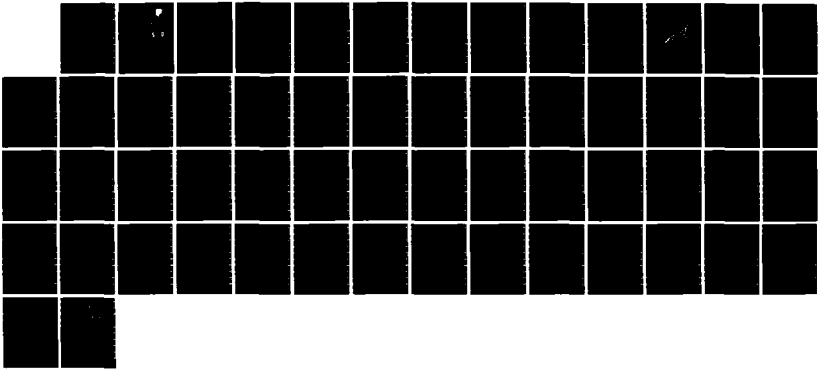
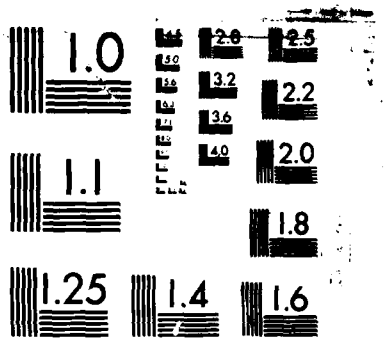
AD-A167 426 SPEECH ANALYSIS BASED ON A MODEL OF THE AUDITORY SYSTEM 1/1
(U) ROCHESTER INST OF TECH NY DEPT OF ELECTRICAL
ENGINEERING H E RHODY ET AL. FEB 86 RADC-TR-85-265
UNCLASSIFIED F3602-81-C-0169 F/G 20/1 NL





(12)

RADC-TR-85-265
Final Technical Report
February 1986

SPEECH ANALYSIS BASED ON A MODEL OF THE AUDITORY SYSTEM

Rochester Institute of Technology

Harvey E. Rhody, Robert A. Houde, Charles W. Parkins
and Sohiel Dianat

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441-5700

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-85-265 has been reviewed and is approved for publication.

APPROVED: James D. Mosko, Ph.D.

JAMES D. MOSKO, Ph.D.
Project Engineer

APPROVED: Walter J. Senus

WALTER J. SENUS
Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: Richard W. Pouliot

RICHARD W. POULIOT
Plans & Programs Division

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA) Griffiss AFB NY 13441-5700. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be returned.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A167426

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS N/A
2a SECURITY CLASSIFICATION AUTHORITY N/A	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE N/A	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) N/A	5 MONITORING ORGANIZATION REPORT NUMBER(S) RADC-TR-85-265
6a NAME OF PERFORMING ORGANIZATION Rochester Institute of Technology 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION Rome Air Development Center (IRAA)
6c ADDRESS (City, State, and ZIP Code) Department of Electrical Engineering Rochester NY 14623	7b ADDRESS (City, State, and ZIP Code) Griffiss AFB NY 13441-5700
8a NAME OF FUNDING / SPONSORING ORGANIZATION AFOSR 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F30602-81-C-0169
8c ADDRESS (City, State, and ZIP Code) Bolling AFB DC 20332	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 61102F; PROJECT NO: 2305; TASK NO: J8; WORK UNIT ACCESSION NO: P9

11 TITLE (Include Security Classification)
SPEECH ANALYSIS BASED ON A MODEL OF THE AUDITORY SYSTEM

12 PERSONAL AUTHOR(S)
Harvey E. Rhody, Robert A. Houde, Charles W. Parkins, Sohiel Dianat

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final	FROM Oct 83 TO Sep 85	February 1986	60

16 SUPPLEMENTARY NOTATION
N/A

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
06	02		Audition, Speech Perception
06	16		Auditory Modelling, Artificial Intelligence

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This report describes a signal processing technique for speech signals based on a model of the processing done in the auditory system. It has been found that the click response of the auditory system is an exponential function. The spectrogram produced by a speech waveform that has been windowed by the same function possesses excellent resolution of both time and frequency parameters of the signal. It is expected that this analysis method will improve the discrimination of phonemic elements, and will prove useful in such applications as speech recognition and speech compression.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED

22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c OFFICE SYMBOL
James D. Mosko, Ph. D.	(315) 330-4024	RADC (IRAA)

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

CONTENTS

INTRODUCTION	1
ANATOMY OF THE AUDITORY SYSTEM	3
SIGNAL GENERATION WITHIN THE AUDITORY SYSTEM	11
AUDITORY SYSTEM MODEL	24
PERFORMANCE OF THE MODEL	27
SPEECH SIGNAL ANALYSIS	30
CONCLUSIONS	45
FUTURE WORK	47
REFERENCES	50

INTRODUCTION

A common problem in any design of a speech recognition system is the large magnitude and high rate of raw data that is produced by any process which digitizes natural speech. As an example, simple pulse-code modulation (PCM) in which speech is sampled at a rate of 12,500 samples per second and quantized to 12 bits per sample generates 150,000 bits per second of data. PCM represents a very simple method for speech encoding, but it also produces data at a rate that would overload any kind of direct recognition process.

It is known that the quantity of data needed to represent speech can be reduced substantially by more sophisticated processes. Vocoder systems can reduce the data rate to less than 1000 bits per second. However, the quality of the representation produced by low-rate vocoders may not be high enough to permit automatic speech recognition.

A goal of this phase of the study, entitled Auditory Spectrum Analysis, is to investigate the methods that are used by animals to reduce the dimensionality of the speech waveform prior to passing it to the brain. It is reasonable to suppose that the representation of speech

that is present on the auditory nerve is fully adequate to preserve all of the information relevant to recognizing and understanding the utterance. A computer algorithm which mimics the natural processing would also preserve the relevant information for an artificial speech understanding system. As a side benefit, it may provide insight which could be exploited to produce a more efficient high-quality vocoder.

It is our expectation that the natural processing preserves the necessary information and also destroys much of the unnecessary information. This is a result of centuries of adaptation, in which the hearing mechanism has become attuned to the sounds in nature that are important for survival. If the natural process does provide such an efficient information filter, then it is sensible to try to emulate it in an artificial system.

The goal of this project is to provide a model for the natural auditory process. The investigators are Dr. Robert A. Houde, Dr. Harvey E. Rhody and Dr. Soheil Dianat of Rochester Institute of Technology, with collaboration and support by Dr. Charles Parkins of the University of Rochester School of Medicine and Dentistry.

The project has three specific tasks:

1. To gather a set of Post Stimulus Time histograms, which form the basis of the analysis of the auditory spectral processing system.

2. To analyze the PST data to develop a model for the processing which takes place in the auditory system.

3. To simulate the auditory spectrum analysis system on a digital computer and compare the results with experimental observations.

These tasks have been carried out successfully, and will be discussed separately in the presentation which follows.

