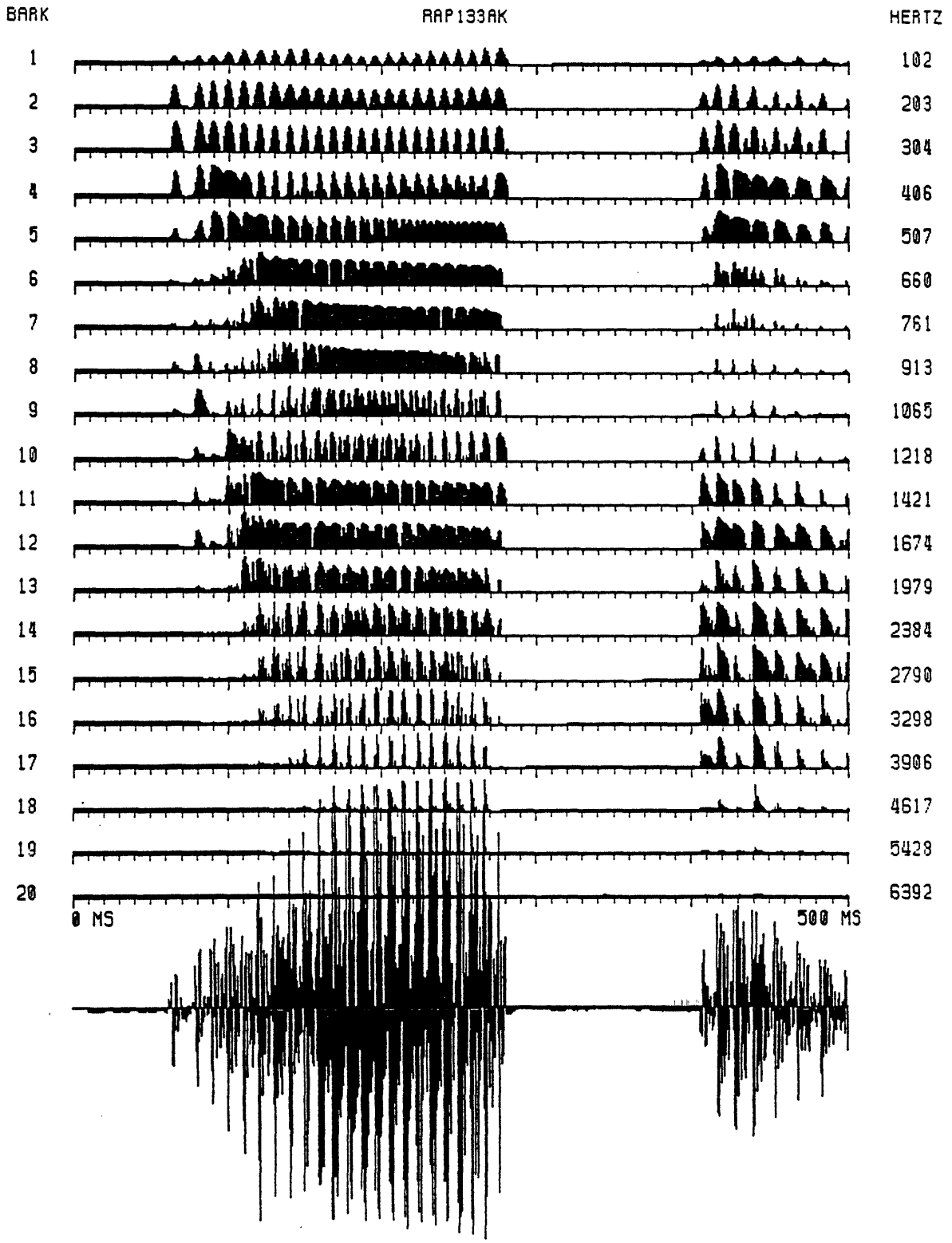

BARK
RAP133AK
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.22

PAM RESPONSE SPECTRA: STOP BURSTS

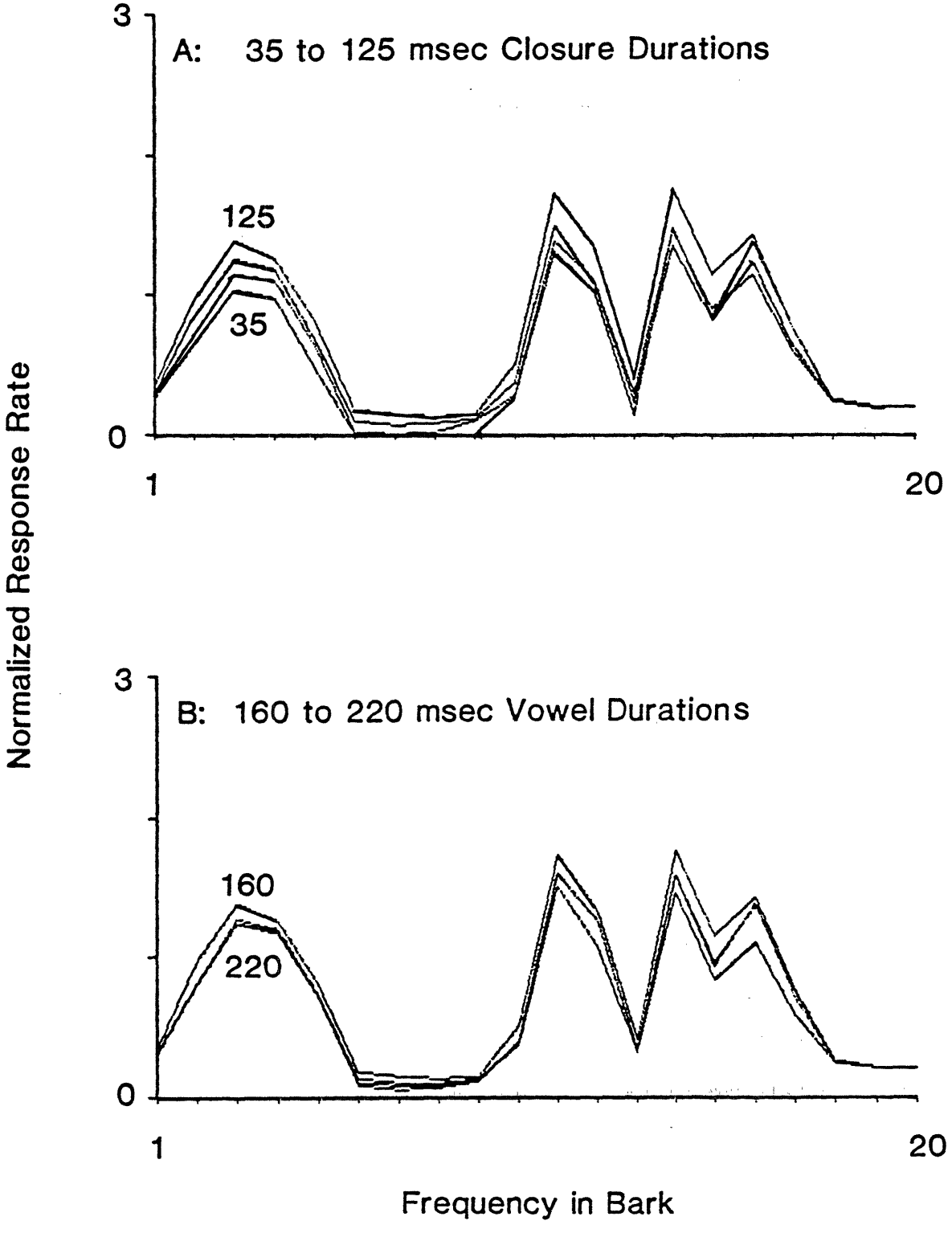


FIGURE 3.23

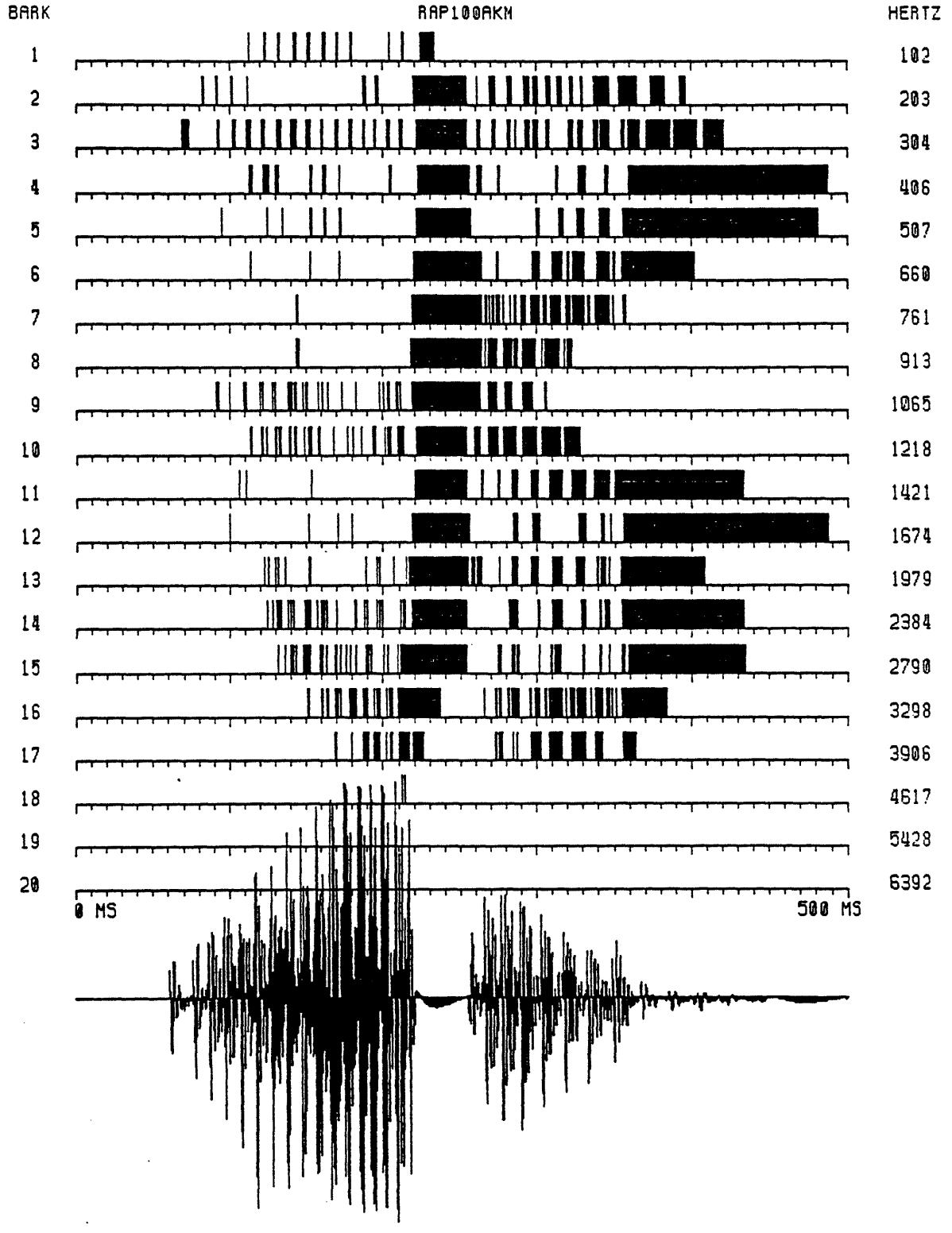

BARK
RAP100AKM
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.24

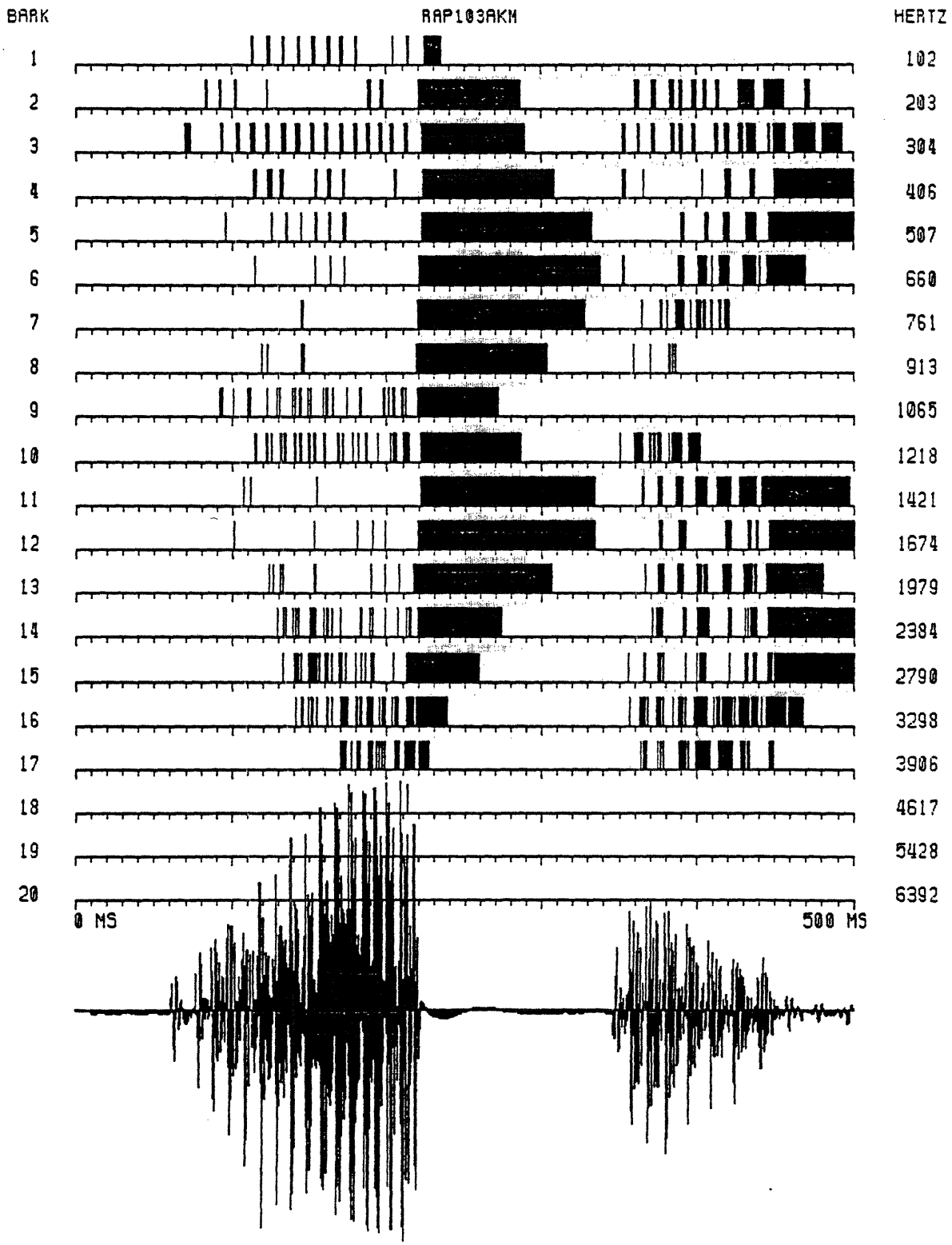

BARK
RAP103AKM
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.25

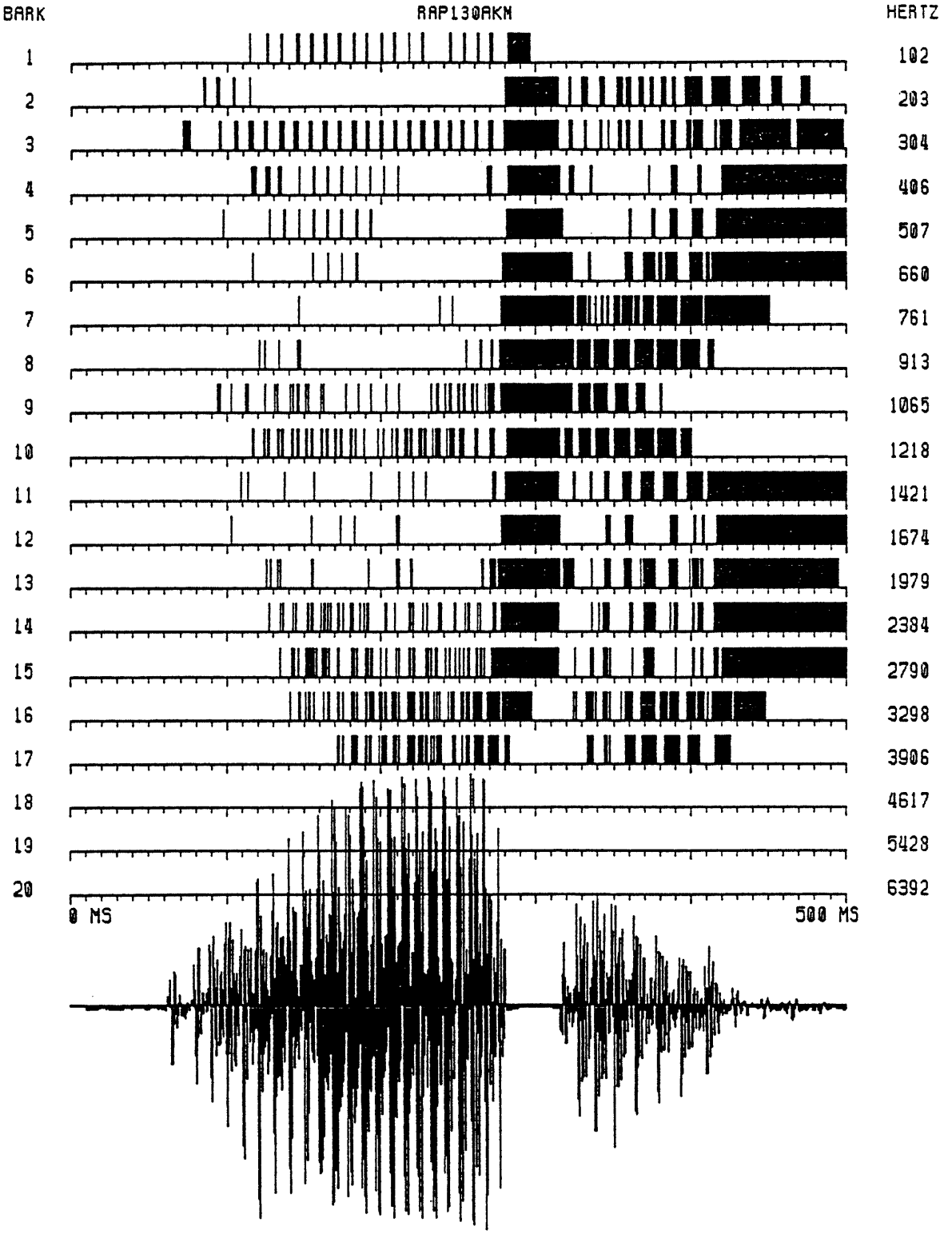
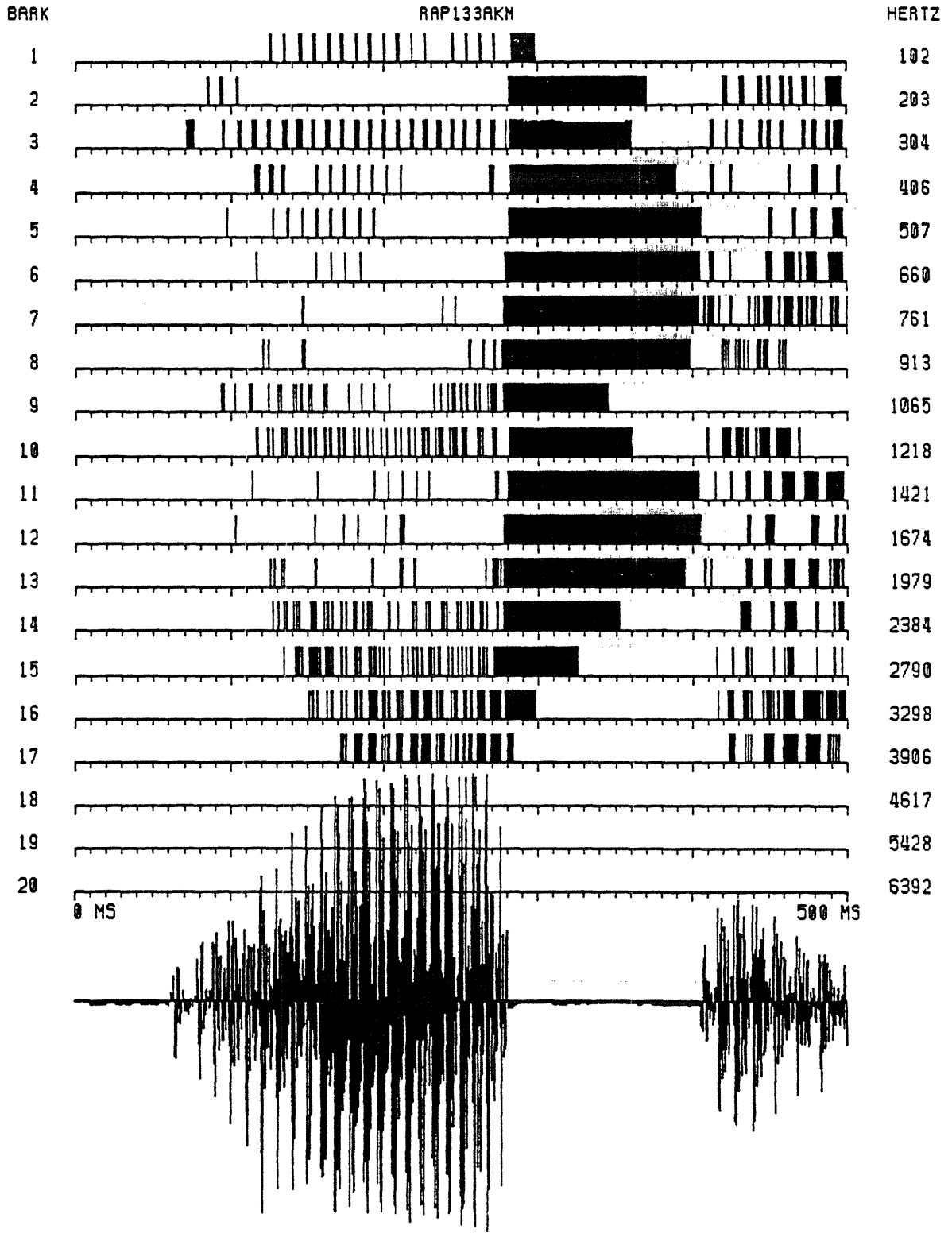

BARK
RAP130AKM
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

FIGURE 3.26


BARK
RAP133AKM
HERTZ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
102
203
304
406
507
660
761
913
1065
1218
1421
1674
1979
2384
2790
3298
3906
4617
5428
6392
0 MS
500 MS