ANATOMY OF THE AUDITORY SYSTEM

The auditory system is very complex. Even a review of the system is of booklength (Keidel, et. al, 1983). Thus, this section will present only a schematic description of the anatomy.

Sound that is picked up by the outer ear is passed through a structure which causes a membrane over the cochlea to vibrate. The cochela is a spiral structure with a membrane running more or less along its axis.

The vibrations are passed to the basilar membrane, which then responds with a travelling wave motion.

The motion pattern of the basilar membrane is such that a sinusoidal excitation produces a travelling wave from the base of the cochlea toward its apex. With increasing distance from the apex the amplitude of the deflection increases slowly to a maximum after which it rapidly drops to zero. The location of the maximum changes with the frequency of the excitation, coming closer to the apex for lower frequencies and closer to the base for higher frequencies.

The auditory nerve is composed of approximately 30,000 fibers, with fiber diameters between 3 and 10 microns. The fibers of the auditory nerve end in the cochlear nucleus, with each fiber making contact with 75 to 100 cells. In turn, each cell of the cochlear nucleus has synaptic contact with many fibers of the auditory nerve. The result is that the total number of cells in the cochlear nucleus is only about three times larger than the number of nerve fibers.

The motion of the basilar membrane is translated into an electrical stimulus of the nerve by a complex activity within the cochlea. The mechanism for this

translation is not well understood, but it is known to be focused on the hair cells. It is known, for example, that damage to the hair cells by streptomycin or noise changes the cochlear microphonics, and they are missing completely in animals with a congenital absence of hair cells. It appears that motion of the basilar membrane causes a bending and torque on individual hair cells, which, in turn, produces an electrical stimulation of the nerve fibers that are connected to that cell.

The terminations of the auditory nerve fibers are distributed over the length of the basilar membrane. Thus, motion that takes place on one part of the membrane is translated into electrical activity on those nerve fibers that are terminated in the hair cells of that region. A schematic diagram of the cochlea and nerve system is shown in Figure 1.

The action potentials on a particular nerve fiber resemble a stochastic process in which pulses of a constant amplitude are generated. The pulse rate is controlled by a complicated biological process in which the bending and twisting of the hair cells causes the secretion of neural chemicals. The rate of secretion is determined by both the direction and magnitude of the bend or twist, but not by the rate of the bend or twist.

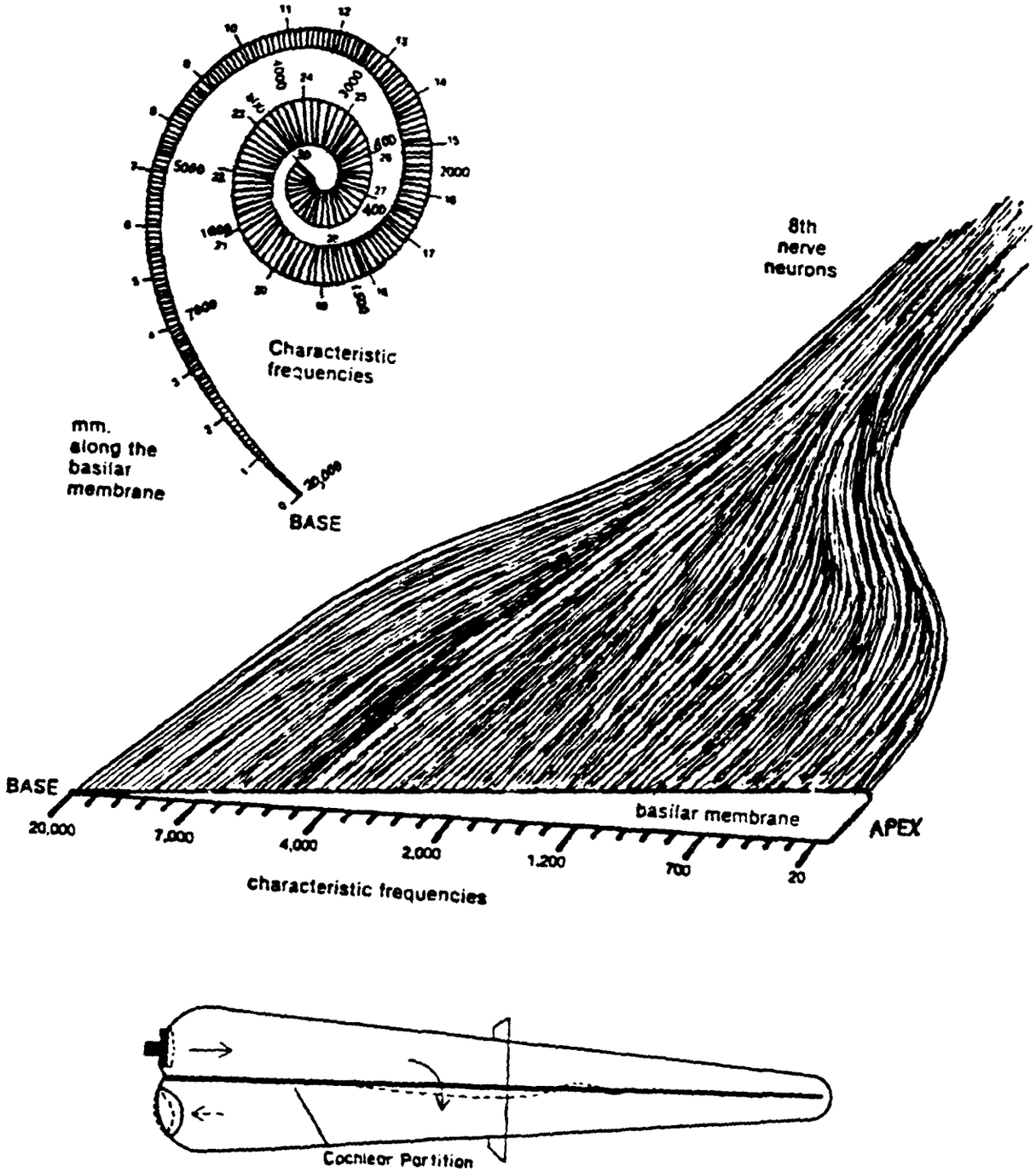


Figure 1. A schematic diagram of the cochlea and auditory nerve system.

When the neural chemical reaches a required concentration, an action potential is generated on the nerve fiber.

A schematic representation of the relationship between movement of a hair cell and the discharge of action potentials is shown in Figure 2. It is assumed that motion of the hair cell in one direction causes an increase in the pulse rate and that motion in the opposite direction causes a decrease in the rate.

The human ear can respond to sound waves in the frequency range of approximately 20 Hz to 20,000 Hz. The intensity of sound waves that lead to auditory perception ranges from approximately 10^{-16} W/cm^2 to 10^{-4} W/cm^2. The sensitivity of the ear to sound is somewhat dependent on frequency, with the greatest sensitivity normally falling in the neighborhood of 4000 Hz.