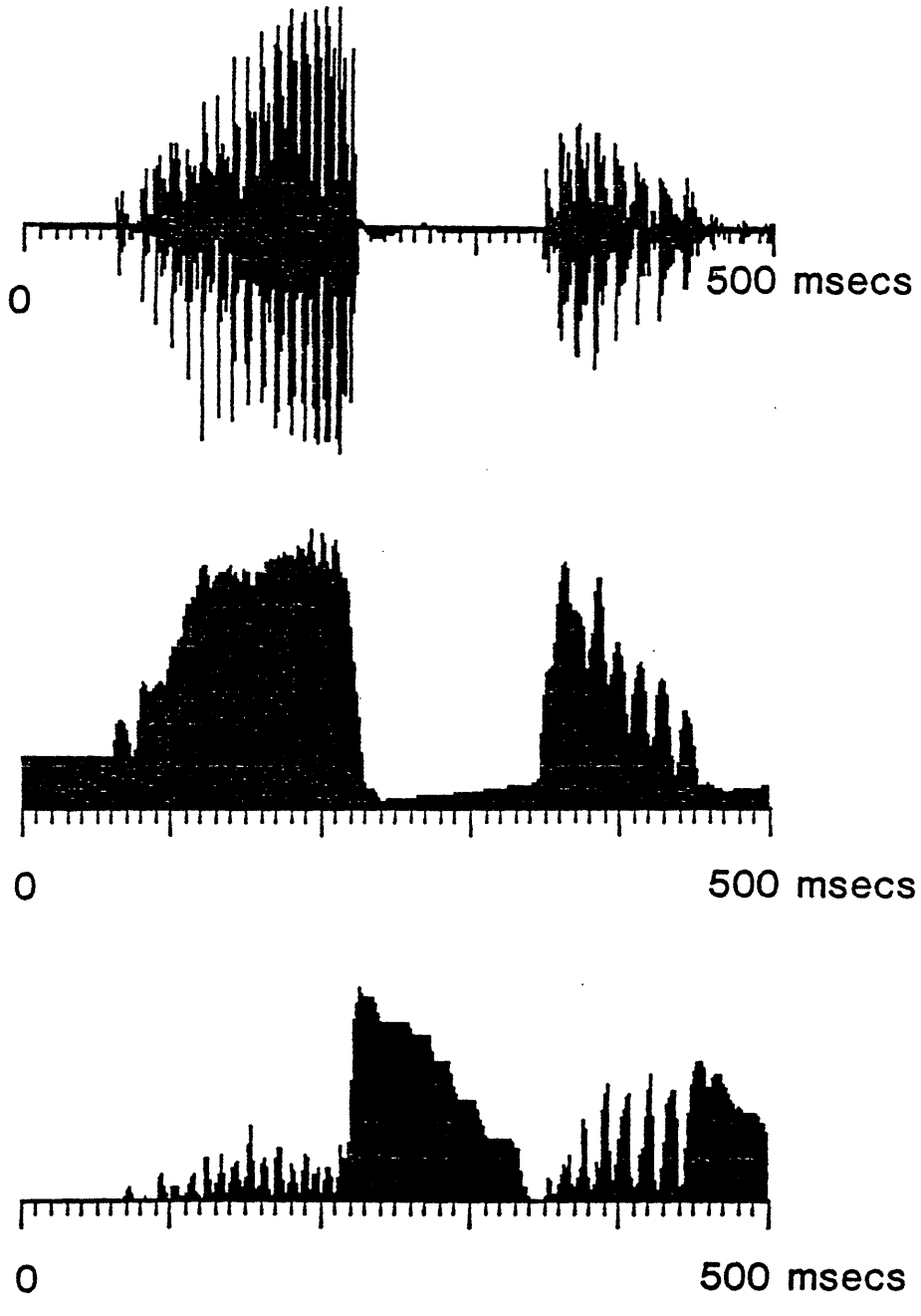
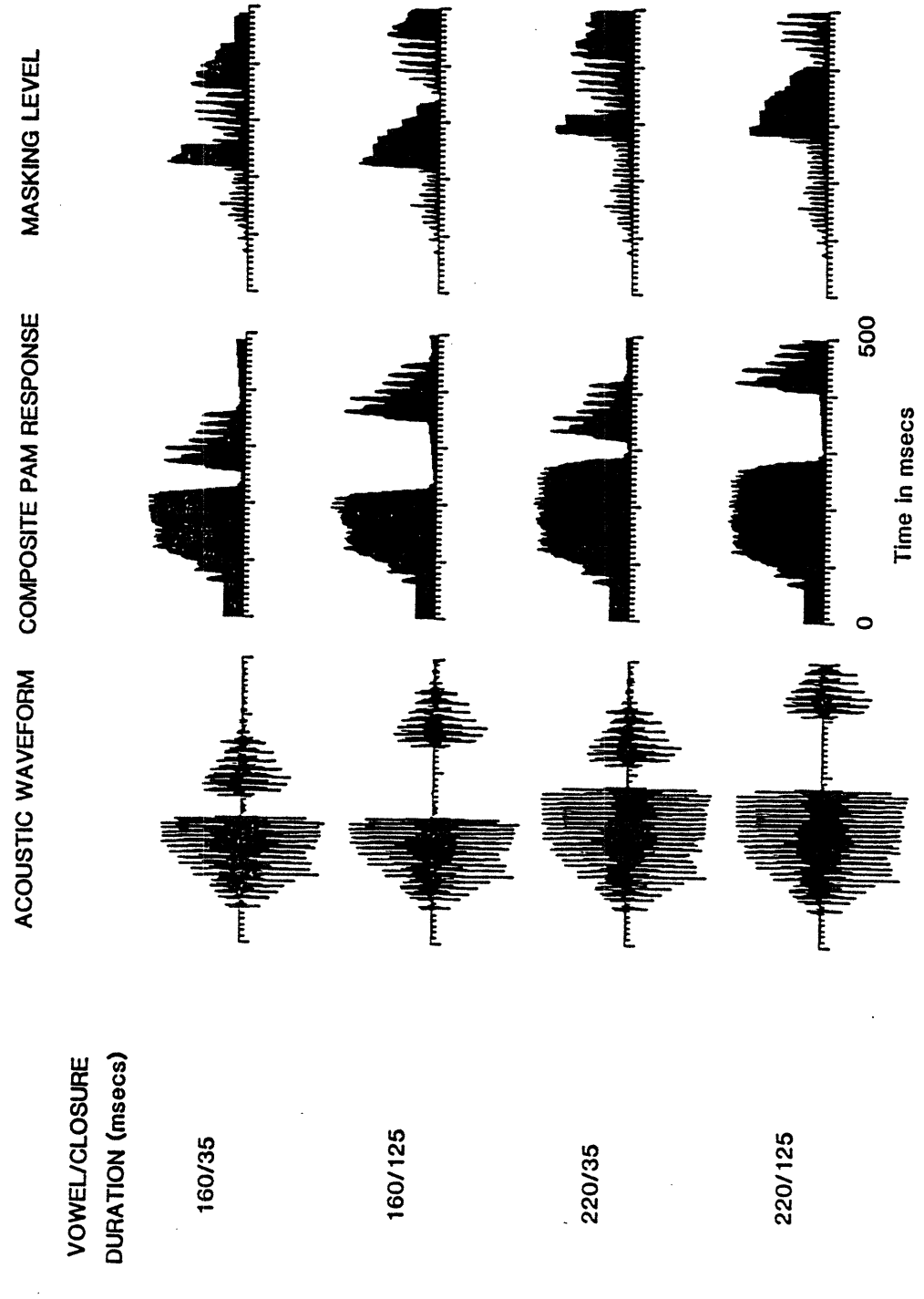

0
500 msecs
0
500 msecs
0
500 msecs

FIGURE 3.28


VOWEL/CLOSURE DURATION (msecs)
ACOUSTIC WAVEFORM
COMPOSITE PAM RESPONSE
MASKING LEVEL
160/35
160/125
220/35
220/125
0
500
Time in msecs

LINEAR FIT OF MASKING LEVEL DURING STOP CLOSURE:

RESPONSE OF PAM FOR YOUNGER SUBJECTS

Token Number	Vowel Duration (msec)	Initial Masking Level	Slope (1/msec) (x .001)	Goodness of Fit	50% Masked Boundary (msec)
1	160	0.89	-6.10	.92	64.6
1	180	0.87	-5.80	.94	64.3
1	200	0.87	-4.84	.93	75.8
1	220	0.87	-4.56	.94	81.4
2	160	0.89	-6.31	.82	61.3
2	180	0.91	-5.96	.84	68.8
2	200	0.91	-5.96	.84	69.0
2	220	0.92	-4.84	.88	87.0
3	160	1.02	-9.67	.94	53.7
3	180	1.06	-8.94	.94	62.1
3	200	1.03	-7.61	.92	69.4
3	220	1.02	-6.74	.92	77.8
4	160	1.03	-10.9	.98	49.1
4	180	1.00	-9.05	.97	55.3
4	200	0.92	-7.23	.97	58.4
4	220	0.90	-6.06	.95	65.5

FIGURE 3.29

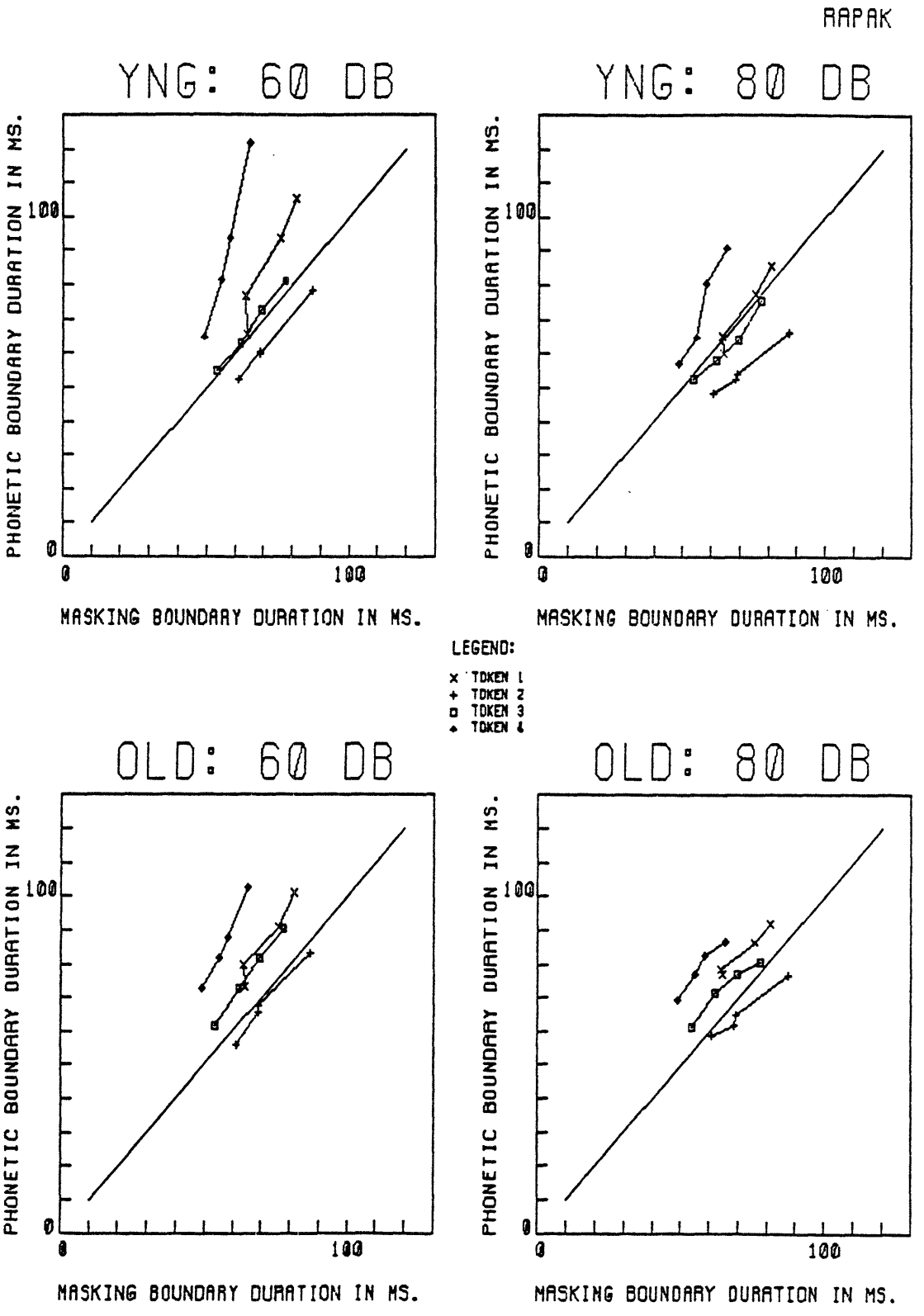

RAPAK
YNG: 60 DB
YNG: 80 DB
PHONETIC BOUNDARY DURATION IN MS.
MASKING BOUNDARY DURATION IN MS.
0
100
LEGEND:
TOKEN 1
TOKEN 2
TOKEN 3
TOKEN 4
OLD: 60 DB
OLD: 80 DB

TABLE 3.5

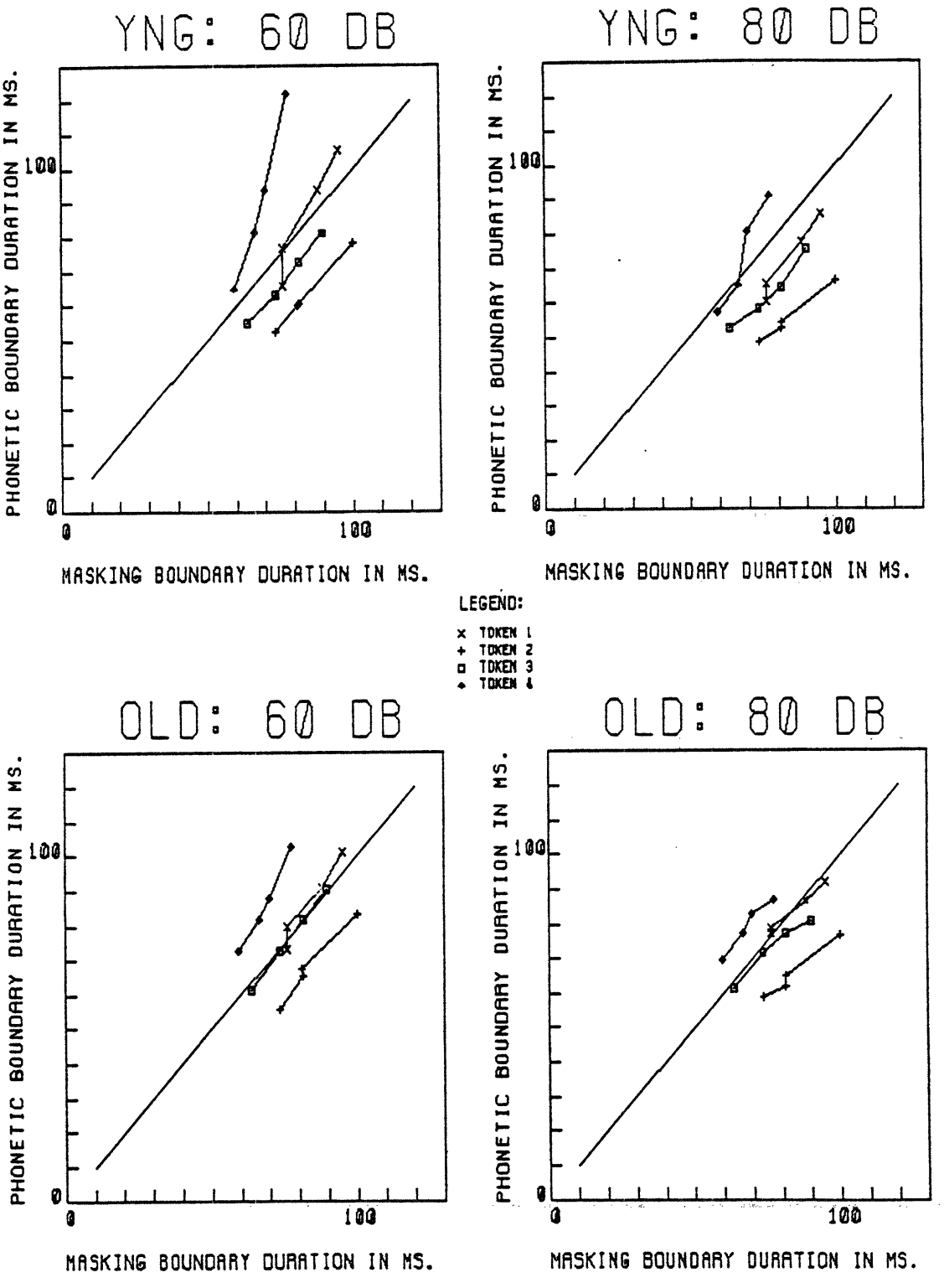
LINEAR FIT OF MASKING LEVEL DURING STOP CLOSURE