The ear has excellent ability to discriminate pitch, with frequency discrimination on the order of .1% being common. This observation makes it appealing to model the basilar membrane as a structure which resonates sharply at a place which is dependent on the excitation frequency. Under this theory, the place of excitation would change with frequency, so that

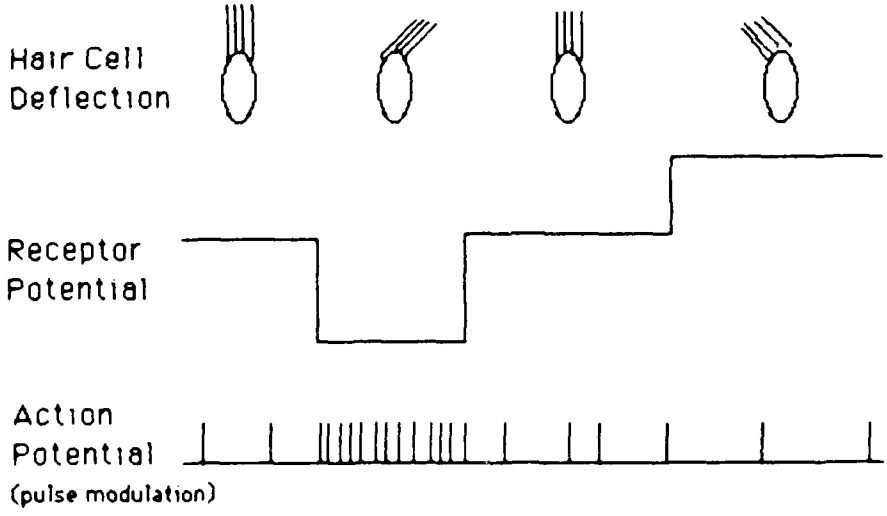


Figure 2. A schematic representation of the relationship between the movement of a hair cell and the discharge of action potentials.

different auditory nerve fibers would be excited by sounds of different frequencies. The number of nerve fibers that are excited by any one tone would be small so that it would be possible for small changes in tone frequency to be detected.

Natural sounds, such as speech, are composed of a large number of sinusoids. However, the individual components could be resolved by the resonant structure; in this way, different complex sounds could be differentiated by the listener.

In addition to the ability to do very fine frequency analysis, the auditory system is also capable of fine temporal resolution. This is evident from an examination of music, in which percussive sounds separated by less than .1 second can easily be resolved.

The resonator model cannot explain both the excellent frequency resolution and the excellent time resolution. To have the required time resolution, it would be necessary for the resonances to be highly damped. This, in turn, would prohibit a "high Q" frequency response. Perception of the quaver effect in music would require damping such that the half-width of the resonance would amount to about half a tone.

Although there is a conflict between the behavior that must be assumed for good time and good frequency resolution, there is no doubt that a relation exists between frequency and site of excitation within the inner ear. In fact, it is now believed that the basilar membrane supports a complicated travelling wave rather than simply resonating. The site of maximum displacement changes with frequency, but the width of the maximum is large compared with the frequency value.

It appears that the natural response of the basilar membrane cannot explain either the excellent frequency discrimination or the excellent time discrimination that is possible with the natural auditory system. Some frequency information is represented by the location of the responding nerve fibers. However, the response bandwidth for each nerve fiber is so large that it is not reasonable to expect that fine resolution should be possible from the place information.

In order to find the mechanism for the preservation of both time and frequency information, we must look in some detail at the structure of the signals that are passed from the auditory system to the brain. The total collection of these signals must contain the required

detail, although it may be somewhat hidden in the signal structure.

SIGNAL GENERATION WITHIN THE AUDITORY SYSTEM

To study the signal processing that takes place within the inner ear, recordings of action potentials from single fibers by means of micro-electrodes are used. The evaluation is done by statistical methods, in which the same sound stimulus is presented many times and the responses are combined statistically. The responses are presented in the form of histograms.

The response of a single nerve fiber to sixteen presentations of a sinusoidal tone burst is shown in Figure 3. Any one response is a stochastic pulse sequence, similar to that shown in Figure 2. However, a time histogram constructed by summing across many responses (200 in this case) shows a definite signal structure. This statistically derived sequence is known as the Post Stimulus Time (PST) histogram.

The PST histogram exhibits a number of features that make it an attractive candidate as a "signal". The pulses repeat at a frequency that is determined by the

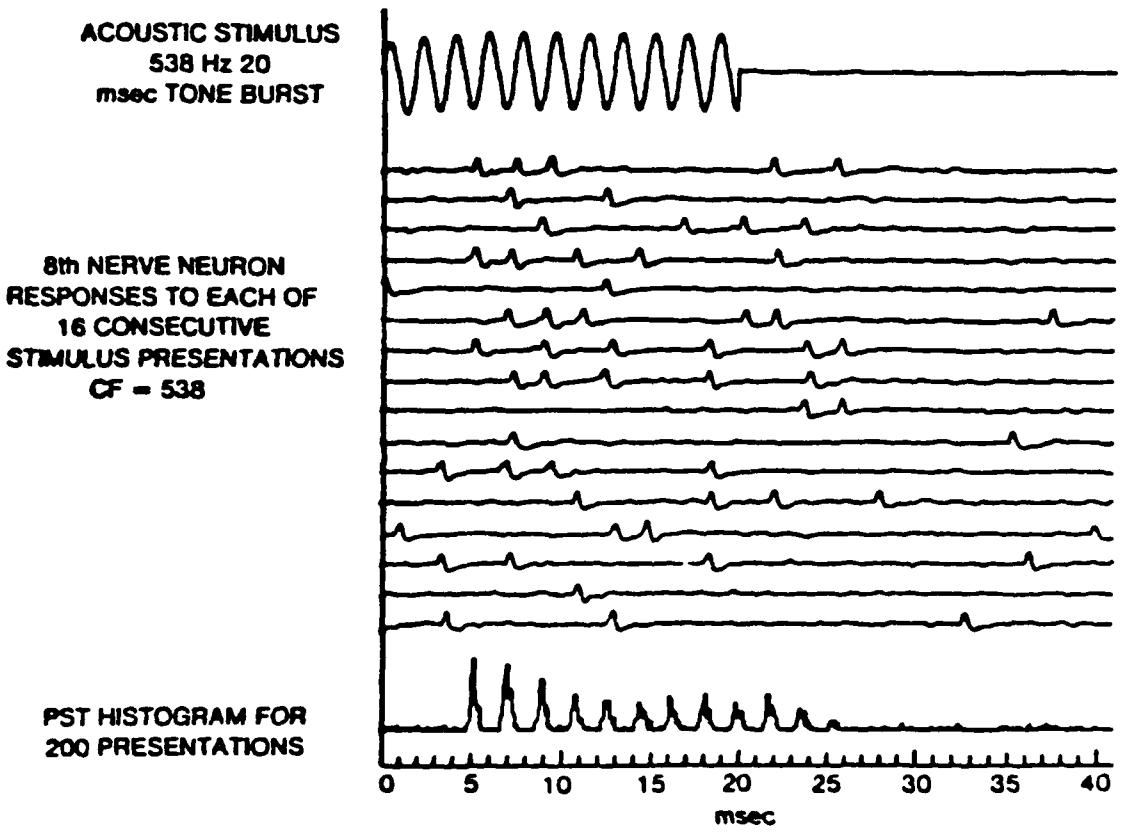


Figure 3. The response of a single nerve fiber to sixteen presentations of a sinusoidal tone burst.

excitation, and thus contain frequency information. The "waveform" has an exponential decay, which will be shown to be important to pitch determination.