RESPONSE OF PAM FOR OLDER SUBJECTS

Token Number	Vowel Duration (msec)	Initial Masking Level	Slope (1/msec) (x .001)	Goodness of Fit	50% Masked Boundary (msec)
1	160	0.89	-5.14	.92	76.2
1	180	0.86	-4.76	.94	76.1
1	200	0.88	-4.32	.93	88.0
1	220	0.87	-3.91	.94	95.0
2	160	0.88	-5.21	.82	73.5
2	180	0.91	-5.06	.84	81.4
2	200	0.92	-5.17	.84	81.3
2	220	0.94	-4.40	.88	100.0
3	160	1.02	-8.14	.94	63.8
3	180	1.07	-7.77	.94	73.6
3	200	1.06	-6.83	.92	81.5
3	220	1.09	-6.60	.92	89.8
4	160	1.04	-9.01	.98	59.5
4	180	1.01	-7.74	.97	66.4
4	200	0.94	-6.25	.97	69.8
4	220	0.93	-5.51	.95	77.6

FIGURE 3.30

RAPAN



CHAPTER 4

EXTENSIONS AND FURTHER EXPERIMENTS

The previous two chapters describe a preliminary examination of the hypothesis that the peripheral auditory system simplifies the encoding of phonetic information in the speech signal. The results seem to indicate that such a hypothesis might indeed be warranted, and that further research is in order. In this chapter I will outline some directions in which the current work could be extended. Two major types of extensions will be described. One is the extension and improvement of the Peripheral Auditory Model. The other is the continued re-examination of acoustic phonetic experiments, particularly those involving temporal effects, from an auditory perspective.

4.1 EXTENSIONS TO THE PERIPHERAL AUDITORY MODEL

The current PAM was developed to test the correspondence theory of peripheral auditory function. Because it was not intended to be a general purpose functional model of the PAS, nor a physiological model, only aspects of the peripheral auditory response that were judged to be most relevant to consonant representation were modeled. In the discussion below, relevance to the representation of speech will continue to be the primary criterion for determining the nature of the

improvements to the current model that will be suggested.

4.1.1 Extending the Spectral Analysis Stage

The PAM should be extended to include the effect of the outer and middle ear on acoustic waveforms. A linear time-invariant filter approximation to these structures should eliminate the need for preprocessing speech signals before using them in the PAM, and would result in a more realistic spectral balance to the PAM response. Typical sources for data on the characteristics of the external and middle ear include Mehrgardt and Mellert (1977), Rabinowitz (1981), Shaw (1974), and Zwislocki (1975).

The current PAM does not model the true phase characteristics and group delay of the basilar membrane. Instead, a linear phase characteristic is used, with a constant group delay across all filter bank channels. This is convenient because it results in a peak response to a burst, for example, that occurs at the same time at the output of each filter bank channel. However, the peripheral auditory system does not respond in this manner. The latency of auditory neural response to an impulse depends on the characteristic frequency of the neuron, with very low CF fibers responding up to 3 msecs later than high frequency fibers (Kiang, 1965).

It would be appropriate to extend the PAM spectral analysis stage to include more accurate cochlear phase characteristics. Pfeiffer and Kim (1975) have derived cochlear phase response data from a Fourier analysis of the discharge patterns of fibers as a function of their position along the basilar membrane. Their data suggest that a linear phase response, with a slope that is dependent on distance along the basilar membrane, is a good model. Figure 4.1 shows the slope of lines fitted by hand to their phase data for fibers at seven locations along the basilar membrane. Although other more recent studies (c.f. Allen, 1983) show much more complicated phase characteristics, even this simple model would result in a group delay that is smaller for higher CF fibers. This in turn might affect the shape of such whole-nerve composite responses as were used in property extraction algorithms in this thesis.

The current PAM spectral analysis stage treats the basilar membrane as an essentially one-dimensional, lumped-parameter, time-invariant linear system, with its position-dependent physical parameters acknowledged only in the use of different parameters for the twenty filter bank channels. All of these assumptions are either open to question or known to be false, and may affect the validity of the spectral response of the PAM to speech. Possible improvements include using transmission line models (e.g., Lyon, 1982; Zweig et al.,

1976) with a larger number of output taps than the 20 channels of the current PAM (Lyon proposes 64 channels); or even the use of two- and three-dimensional models, such as those discussed by Allen and Sondhi (1979), and Steele and Taber (1979a,b).

An even more important extension for speech processing might be to model the nonlinear nature of basilar membrane response. Nonlinearities have been observed in the ratio of middle ear to basilar membrane response at characteristic frequencies (Rhode, 1971), indicating that basilar membrane mechanics might contribute to a nonlinear "sharpening" of the response of fibers to spectral peaks (such as formants) which occur at their CF. Further, nonlinearities involving the interaction of spectral components in complex tones have been observed in both physiological and psychophysical experiments. These include the reduction in response to one tone caused by the presence of another (two-tone suppression), and the generation of a third tone caused by the presence of two primary tones (combination tones). Basilar membrane nonlinearities have been implicated in both of these phenomena (c.f. Rhode and Robles, 1974; Kim et al, 1980). It is reasonable to assume that these nonlinearities would have an effect on the auditory representation of speech sounds containing strong formant frequencies, especially if the formants were close together. A number of nonlinear cochlear models have been

developed (Allen and Sondhi, 1979; Hall, 1977; Kim et al, 1980) which account for these phenomena with varying degrees of success.

4.1.2 Extending the Transduction Stage

In this section we describe some possible extensions to the transduction stage of the PAM. In its current form the combination of the filter bank and transduction stage can be viewed as a channel vocoder (c.f. Rabiner and Schafer, 1978, Sec. 6.7.3). A schematic diagram of such a system is shown in Figure 4.2. Following convolution of the input signal with the impulse response of each channel, the magnitude of the complex output is calculated and decimated. This constitutes the analysis stage of the vocoder. In the PAM, the resulting parametric representation of the speech is then modified in the transduction stage using a nonlinearity function derived from a separate transduction model. This is an example of a common use of vocoders: the transformation of a speech waveform into an alternative representation which is then modified in accordance with some model. Whether or not the speech waveform is then reconstructed (that is, whether or not the resynthesis stage of the vocoder is implemented) depends on the purpose of the vocoding. In our case, the modified channel magnitudes are used directly as input to the adaptation stage, and no resynthesis is necessary.

An important extension of the current PAM would be to model the synchronization of neural discharges to spectral components of the acoustic signal within a critical band around the CF of each fiber. A number of studies have shown that measures of the degree of synchrony of such fibers to various spectral components of vowel sounds result in a good auditory representation of vowel spectra and fundamental frequency (Young and Sachs, 1979; Delgutte and Kiang, 1984a,b,c; Seneff, 1985).

To model this synchrony, both magnitude and phase information about the output of each PAM filter bank channel must be preserved and properly modified by the transduction stage. This is equivalent to converting the filter bank/transduction stage system into a phase vocoder (c.f. Rabiner and Schafer, 1978, Sec. 6.7.2). In a phase vocoder, speech signals are represented by both the magnitude and instantaneous frequency of each filter bank channel output. The original signal can be resynthesized by adding together sine waves that are amplitude and frequency modulated by the vocoder channel parameters. Note that a channel vocoder is simply a phase vocoder with the instantaneous frequency parameters set to the center frequency of each channel. In this proposed extension to the PAM, the resynthesis stage of the vocoder could be used to synthesize the proper haircell response. Various synchrony measures could then be applied to the resulting fine-time sig-

nal. Alternatively, equivalent synchrony values could be calculated directly from the vocoder parameters.

The phase vocoder organization of the transduction stage has another advantage: it forms the basis for a flexible, easily controlled extension in modeling the transduction characteristics of auditory haircells. A schematic diagram of such a system is shown in Figure 4.3. If we assume that the phase vocoder parameters of speech signals vary slowly in comparison to the center frequencies of the filter bank channels, we can treat the output of each channel, to a first approximation, and over a short period of time, as a tone burst. If we have data, derived either from a transduction model or from physiological studies, of the transduction characteristics of haircells in response to tones, we can use that data to modify the vocoder parameters accordingly. In Section 2.3 of this thesis we took the first step in this approach by calculating the DC transfer function of a simple transduction model in response to a tone, and using the resulting transfer function to modify the channel vocoder parameters.

In a similar way, transfer functions specifying the magnitude and phase characteristics of the fundamental and higher harmonic components of the response of a model, or an actual haircell, to sine wave stimuli, can be used to modify both the magnitude and phase parameters of a phase vocoder. Holton and Weiss and their colleagues (1983) have provided exactly this

kind of information regarding the response of alligator lizard haircells, including transfer functions based both on experimental data, and on a model of the alligator lizard cochlea (Weiss and Leong, in preparation). The organization of the transduction stage as a phase vocoder fits well with the specification of transduction nonlinearities using transfer functions of harmonic components. With such a system it would be relatively easy to investigate the effect of specific transfer function characteristics on the response to speech signals, because the transfer functions are used directly to modify the vocoder parameters.

Extending the first two stages of the PAM to a phase vocoder would provide a convenient facility for evaluating extensions to the underlying transduction model. As discussed in Section 2.3, the current model consists of a simple memoryless saturating nonlinearity. The DC transfer functions of this model in response to sine waves with a DC bias exhibit a number of characteristics in common with neural rate-level curves. However, many questions remain unanswered, and some differences do exist between the response of this transduction model and the response of auditory haircells.

One question that deserves more attention is the relationship between the underlying transduction function and the characteristics of the DC transfer function. For instance, preliminary analysis shows that the dynamic range of the DC

transfer function can be expressed analytically, and is independent of the exact transduction function under a variety of reasonable assumptions. But this preliminary analysis also indicates that it might be possible to specify transduction functions with an increased dynamic range. This possibility is of interest because studies by Schalk and Sachs (1980) and Evans and Palmer (1980) indicate that hair cells with low spontaneous firing rates can have dynamic ranges that are at least 10 to 20 dB larger than our simple model, which accurately models the dynamic range of most high spontaneous rate fibers.

A second difference between the current tranduction model and physiological data is the shape of the response to high amplitude stimuli. Holton and Weiss (1983) show receptor potential waveforms in response to high-amplitude, low-frequency tone bursts that fail to square off even 20 dB above the input amplitude at which the size of the response saturates. Under similar conditions of high input amplitude, our transduction model generates a square wave response. This difference in behavior is irrelevant in the current model, but would be an important problem if the PAM were extended to model neural synchrony. The question, then, is which elements of the transduction model must be changed to prevent this?

A third problem with the transduction model is that fibers with a range of spontaneous rates apparently terminate

on the same haircell (Liberman, 1982a,b). Although there is nothing explicit in the transduction model that conflicts with this, there is a general implication that spontaneous rate is a function of some characteristic of the haircell, not of auditory neurons. Suggesting a plausible relationship between the DC bias in the transduction model and some aspect of the haircell-neuron complex would enhance the validity of the transduction model.

A final extension to the current transduction stage of the PAM would be to model the existence of neurons with a variety of relative thresholds. The current model generates a variety of thresholds, but only fibers with a single threshold are actually used. Modeling as few as four different properly chosen thresholds could increase the effective dynamic range of the PAM to 80 dB or more. Since relative threshold and spontaneous rate are inversely related, modeling fibers with a variety of thresholds would result in response patterns exhibiting both high and low spontaneous rates. Since it appears that spontaneous firing--or the lack thereof--may play an important role in the decoding of speech signals, modeling the full variety of spontaneous rates observable in the peripheral auditory system might be an important step.