Each nerve fiber in the auditory system responds to a range of frequencies. This is exhibited by the sequence of excitation and response curves shown in Figure 4. The experiment was carried out by electrically probing one auditory nerve fiber that responded at the characteristic frequency of 420 Hz at a threshold of 52 dB SPL. Tone burst excitations of 20 ms duration and frequencies of 141 to 800 Hz at a sound pressure level of 77.5 dB were used. The response curve in each case is the PST histogram.

The 141 Hz excitation is observed to cause a spike response due to the beginning and end of the burst. This is the "click" response that is generated on all nerve fibers by any large transient.

The 200 Hz excitation is seen to generate a PST histogram in which response pulses are spaced by the sinusoidal period of 5 ms. The initial spike is about 4 times as large as the steady-state level of spikes 2 through 4. A final large spike is generated by the end of the burst.

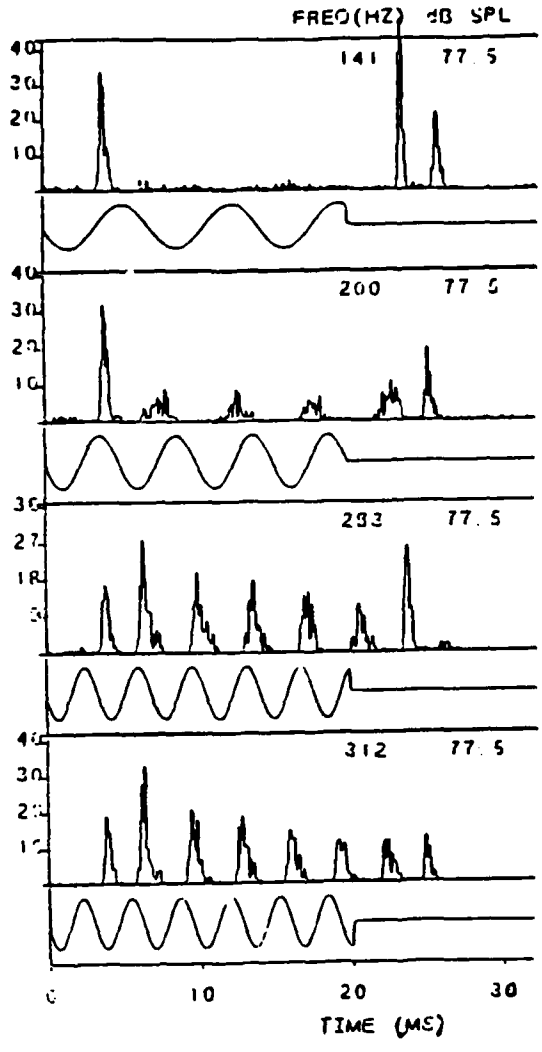


Figure 4. Response of a single auditory nerve that has a characteristic frequency of 420 Hz and a threshold of 52 dB to 20 ms tone burst excitations at a sound pressure level of 77.5 dB. The responses to a sequence of stimuli in which center frequencies of 141 through 800 Hz are shown.

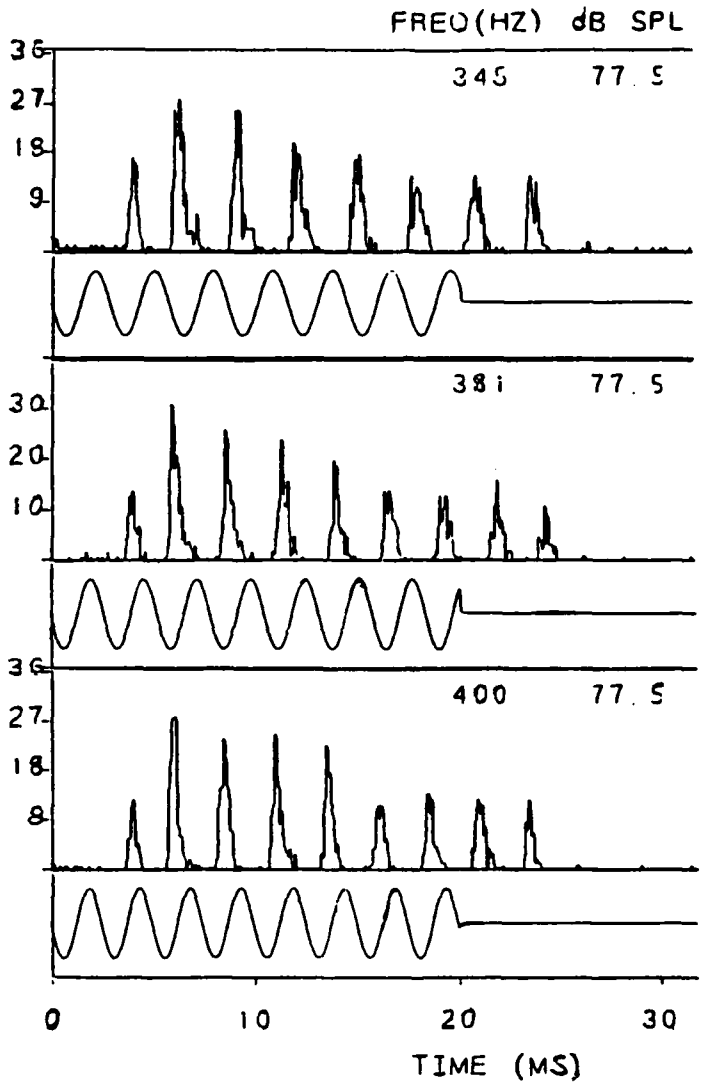


Figure 4. (Continued)

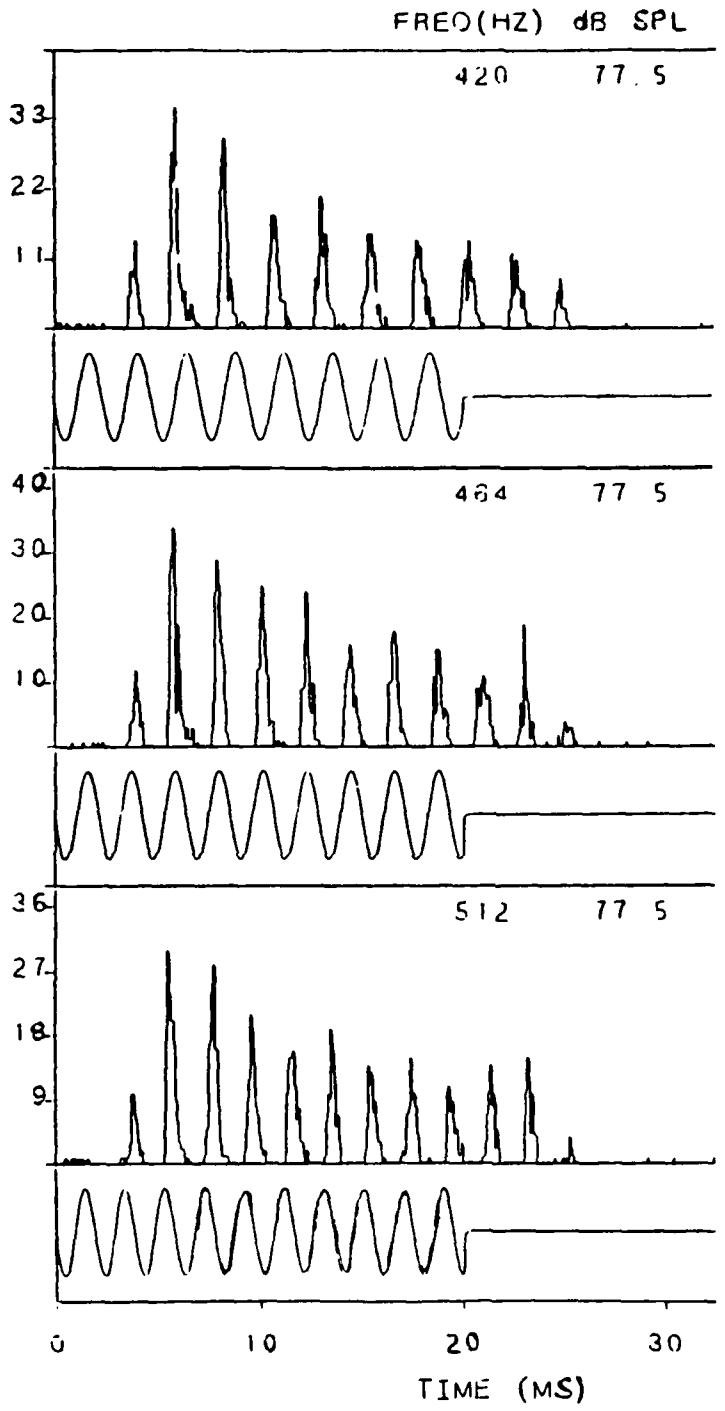


Figure 4. (Continued)

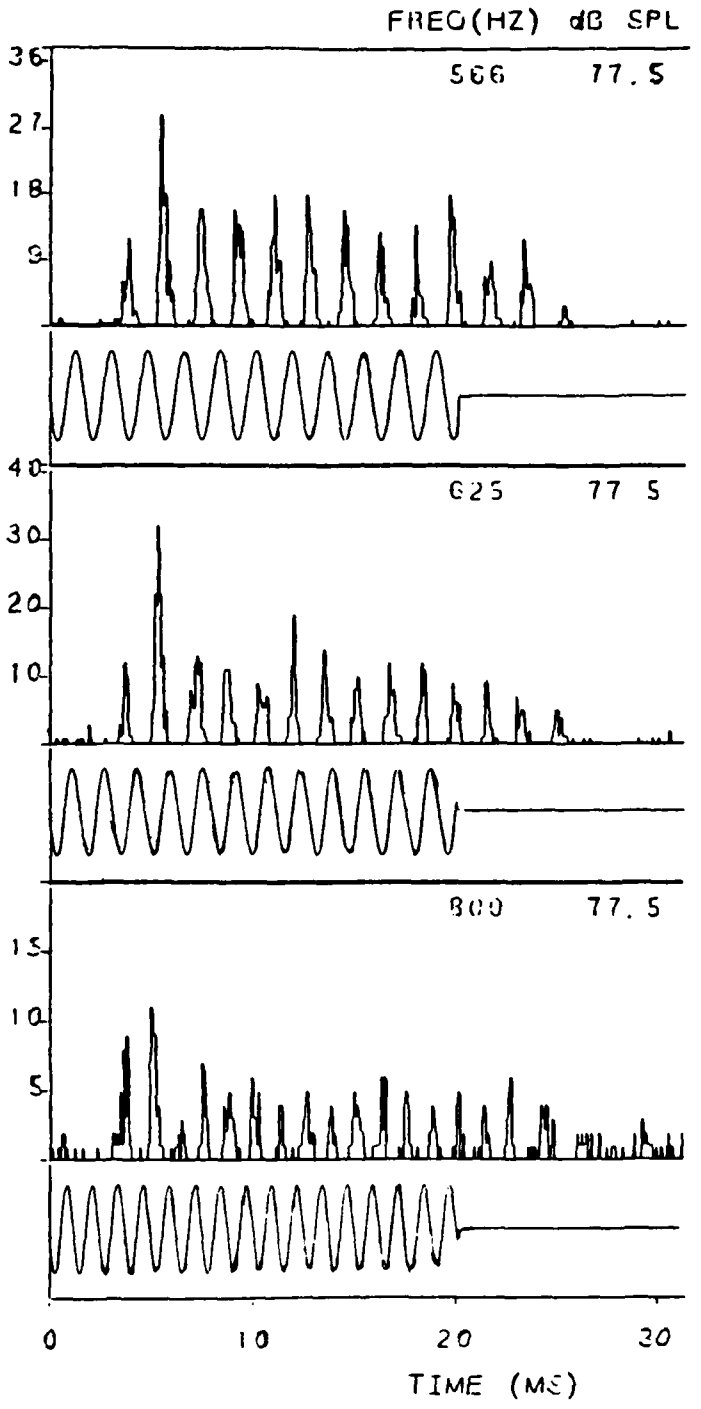


Figure 4. (Continued)

The responses to the 283 through 800 Hz tone bursts follow a general pattern in which a pulse sequence at the frequency of the excitation is generated. The delay of the pulse sequence is the same, about 4 ms, in all cases, which produces a phase offset which is a linear function of the excitation frequency. The leading pulse is somewhat smaller than the second, after which there is an exponential decrease to a steady-state level. The steady-state level is approximately the same for all sequences. The pulse sequences persist for a time equal to the initial delay, about 4 ms, after the excitation ends.

The experiment was repeated at a sound pressure level of 92.5 dB and frequencies ranging from 283 to 1131 Hz. The results in this case, shown in Figure 5, follow the same pattern, except that the largest pulse in each sequence is typically the first. The response to the 1131 Hz burst exhibits a pair of pulses that represent the click response, followed by a random output that is little different from the background level.