4.1.3 Extending the Adaptation Stage

The adaptation stage is the most carefully developed stage in the current PAM, but a number of important questions remain. Some of these questions relate to the nature of rapid adaptation. Is it a linear phenomenon, or is it, as Delgutte suggests (1980), more prominent at higher stimulus levels? Is it qualitatively the same as short-term adaptation, or is it related to the refractory period of neurons? Other questions relate to the nature of long-term adaptation. Are there a series of long time constants, or is a single-constant process of some sort a better model? Is short-term adaptation qualitatively the same as long-term adaptation? Answering these questions could result in a better model of an auditory transformation that appears to be very important to speech perception.

4.2 FURTHER AUDITORY PHONETIC EXPERIMENTS

In Chapter Three we reexamined some acoustic phonetic relationships using the PAM. The results indicate that the peripheral auditory system may play a role in the decoding of speech, and suggest that it might be worthwhile to reexamine other acoustic phonetic relationships from an auditory perspective. The following sections present some likely candidates, and indicate the sort of auditory properties which might prove to be related to phonetic content.

4.2.1 Further Temporal Effects

The "rabid-rapid" experiment described in Chapter Three is only one example of a rich collection of phonetic contrasts that can be produced by varying the duration of a silence gap within an utterance. Other minimal pairs that have been studied include "slit-split" (Dorman et al., 1979; Fitch et al., 1980; "say-stay" (Repp, 1981; Best et al., 1981); "shop-chop" (Dorman et al., 1979; Repp, 1981); and "goat-coat" (Repp, 1981).

4.2.1.1 The "Slit-Split" Contrast

In the case of "slit-split", the contrast is produced by introducing a variable amount of silence between the initial [s] and the following glide. In one version of this experiment Dorman et al. report that their subjects heard "slit" when the duration of the silence was less than 60 msecs. For durations between 60 and 250 msecs, a majority of subjects reported hearing "split". As the silence was lengthened still further, the stimulus was more and more likely to be heard as [s] followed by "lit".

Figure 4.4 shows the PAM response to a syllable synthesized in the same way that stimuli were prepared for the Dorman experiment: a male speaker produced the fricative noise [s] and the syllable "lit". Digital recordings of these

tokens were spliced together with 200 msecs of silence between them. Two other stimuli were produced from the same pieces. In one, the pieces were separated by 0 msecs of silence, in the other by 500 msecs. Informal listening revealed that the 0 msec token sounded like "slit", the 200 msec token like "split", and the 500 msec token like [s] followed by "lit".

As Figure 4.4 shows, most of the energy of the initial sibilant occurs in the five highest frequency channels of the PAM. Consequently these channels might be appropriate places to look for auditory properties that cue the presence of a stop. Figure 4.5 shows the original waveform of the three synthesized tokens, and the response of PAM channels 16 through 20, postprocessed and smoothed in the same way that stimuli were in the experiments reported in Chapter Three, except that only the high frequency channels are included. The left panel of Figure 4.5 shows the waveforms of the three stimuli. The middle panel shows the composite response of the high frequency PAM channels. The right panel shows the composite masking level of the same channels.

At the bottom of the middle and right panel are shown listener responses from Dorman et al.: percent of "split" judgements as a function of silence duration. The response data has been scaled and positioned so that for any silence duration of interest, perceptual data and PAM response values are vertically aligned. Throughout the following discussion

it should be remembered that the Dorman perceptual response data are not based on the stimuli shown.

An investigation of the relationship between perceptual responses and PAM responses to these stimuli might begin with the following observations:

1. For silence durations less than 60 msecs, Dorman's subjects reported no stop perceptions. During this same period, the onset response to the glide is substantially smaller than it is for a longer silence duration, because of adaptation to the preceding sibilant. The degree of masking during this period is correspondingly high.

2. For silence durations between 60 and 250 msecs, the percentage of stop judgements monotonically increases. During approximately the same period, the degree of masking in the high frequency channels falls to zero.

3. For silence durations greater than 250 msecs, the percentage of stop judgements decreases steadily, reaching zero at 650 msecs. During this same period of time the spontaneous firing rate of the high frequency PAM channels are approaching their quiescent value.

4.2.1.2 The "Shop-Chop" Contrast

The contrast between an initial fricative or affricate in the words "shop" and "chop" can be affected by the duration of silence preceding the word when it is spoken in a carrier phrase such as "Say _____" (Dorman et al., 1979; Repp, 1981). Silences less than 40 msecs tend to cue the fricative variant. Longer silences cue the affricate. A second cue to this manner distinction is the rise time of the frication (Cutting and Rosner, 1974; Howell and Rosen, 1983). A faster rise time tends to signal the affricate manner, while a longer rise time signals a fricative. Cutting and Rosner show a 40 msec rise time as the phonetic boundary between these two manners; Howell and Rosen found that the boundary was context dependent. Finally, Repp and his colleages (1978, 1981) report that lengthening frication duration results in more fricative responses.

All of these acoustic correlates to the affricate/fricative manner distinction have one thing in common: through adaptation, they affect the amplitude of the peripheral auditory response to the stimulus. Figure 2.27 shows how silence duration, rise time, and stimulus duration interact to affect the timing and magnitude of peripheral auditory response. The acoustic account of the interaction of these cues is confused and highly context dependent. It is possible that the relationship between the corresponding

auditory properties and the perceptual response will prove to be more straightforward.

4.2.1.3 Speaking Rate, Auditory Response, and Phonetic Perception

The experiments reported in Chapter Three concentrate on the auditory and perceptual effects of short term temporal changes: variations in the duration and rise times of the target phoneme itself, or its immediately surrounding acoustic context. A number of experiments have shown that longer range temporal factors can also affect phonetic judgements. For instance, Miller and Grosjean (1981), and Miller, Aibel, and Green (1983) report that the perception of "ba-wa" syllables similar to the ones discussed in Section 3.2 can be affected by temporal changes to adjacent syllables, changes in the duration of more remote syllables in a carrier phrase, and changes in the duration of pauses inserted in the carrier phrase. In general, the more remote the change, the less it seems to affect phonetic perception. Similarly, Port (1977, 1979) showed that the speaking rate of a carrier phrase could affect the voiced/unvoiced stop closure boundary in "rabid-rapid" stimuli. Once again, the size of the boundary shift was relatively small (7 to 10 msecs) compared with the effect of changing preceding vowel duration.

It would be interesting to determine the extent to which these results are predicted by the PAM. If speaking rate influences these and similar phonetic judgements through the medium of long-term adaptation, it might be possible to explain many rate-related effects solely through the response patterns of the peripheral auditory system, without recourse to higher-order auditory processes.

4.2.2 Further Investigations of Auditory Correlates to Voicing in Stops

The acoustic properties which cue the phonetic distinction of voicing in stop consonants have probably received as much attention as any other area of acoustic phonetics. Klatt (1975) identifies six major cues for initial stops: voice onset time, amount of low-frequency energy following the release, burst amplitude, fundamental frequency, prevoicing during the closure, and the duration of the preceding segment. Edwards (1981) identifies ten acoustic properties that can cue intervocalic voicing mode. From an acoustic phonetic point of view, a unified theory of stop consonant voicing is hard to imagine.

A peripheral auditory account of voicing perception might be considerably simpler than the acoustic account. The peripheral response to stimuli which cue voicing distinctions through a variety of acoustic properties could be modeled via

the PAM, and the PAM response patterns examined with an eye to finding simple properties that are related to voicing perception.

A ratio of low frequency energy to non-low frequency energy is a possible characteristic of an auditory phonetic account of stop consonant voicing perception. More low frequency energy (such as prevoicing, and an early onset of the first formant following voicing onset) seems to cue a voiced stop. More high frequency energy (such as a strong burst, or higher fundamental frequency), or less low frequency energy, tends to cue a voiceless stop.

In using any metric based on auditory measures of spectral energy, an important question arises regarding the treatment of masked fibers--fibers that are not firing despite the fact that they have non-zero spontaneous rates. It would be interesting to investigate the consequences of treating these fibers as if they were strongly stimulated. A fiber that is firing at its spontaneous rate is indicating that not much energy is present within its frequency band. A fiber that is firing near its maximum rate is indicating the presence of a large amount of energy. It is unclear how much energy is present within the frequency band of a fiber that is not firing at all despite the fact that it has a high spontaneous rate. There could be very little energy, or there could be a considerable amount: the fiber response is ambiguous. In

fact, the only certain conclusion is that there was a significant amount of energy present in the last 100 msecs or so.

One could, of course, interpret masking as the absence of energy, but an equally plausible alternative is to interpret masking as being equivalent to the presence of energy. Under such a scheme, the output of the PAM masking detector postprocessor used in Section 3.3 would be treated as energy, and could contribute to the ratio of low and non-low frequency energy. It is possible that such a generalization of the notion of auditory spectral balance could greatly simplify an auditory explanation of stop consonant voicing cues.

4.2.3 Auditory Correlates to Place of Articulation in Stops

The final experiment that we shall propose is a study of peripheral auditory correlates to place of articulation in stop consonants. A considerable amount of work has been done to specify the nature of invariant acoustic correlates of place of articulation in word-initial stop consonants. Blumstein and Stevens (1979; Stevens and Blumstein, 1981) proposed that the spectral shape of the stop release burst could reliably identify the place of articulation. As an alternative, Kewley-Port (1983a,b) has proposed substituting a set of dynamic properties for Blumstein and Stevens's static spectral shapes. Kewley-Port's proposed properties remain invariant in the sense that they are independent of the vowel which follows

the stop, but are dynamic in that they reflect the changes in the spectrum during the first 40 msecs following the stop release. In the domain of auditory modeling, Searle, Jacobson, and Rayment (1979) showed that stop consonants in CV syllables could be correctly identified with reasonable accuracy using vowel-independent features extracted from a filter bank whose design was based on auditory characteristics.

Most recently, Lahiri, Gewirth, and Blumstein (1984) have proposed an acoustically invariant metric for place of articulation in diffuse stops (labial versus alveolar and dental) that measures the difference between the spectral balance of the release burst and the spectral balance following the onset of voicing. Simply stated, Lahiri et al. suggest that dental and alveolar stops are characterized by a spectral shape that contains a substantially larger proportion of high frequency energy than the following vowel: either the amount of high frequency energy declines during the transition into the vowel, or the amount of low frequency energy increases. Labial consonants, on the other hand, are characterized by a constant proportion of low and high frequency energy in the burst and the vowel, or an increase in the proportion of high frequency energy during the transition into the vowel. Thus, unlike the earlier Blumstein and Stevens spectral shape property, it is not the absolute spectral shape of the burst that matters, but its shape relative to the shape of the following

vowel--still regardless of the identity of that vowel.

The top panels of Figure 4.6 exemplify the type of metric that Lahiri et al. propose. These panels show the output of the PAM filter bank in response to two tokens of natural speech: /ba/ and /da/ as uttered by a male speaker. (The filter bank output has been smoothed with a 10 msec left half hamming window.) Two spectra are shown in each panel. One is taken from the peak of the release burst. The other is from the second glottal pulse following the onset of voicing. Lines ab in each figure touch the second and fourth formant peaks of the burst spectra. Lines cd in each figure touch the second and fourth formant peaks of the vowel spectra. The vertical lines are positioned at the center frequency of the ninth and sixteenth filter bank channel locations. These are the channels closest to the arbitrary frequencies of 1500 Hz and 3500 Hz which Lahiri et al. choose to use in measuring the slopes.

Thus the slopes of ab and cd are determined by the height and location of the second and fourth formants, and measured by the relative lengths of ca and db. Lahiri et al. chose the ratio

$$r = (d-b)/(c-a)$$

to measure the change in the spectral balance between the burst and the vowel. They found that values of $r > 0.5$, or

negative values resulting from a negative denominator, indicated a labial consonant. For our /ba/ stimulus, r equals 1.8. Positive values of r < Ø.5, or negative values resulting from a negative numerator, indicated an alveolar or dental consonant. For our /da/ stimulus, r equals Ø.55: slightly above the criterion value.

Lahiri et al. construct their metrics from LPC spectra produced from speech windowed by 1Ø msec Hamming windows. The frequency dimension of their spectra is linear in Hz. As described above, our spectra are generated by the PAM spectral analysis stage, and frequency is linear in Bark. In other respects, the spectra in the top panels of Figure 4.3 are similar to those of Lahiri and her colleagues.

The question now arises, what would be the effect on the specified slope ratio r of putting the stimuli through the rest of the PAM? The bottom panels of Figure 4.3 were produced by doing exactly that: they are spectra generated from the output of the PAM, smoothed in the same way that the top spectra were smoothed. The metric r was recalculated for these new spectra. For the /da/ stimulus, r decreases from Ø.55 to -Ø.9: a value that is farther from the criterion point in the alveolar direction. For the /ba/ stimulus, r increases from 1.8 to 2.3: a value that is farther from the criterion point in the labial direction.

These examples suggest that it would be worthwhile to do an experiment to discover whether the peripheral auditory system emphasizes shifts in spectral balance between bursts and following vowels, relative to the shifts seen in acoustic spectra. If so, and if the metric proposed by Lahiri et al. is a reliable metric, the transformations of the peripheral auditory system may serve to make labial and alveolar stops more distinct.

FIGURE CAPTIONS FOR CHAPTER 4

FIGURE 4.1

Slope of phase response of auditory fibers. The data points show the slope of straight lines fitted by hand to phase response data published by Pfeiffer and Kim (1975). The straight line is a linear fit to the six data points representing fibers with characteristic frequencies between 3 and 17 Bark. Note the data point at 0.5 Bark, with a slope of zero.

FIGURE 4.2

Schematic diagram of one channel of a channel vocoder system. $h_i(n)$ is the impulse response of the i'th channel. After the analysis stage, the channel signal could be modified in some way (for example by the PAM transduction and adaptation stages) and then, if desired, resynthesized.

FIGURE 4.3

Schematic diagram of a single channel of the proposed PAM phase vocoder-based transduction stage. The magnitude of the channel band pass filter output is used as the input to the signal modification stage. In this stage, magnitude and phase transfer functions for the harmonic response of an underlying transduction model are indexed by the channel magnitude value. The harmonic magnitude and phase information is used to reconstruct the modified signal using a harmonic synthesizer.

FIGURE 4.4

PAM response to stimulus perceived as "split". The stimulus was created by concatenating [s] and "lit". The nominal center frequencies of each channel in Hertz and Bark are listed along the sides. The vertical dimension is expected firing rate. The vertical distance between the zero lines of adjacent channels corresponds to an expected firing rate 3.0

times the maximum steady state firing rate. The stimulus waveform is shown at the bottom, delayed by 9.9 msecs, the group delay of the PAM filters, so that glottal pulses in the waveform and in the PAM output channels are aligned vertically.

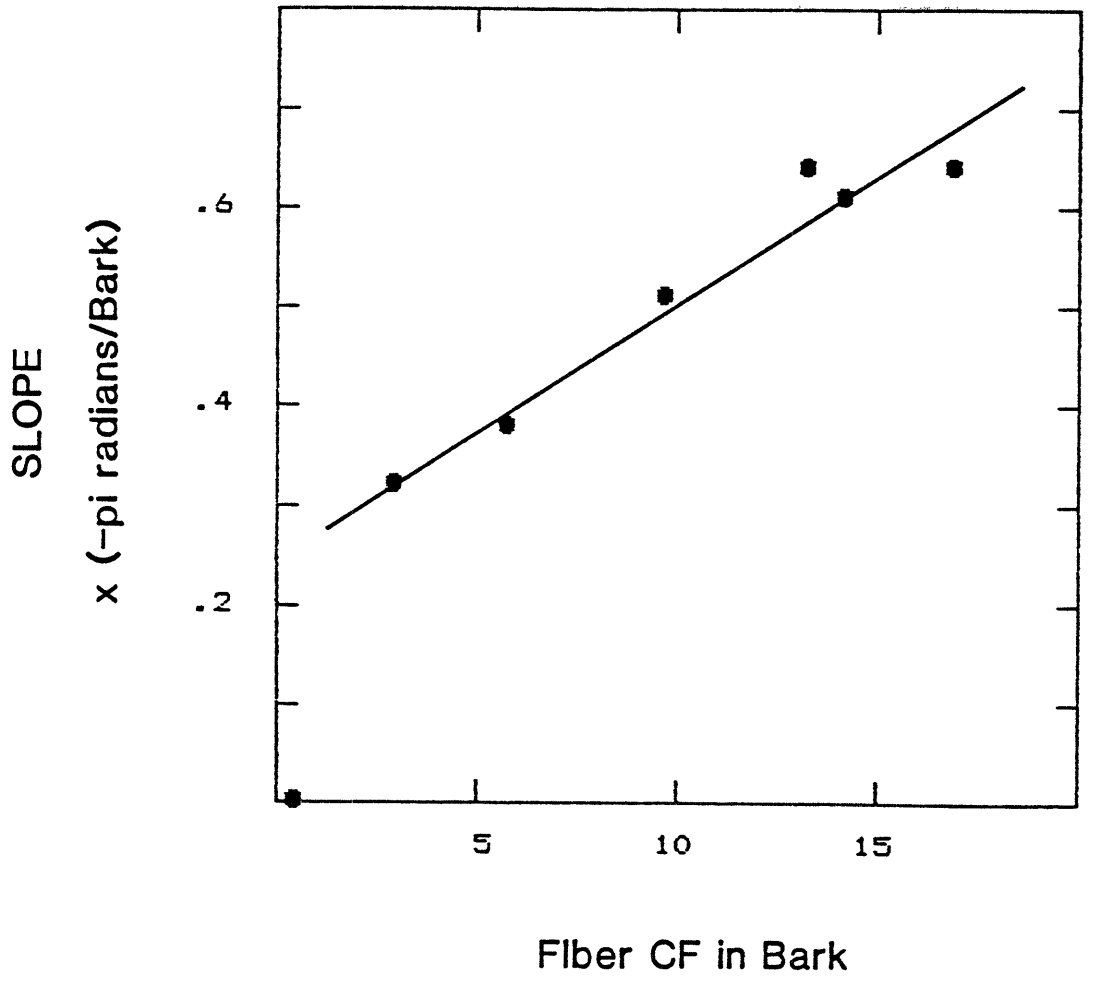
FIGURE 4.5

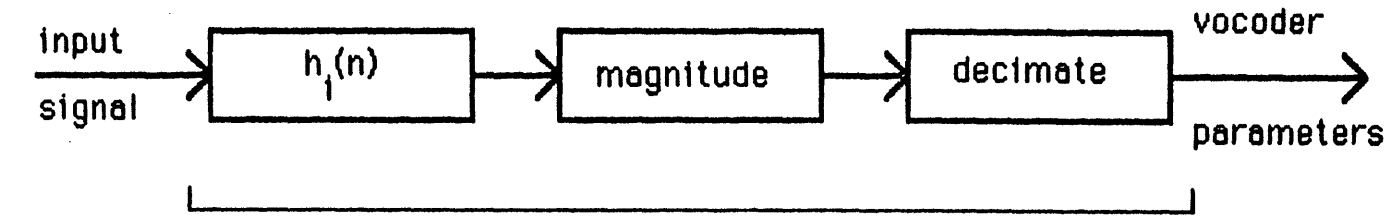
Three representations of "slit/split/s-lit" stimuli. The signals shown were created by concatanating [s] and "lit" with 0, 200, and 500 msecs of silence separating the sibilant from the glide. The signals on the left are the acoustic waveforms. the signals in the middle are the smoothed composite PAM response patterns of the five highest frequency channels, smoothed using a 20 msec half hamming window. The signals on the right are the composite masking detector outputs of the five highest frequency channels, smoothed using a 5 msec half hamming window. The two panels at the bottom of the the picture in the middle and on the right summarize data reported by Dorman et al. (1979) regarding subjects' responses to stimuli similar to those shown here. The panels show the percent of "split" responses as a function of gap duration.

FIGURE 4.6

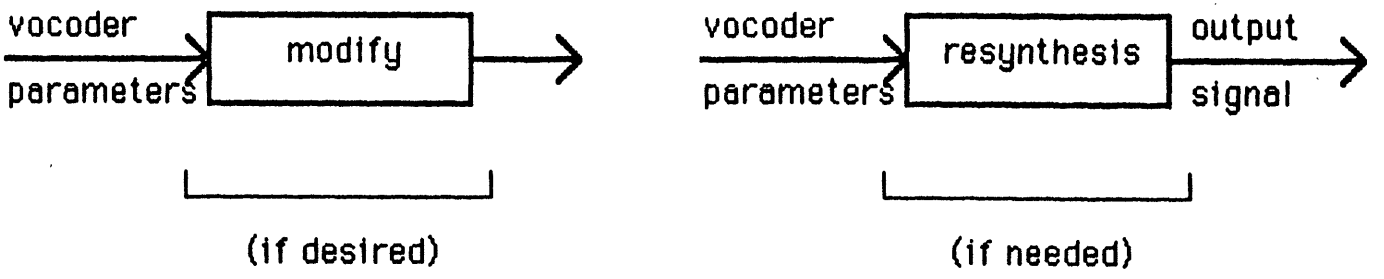
Two representations of the spectral shape of the burst and second glottal pulse following voice onset for /ba/ and /da/. The top panels show the magnitude of the output of the PAM spectral analysis stage. The bottom panels show the output of the PAM adaptation stage. Lines ab touch the F2 and F4 peaks of the burst spectrum. Lines cd touch the F2 and F4 peaks of the second glottal pulse following the onset of voicing.