The frequency range that can be associated with a single nerve fiber can be determined by repeating the experiment using sinusoids of different frequencies. This experimentation reveals that the frequency range

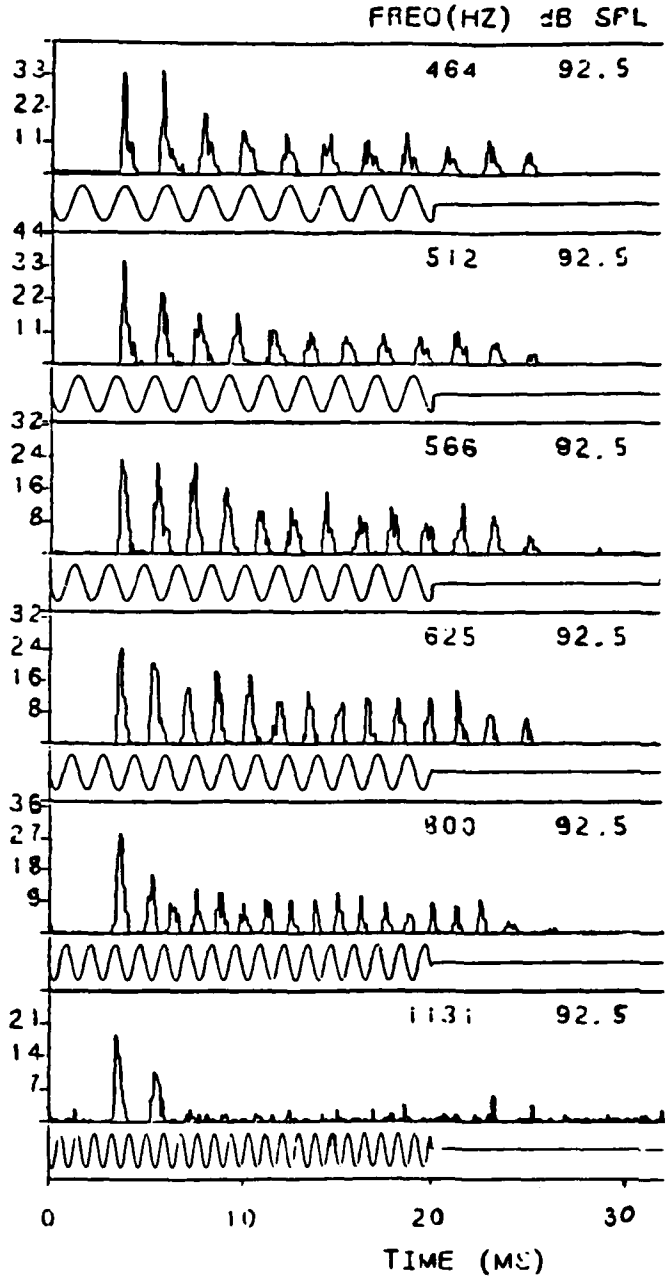


Figure 5. Response of a single auditory nerve that ... a characteristic frequency of 420 Hz ... threshold of 52 dB to 20 ms tone ... excitations at a sound pressure level of ... dB. The responses to a sequence of st... in which center frequencies of 283 thr... 1131 Hz are shown. (Data from Parkins).

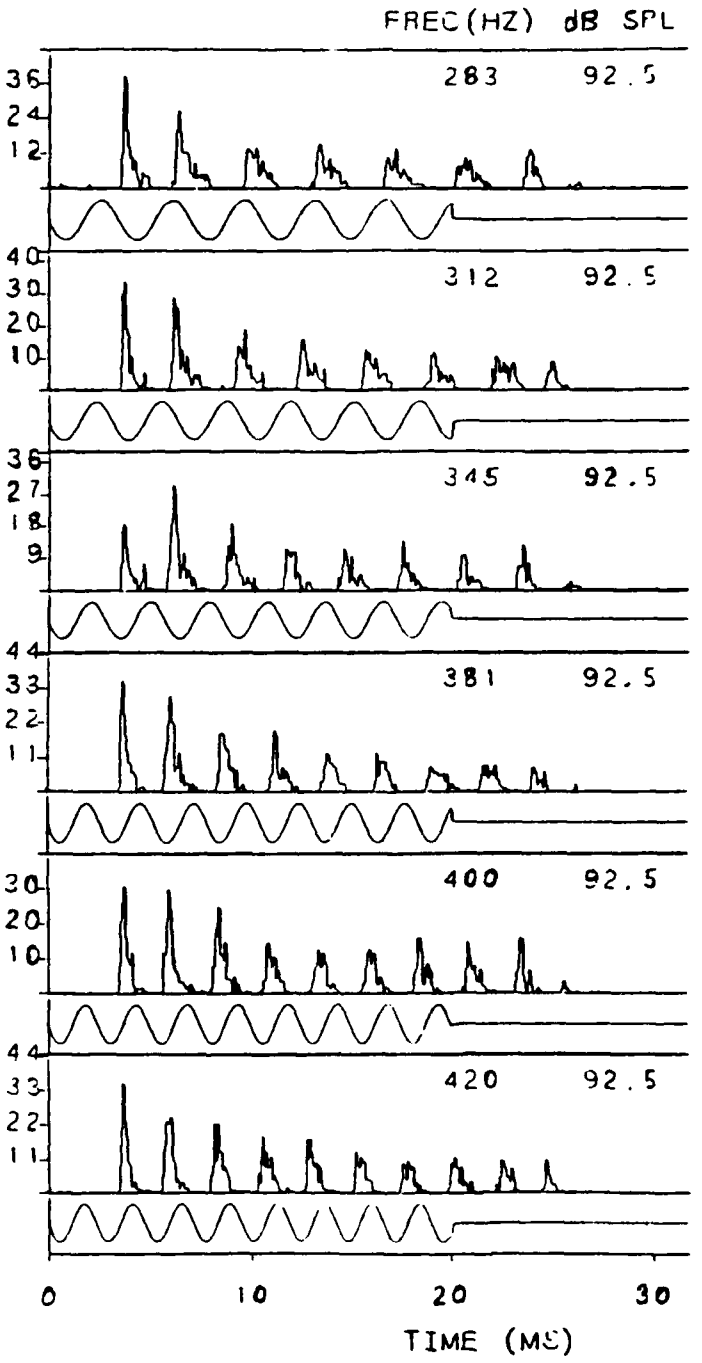


Figure 5. (Continued)

associated with any single nerve fiber is rather broad, on the order of several hundred hertz. Because of the structure of the system, the responses of adjacent nerve fibers would overlap considerably.

The response bandwidth observed on this particular nerve fiber ranges from about 200 Hz to about 800 Hz. This is a much larger bandwidth than would be expected if the frequency resolution was to be produced by a sharply resonant system. The frequency resolution needed to separate pairs of tones is present in the PST process itself, and not in the response of the individual nerve.

It may seem that there is a difficulty in using the PST histogram as a signal that is presented to the higher level processing system. The PST histogram is derived by repeated presentations of the same excitation. This is not a realistic model of nature, in which the ear must respond to a single presentation of a sound. However, something like the PST histogram may actually be present in the human auditory system. It is to be recalled that every cell of the cochlear nucleus makes contact with many nerve fibers. Each of the fibers would be excited in the same way, so that each would carry a stochastic pulse sequence. The stochastic

pulse sequences would be presented simultaneously to the brain, which could form a histogram by summing over the ensemble of fibers. Thus, the PST "signal" could be derived at the brain.

The development which follows is based on the assumption that the information needed to decode the speech signal is contained in the PST histogram. For all intents and purposes, the PST histogram is treated as a time waveform.