SLOPE OF PHASE OF FIBER FREQUENCY RESPONSE





ANALYSIS STAGE OF VOCODER



Single Channel of a Channel Vocoder

FIGURE 4.2

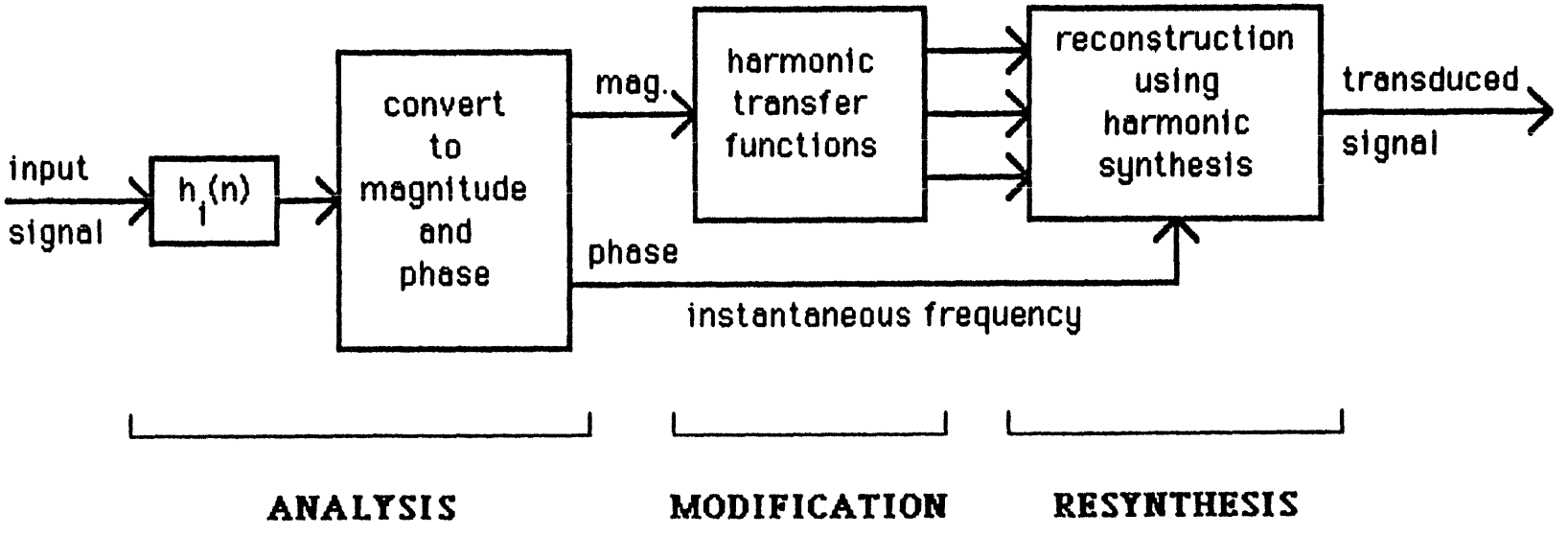


FIGURE 4.3

Single Channel of PAM phase vocoder-based Transduction System

FIGURE 4.4

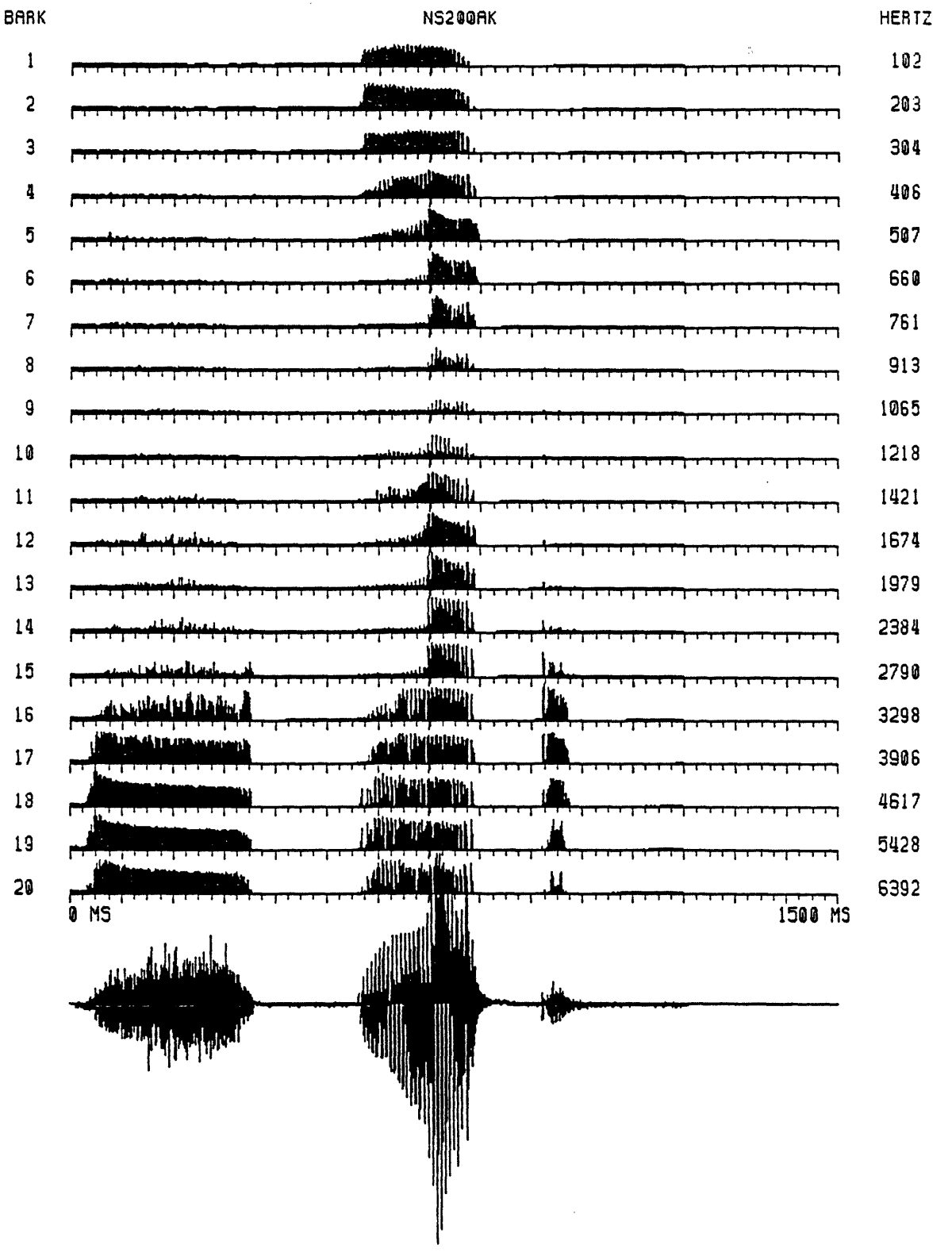
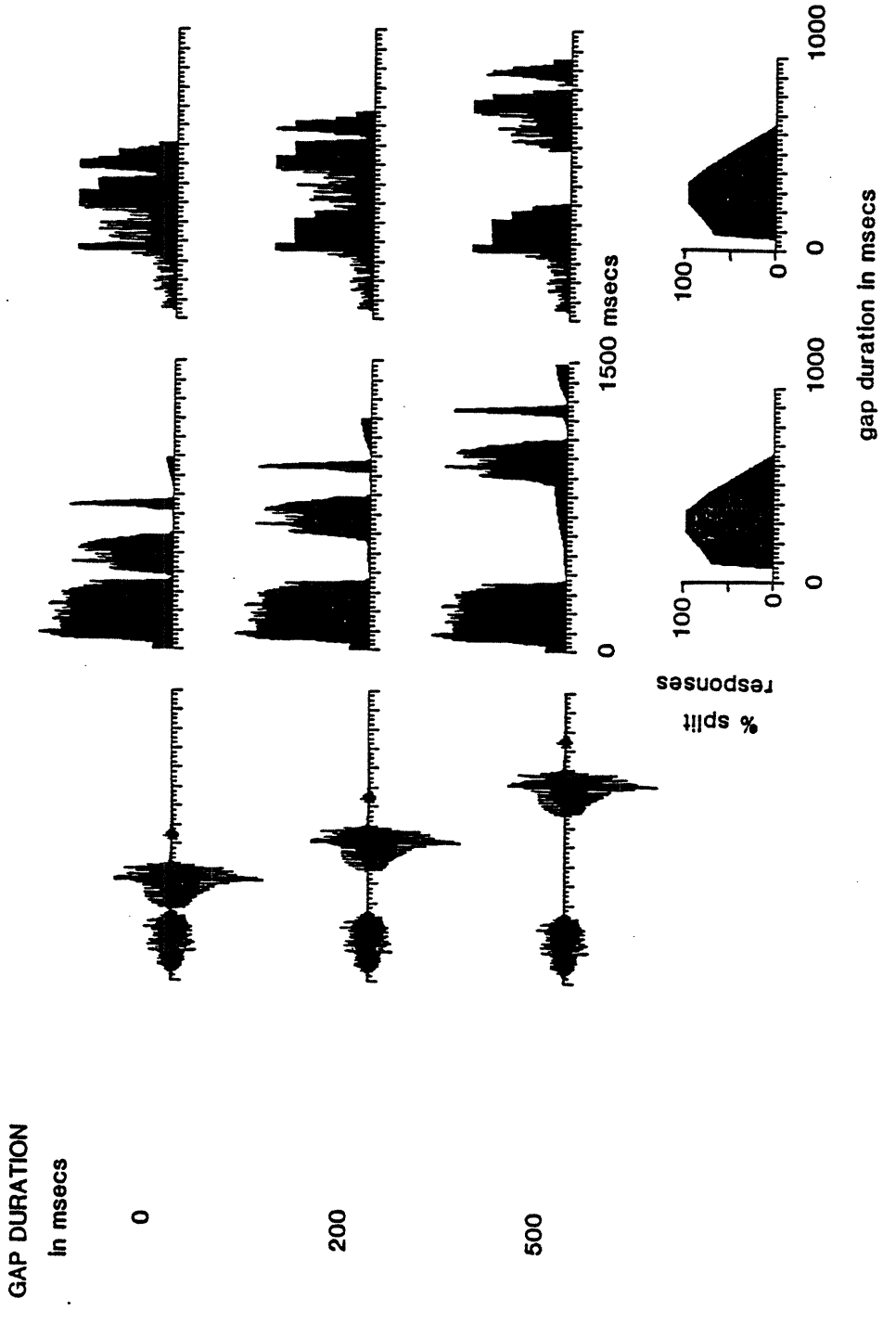

BARK
NS200AK
HERTZ
1
102
2
203
3
304
4
406
5
507
6
660
7
761
8
913
9
1065
10
1218
11
1421
12
1674
13
1979
14
2384
15
2790
16
3298
17
3906
18
4617
19
5428
20
6392
0 MS
1500 MS

FIGURE 4.5


ACOUSTIC WAVEFORM
COMPOSITE PAM RESPONSE
MASKING LEVEL
GAP DURATION
in msecs
0
200
500
0
1500 msecs
% split
responses
100
0
0
1000
gap duration in msecs

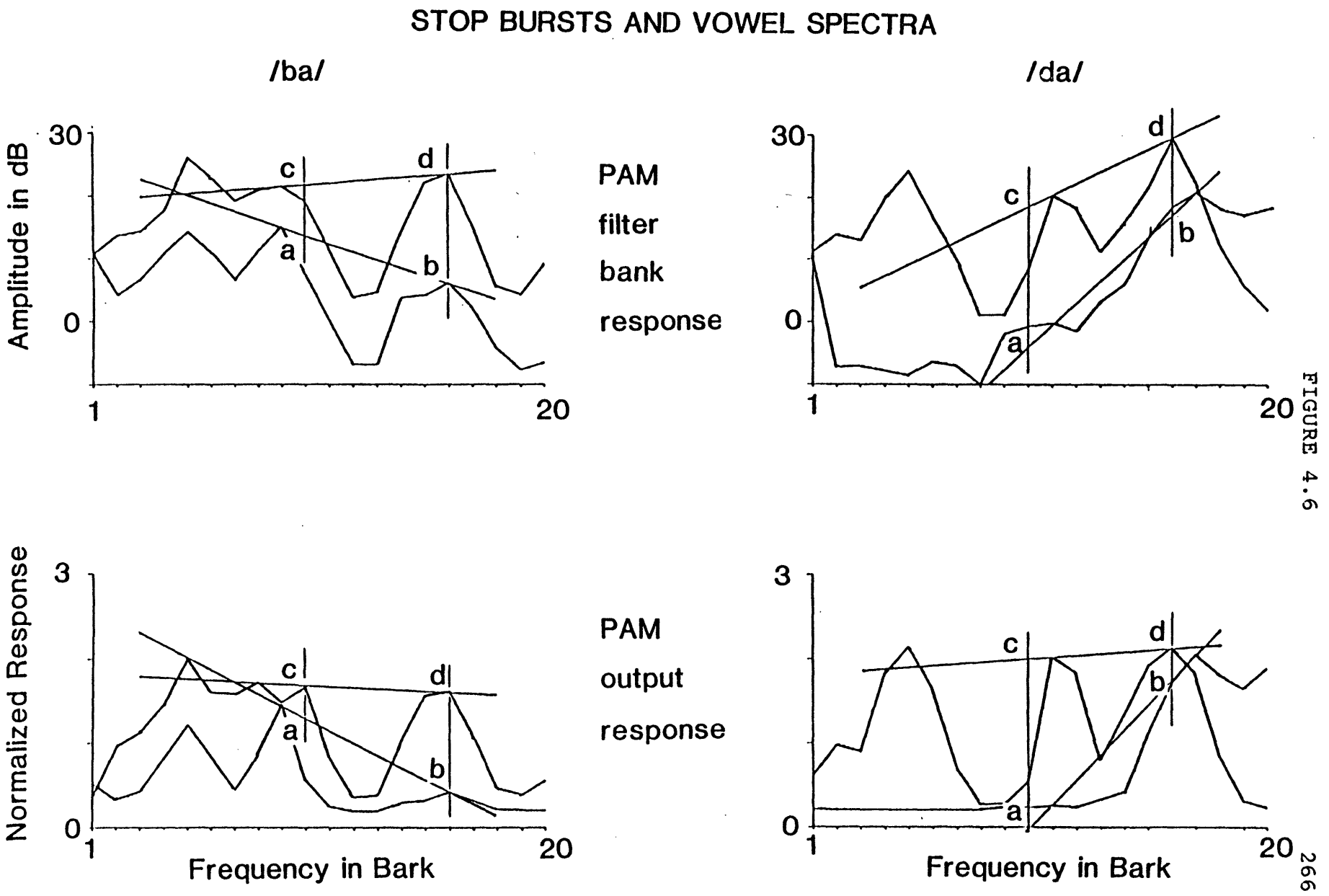

STOP BURSTS AND VOWEL SPECTRA
/ba/
/da/
Amplitude in dB
30
0
1
20
a
b
c
d
PAM
filter
bank
response
Normalized Response
3
0
1
20
Frequency in Bark
PAM
output
response

FIGURE 4.6

BIBLIOGRAPHY

Abbas, P. and M. Gorga (1981), "AP responses in forward-masking paradigms and their relationship to responses of auditory-nerve fibers," JASA **69**(2), 492-499.

Allen, J. B. (1983), "A Hair Cell Model of Neural Response," in Mechanics of Hearing, E. de Boer and M. Viergever, eds., Delft University Press (Martinus Nijhoff Publishers).

Allen, J. and M. Sondhi (1979), "Cochlear macromechanics: Time domain solutions," JASA **66**(1), 123-132.

Allen, J. (1983), "Magnitude and phase--frequency response to single tones in the auditory nerve," JASA **73**(6), 2071-2092.

Art, J. and R. Fettiplace (1984), "Efferent Desensitization of Auditory Nerve Fibre Responses in the Cochlea of the Turtle Pseudemys Scripta Elegans," J. Physiol. **356**, 507-523.

Beasley, D. and J. Maki (1976), "Time- and frequency-altered speech," in Contemporary Issues in Experimental Phonetics, ed. by N. Lass, Academic Press, New York, 419-458.

Best, C., B. Morraongiello, and R. Robson (1981), "Perceptual equivalence of acoustic cues in speech and nonspeech perception," Perception & Psychophysics, **29**(3), 191-211.

Blomberg, M., R. Carlson, K. Elenius, and B. Granstrom (1982), "Experiments with Auditory Models in Speech Recognition," in The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granstrom, eds., Elsevier Biomedical Press, 197-201.

Blomberg, M., R. Carlson, K. Elenius, and B. Granstrom (1984), "Auditory Models in Isolated Word Recognition," Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, San Diego, 17.9.1-4.

Blumstein, S. and K. Stevens (1979), "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," JASA **66**(4), 1001-1017.