The frequency content of the excitation is contained in the basic pulse rate of the PST histogram. A mechanism by which a signal processor can extract the detailed frequency information is based on short-time spectral analysis. This will be presented later in this report. For now, it is sufficient to observe that the detailed frequency information is indeed preserved.

Time information is preserved in the responses of all of the nerve fibers. The PST histogram is a time function, and therefore contains time information directly. Those nerve fibers which correspond to the excitation frequency respond with a pulse sequence that is delayed in time by a characteristic offset and in which the pulse rate corresponds to the stimulation

frequency. Those fibers which are not in the excitation band respond with their characteristic click response. The click response for a given nerve has a characteristic pulse train in which the initial pulse is delayed from the time of stimulation by a time that is the same for all click stimuli and in which the pulse rate is at a characteristic frequency for that nerve fiber. The characteristic frequency is typically the frequency at which that nerve is most sensitive to stimulation. The pulse train decays with a characteristic exponential shape, with a time constant on the order of 4 ms.

All nerve fibers respond to a rapid change in signal level, be it a step or a pulse. Thus, the system naturally posesses the information for time resolution; it would appear that the time resolution should be at least 1 ms and perhaps less.

This excellent temporal resolution is related to the properties of the stochastic signals that are generated by the hair cells, rather than by the properties of the basilar membrane.

AUDITORY SYSTEM MODEL

The PST histogram, viewed as a signal, could be generated by a system with the block diagram shown in Figure 6. The incoming sound is first presented to a parallel set of bandpass filters. These filters have a rather broad response, corresponding to the frequency range to which an individual auditory nerve fiber would respond. The filter passbands are distributed in frequency with overlapping responses. At frequencies below about 3 kHz the filters all have the same bandwidth, and above 3 kHz the bandwidths increase exponentially. This distribution of bandwidths corresponds to experimental observation.

The second block in each channel is a half-wave rectifier. This corresponds to the unipolar behavior of the PST histogram, in which each peak corresponds to a peak in the stimulus. A low-pass filter with a cutoff frequency of about 1 kHz is included in the rectifier block. The function of the filter is to remove the high-frequency detail (sharp edges) which would be produced by ideal rectification.

The third block in each channel is a time-window function. The time window provides an exponential

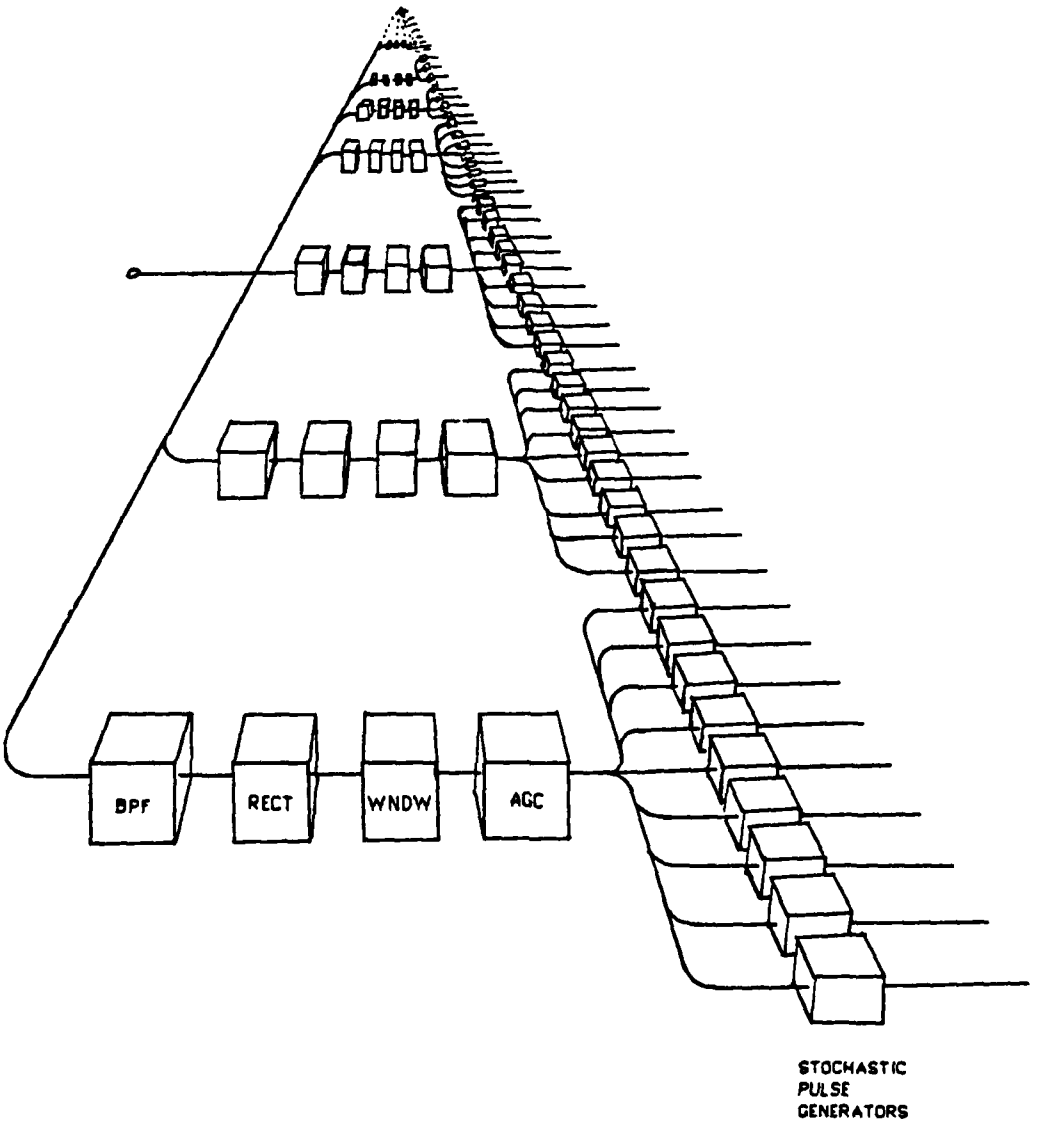


Figure 6. Block diagram of the signal processing model for the auditory system. (after Parkins, 1983).

weighting of the rectified signal, with the most recent portion given the greatest emphasis. An appropriate time constant for the exponential window is approximately 4 ms.

The final block in each processing channel is an automatic gain control. The function of this block is to provide for the adaptation that is observed in the auditory system to the general level of the incoming signal. Parkins, et. al. (1983) have shown that the dynamic range of the AGC should be on the order of 30 dB.

The output of each AGC is presented to a set of stochastic pulse generators. Each pulse generator produces a sequence of pulses of uniform amplitude but with firing time determined by the instantaneous level of the AGC output; the probability of a pulse generator firing at a given instant increases as the amplitude of the AGC output increases. Once a pulse generator has fired, it must wait for at a minimum dead time before it can again fire. The minimum dead time depends upon the characteristic frequency of the transmitter, but prohibits firing more than once per cycle.