Carlson, R, and B. Granstrom (1982), "Towards an Auditory Spectrogram," in The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granstrom, eds., Elsevier Biomedical Press, 109-114.

Chistovich, L. A., M. P. Granstrem, V. A. Kozhevnikov, L. W. Lesogor, V. S. Shupljakov, P. A. Taljasin, and W. A. Tjulkov (1974), Acoustica **31**, 349-353.

Chistovich, L. A., V. V. Lublinskaya, T. G. Malinnikova, E. A. Ogorodnikova, E. I. Stoljarova, and S. Ja. Zhukov (1982), "Temporal Processing of Peripheral Auditory Patterns of Speech," in The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granstrom, eds., Elsevier Biomedical Press, 165-180.

Cutting, J. and B. Rosner (1974), "Categories and boundaries in speech and music," Perception & Psychophysics **16**(3), 564-570.

Delgutte, B. (1980), "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers," JASA **68**(3), 843-857.

Delgutte, B. (1981), Representation of Speech-like Sounds in the Discharge Patterns of Auditory-nerve Fibers, M.I.T. Doctoral Thesis.

Delgutte, B. (1982), "Some correlates of phonetic distinctions at the level of the auditory nerve," The Representation of Speech in the Peripheral Auditory System, ed. R. Carlson & B. Granstrom, Elsevier Biomedical Press, 131-149.

Delgutte, B. and N. Kiang (1984a), "Speech coding in the auditory nerve: I. Vowel-like sounds," JASA **75**(3), 866-878.

Delgutte, B. and N. Kiang (1984b), "Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds," JASA **75**(3), 879-886.

Delgutte, B. and N. Kiang (1984c), "Speech coding in the auditory nerve: V. Vowels in background noise," JASA **75**(3), 908-918.

Dorman, M., L. Raphael, and A. Liberman (1979), "Some experiments on the sound of silence in phonetic perception," JASA **65**(6), 1518-1532.

Durlach, N. (1968), A Decision Model For Psychophysics, Unpublished notes, Department of Electrical Engineering, MIT.

Edwards, T. (1981), "Multiple features analysis of intervocalic English plosives," JASA **69**(2), 535-547.

Evans, E. and A. Palmer (1980), "Relationship Between the Dynamic Range of Cochlear Nerve Fibres and Their Spontaneous Activity," Exp Brain Res **40**, 115-118.

Fant, G. (1960), Acoustic Theory of Speech Production, Mouton & Co., 's-Gravenhage, The Netherlands.

Fitch, H., T. Halwes, D. Erickson, and A. Liberman (1980), "Perceptual equivalence of two acoustic cues for stop-consonant manner," Perception & Psychophysics, **27**(4), 343-350.

Hall, J. (1977), "Two-tone suppression in a nonlinear model of the basilar membrane," JASA **61**(3), 802-810.

Harris, D. and P. Dallos (1979), "Forward Masking of Auditory Nerve Fiber Responses," J of Neurophysiology **42**(4), 1083-1107.

Holton, T. and T. Weiss (1983), "Receptor Potentials of Lizard Cochlear Hair Cells with Free-Standing Stereocilia in Response to Tones," J. Physiol. **345**, 205-240.

Howell, P. and S. Rosen (1983), "Production and perception of rise time in the voiceless affricate/fricative distinction," JASA **73**(3), 976-984.

Kewley-Port, D. (1983a), "Time-varying features as correlates of place of articulation in stop consonants," JASA **73**(1), 322-335.

Kewley-Port, D. (1983b), "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," JASA **73**(5), 1779-1793.

Kiang, N., T. Watanabe, E. Thomas, and L. Clark (1965), Discharge Patterns of Single Fibers in the Cat's Auditory Nerve, Research Monograph no. 35, The M.I.T. Press, Cambridge.

Kiang, N. (1980), "Processing of speech by the auditory nervous system," JASA **68**(3), 830-835.

Kim, D., C. Molnar, and J. Matthews (1980), "Cochlear mechanics: Nonlinear behavior in two-tone responses as reflected in cochlear-nerve-fiber responses and in ear-canal sound pressure," JASA **67**(5), 1704-1721.

Khanna, S. and D. Leonard, (1982), "Basilar Membrane Tuning in the Cat Cochlea," Science **215**, 305,306.

Klatt, D. (1975), "Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters," J of Speech and Hearing Research **18**(4), 686-706.

Klatt, D. (1979), "Speech perception: a model of acoustic-phonetic analysis and lexical access," J of Phonetics **7**, 279-312.

Klatt, D. (1980), "Software for a cascade/parallel formant synthesizer," JASA **67**(3), 971-995.

Klatt, D. (1982), "Speech processing strategies based on auditory models," The Representation of Speech in the Peripheral Auditory System, ed. R. Carlson & B. Granstrom, Elsevier Biomedical Press, 181-196.

Koenig, W., H. Dunn, and L. Lacy (1946), "The Sound Spectrograph," JASA **17**(1), 19-24.

Lahiri, A., L. Gewirth, and S. Blumstein (1984), "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," JASA **76**(2), 405-410.

Landahl, K. and E. Maxwell (1983), "The /b/-/w/ contrast and considerations of syllable duration," Papers from the Nineteeth Regional Meeting, ed. by Chukerman, Marks, and Richardson, Chicago Linguistic Society, 234-243.

Liberman, M. C. (1978), "Auditory-nerve response from cats raised in a low-noise chamber," JASA **63**(2), 442.

Lindsey, P. and D. Norman (1972), Human Information Processing, Academic Press.

Lisker, L. (1957), "Closure Duration and the Intervocalic Voiced-Voiceless Distinction in English," Language **33**(1), 42-49.

Lisker, L. (1978), "Rapid vs. Rabid: A Catalogue of Acoustic Features That May Cue the Distinction," Haskins Laboratories: Status Report on Speech Research SR-54, 127-132.

Ludlow, C., E. Cudahy, and C. Bassich, (1982), "Developmental, age, and sex effects on gap detection and temporal order," JASA **71** Suppl. 1, S47.

Lyon, R. (1982), "A Computational Model of Filtering, Detection, and Compression in the Cochlea," Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, Paris, 1282-1285.

Lyon, R. (1983), "Computational Models of Neural Auditory Processing," Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, San Diego, 36.1.1-4.

Makhoul, J. and J. Wolfe (1972), Linear Prediction and the Spectral Analysis of Speech, BBN Report No. 2304, Bolt Beranek and Newman, Cambridge, MA.

Mehrgardt, S. and V. Mellert (1977), "Transformation characteristics of the external human ear," JASA **61**(6), 1567-1576.

Miller, J. and A. Liberman (1979), "Some effects of later-occurring information on the perception of stop consonant and semivowel," Perception and Psychophysics **25**, 457-465.

Miller, J. and F. Grosjean (1981), "How the Components of Speaking Rate Influence Perception of Phonetic Segments," Journal of Experimental Psychology: Human Perception and Performance **7**(1), 208-215.

Miller, J., I. Aibel, and K. Green (1983), "On the Nature of the Listener's Adjustment for Speaking Rate during Phonetic Perception," JASA, Suppl. 1 **73**, S67.

Moore, B. (1982), An Introduction to the Psychology of Hearing, 2nd edition, Academic Press.

Ohde, R. and K. Stevens (1982), "Effect of burst amplitude on the perception of stop consonant place of articulation," JASA **74**(3), 706-714.

Patterson, R. (1976), "Auditory filter shapes derived with noise stimuli," JASA **59**(3), 640-654.

Pickles, J. (1982), An Introduction to the Physiology of Hearing, Academic Press.

Port, R. (1977), The Influence of Speaking Rate on the Duration of Stressed Vowel and Medial Stop in English Trochee Words, U. of Conn doctoral disseration, published by IU Linguistics Club, Indiana University.

Port, R. (1979), "The influence of tempo on stop closure duration as a cue for voicing and place," Journal of Phonetics **7**, 45-56.

Price, P. and H. Simon (1984), "Perception of temporal differences in speech by 'normal-hearing' adults: Effects of age and intensity," JASA **76**(2), 405-410.

Pfeiffer, R. and D. Kim (1975), "Coclear nerve fiber responses: Distribution along the coclear partition," JASA **58**(4), 867-869.

Rabiner, L. and R. Schafer (1978), Digital Processing of Speech Signals, Prentice Hall.

Rabinowitz, W. (1981), "Measurement of the acoustic input immittance of the human ear," JASA **70**(4), 1025-1035.

Repp, B., A. Liberman, T. Eccardt, and D. Pesetsky (1978), "Perceptual Integration of Acoustic Cues for Stop, Fricative, and Affricate Manner," Journal of Experimental Psychology: Human Perception and Performance **4**(4), 621-637.

Repp, B. (1981), "Phonetic and Auditory Trading Relations between Acoustic Cues in Speech Perception: Preliminary Results," Haskins Laboratories: Status Report on Speech Research SR-67/68, 165-189.

Rhode, W. (1971), "Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys using the Mossbauer Technique," JASA **49**(4), 1218-1231.

Rhode, W. and L. Robles (1974), "Evidence from Mossbauer experiments for nonlinear vibration in the cochlea," JASA **55**(3), 588-596.

Sachs, M. and E. Young (1979), "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate," JASA **66**(2), 470-479.

Schalk, T. and M. Sachs (1980), "Nonlinearities in auditory-nerve fiber responses to bandlimited noise," JASA **67**(3), 903-913.

Schroeder, M., B. Atal, and J. Hall (1979), "Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception," in Frontiers of Speech Communication Research, edited by B. Lindblom and S. Ohman, Academic Press, 217-229.

Searle, C., J. Jacobson, and S. Rayment (1979), "Stop consonant discrimination based on human audition," JASA **65**(3), 799-809.

Seneff, S. (1983), "Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model," Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, San Diego, 36.2.1-4.

Seneff, S. (1985), Pitch and Spectral Analysis of Speech based on an Auditory Synchrony Model, MIT Doctoral Thesis.

Shaw, E. (1974), "Transformation of sound pressure level from the free field to the eardrum in the horizontal plane," JASA **56**(6), 1848-1861.

Smith, R. (1979), "Adaptation, saturation, and physiological masking in single auditory-nerve fibers," JASA **65**(1), 166-178.

Smith, R. and M. Brachman (1980), "Operating range and maximum response of single auditory nerve fibers," Brain Research **184**, 499-505.

Smith, R. and J. Zwislocki (1975), "Short-Term Adaptation and Incremental Responses of Single Auditory-Nerve Fibers," Biol. Cybernetics **17**, 169.

Steele, C. and L. Taber (1979a), "Comparison of WKB caclulations and experimental results for a two-dimensional cochlear model," JASA **65**(4), 1001-1006.

Steele, C. and L. Taber (1979b), "Comparison of WKB caclulations and experimental results for three-dimensional cochlear models," JASA **65**(4), 1007-1018.

Stevens, K. and A. House (1961), "An Acoustic Theory of Vowel Production and Some of its Implications," JASA **4**(4), 303-320.

Stevens, K. and D. Klatt (1974), "Role of formant transitions in the voiced-voiceless distinction for stops," JASA **55**(3), 653-659.

Stevens, K. (1980), "Acoustic correlates of some phonetic categories," JASA **68**(3), 836-842.

Stevens, K. and S. Blumstein (1981), "The Search for Invariant Acoustic Correlates of Phonetic Features," Perspectives on the Study of Speech, ed. P. Eimas & J. Miller, Lawrence Erlbaum Associates, 1-38.

Weiss, T. and R. Leong (in preparation), "A Model for Signal Transmission in an Ear Having Hair Cells with Free-Standing Stereocillia: IV. Mechanoelectric Transduction Stage".

Young, E. and M. Sachs (1973), "Recovery from sound exposure in auditory-nerve fibers," JASA **54**(6), 1535-1543.

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

Zhukov, S. Ya, M. G. Zhukova, and L. A. Chistovich (1974), "Some new concepts in the auditory analysis of acoustic flow," Sov. Phys. Acoust. **20**(3), 237-240.

Zue, V. (1976), Acoustic Characteristics of Stop Consonants: A Controlled Study, MIT Doctoral Thesis.

Zweig, G., R. Lipes, and J. Pierce (1976), "The cochlear compromise," JASA **59**(4), 975-982.

Zwicker, E. (1961), "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," JASA 33(2), 248.

Zwicker, E. (1970), "Masking and Psychological Excitation as Consequences of the Ear's Frequency Analysis," Frequency Analysis and Periodicity Detection in Hearing, ed. R. Plomp & G. Smoorenburg, A.W. Sijthoff, Leiden.

Zwicker, E., E. Terhardt, and E. Paulus (1979), "Automatic speech recognition using psychoacoustic models," JASA 65(2), 487-498.

Zwislocki, J. (1975), "The Role of the External and Middle Ear in Sound Transmission," in The Nervous System, Vol 3: Human Communication and Its Disorders, D. Tower (ed.), Raven Press, New York.