The role of the parallel stochastic channels is to

transmit the information contained in the AGC output to the brain. The parallel channel bundle constitutes a transmission system within which the signals are pulse trains. This transmission channel can do no more than preserve the information which is in the AGC output waveform. This transmission system is not necessary in a system which is to duplicate the signal processing operations of the ear.

PERFORMANCE OF THE MODEL

The output of each of the processing stages of the model for a 538 Hz tone burst of 20 ms duration is shown in Figure 7. The simulated input is shown on the first line, with the onset of the burst corresponding to the left margin of the figure.

The output of a bandpass filter that was shaped to the neuron's tuning curve is shown on the second line. The filter phase characteristic may be adjusted to account for the latency of the neuron's response and the propagation delay along the basilar membrane. Here the delay has been adjusted to 4 ms. The bandlimiting effect of the filter produces a somewhat gradual rise

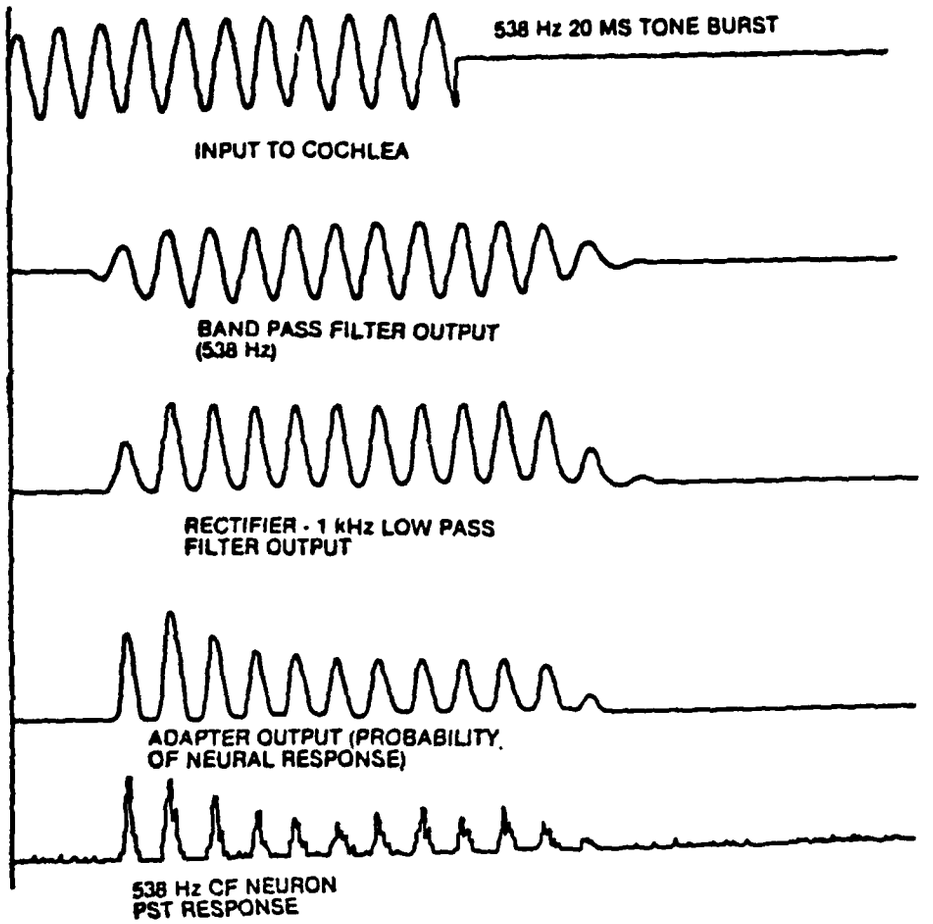


Figure 7. A comparison of signals generated by the auditory system model and an observed PST histogram (after Parkins).

and fall time on the tone burst. The filter bandwidth is on the order of 400 Hz, giving rise to a response time on the order of 2.5 ms.

The rectified and lowpass filtered output is shown on the third line. The positive polarity cycles produce output pulses, with the negative pulses being clipped at the zero baseline. The sharp corners that would result from clipping have been removed by the 1 kHz lowpass filter which is combined with the half-wave rectifier.

The result of the window and gain control operation is shown on the fourth line. Note that the short leading pulse has been enhanced by nearly a factor of two relative to the others, and that the third pulse is slightly reduced in size. Since the first and third pulses are spaced by about 4 ms, this would correspond to an exponential weighting with about a 4 ms time constant if the ratio of the first and third pulse changes was about e, which is approximately correct.

The fifth line of the figure is an experimental PST histogram obtained by aural stimulus with a 20 ms duration tone burst at 538 Hz. The strong correspondence between the experimental curve and the output of the auditory model is evident. Both the model and the

experiment produce pulse sequences that are at the same frequency as the excitation. Both sequences exhibit a large initial value, which decays exponentially to a low steady-state value and vanishes after a delay equal to the latency time after the excitation ends.

SPEECH SIGNAL ANALYSIS

Speech signal analysis may be carried out in many ways. A goal is to extract a set of parameters from the signal in such a way that the higher level processes in a speech recognition system can recognize the utterance. There is no requirement that the processing follow any "natural" process; however, it is likely to be the case that a substantial amount of guidance can be derived by observing the natural operations.

An important set of speech parameters can be derived from the speech spectrogram. It has long been recognized that the speech spectrogram provides one of the best parameter sets for recognition of speech sounds. Although other parameter families are possible, those which can be derived from the speech spectrogram have been chosen for this study. This does not mean that other sets should be ignored; only, that a reason-

able set had to be chosen as a starting point. It is not yet known whether a sufficient parameter set can be obtained from the speech spectrogram. However, the speech spectrogram seems to be the most successful basis for human speech reading. Later developments may lead to the need to at least augment the spectrogram analysis.

A modern version of the speech spectrogram can be produced by repeated short-time spectral analysis of the speech sound. Many versions of this process have been implemented.

The innovation that is presented in this report is the use of an exponential window prior to the spectral analysis. This approach mimics the processing that is done in the auditory system model. Windowing is intended to produce the same weighting that is represented by the click response of the auditory system. The click response is equivalent to the system impulse response, and represents the filtering that is done by the auditory system. The spectral analysis is then used to observe the frequency content from the windowed time function.

The effect of the exponential window is to provide a spectrogram in which both time and frequency information are preserved and made evident.

A block diagram of the signal processing system is shown in Figure 8. The speech is sampled at a rate of 12,500 samples per second using a 12 bit A/D converter. The samples are then multiplied by an exponential window. Approximately 10 ms (128 samples) of the windowed output are then processed using a 128 point FFT. The magnitudes of the frequency samples are computed and saved.

The spectrum magnitude samples may be processed to reduce the signal dimensionality by grouping magnitude components into frequency bins. The number and size of the frequency bins is a variable that must be manipulated to optimize system performance. A set of equal width bins up to 2 or 3 kHz and then of exponentially increasing width up to about 6 kHz will probably be found to be useful. The number of bins required will probably be in the range of 16 to 32.

The masking effect of the ear may be simulated by reducing the spectral amplitude of bins adjacent to a bin with a very large spectral value.

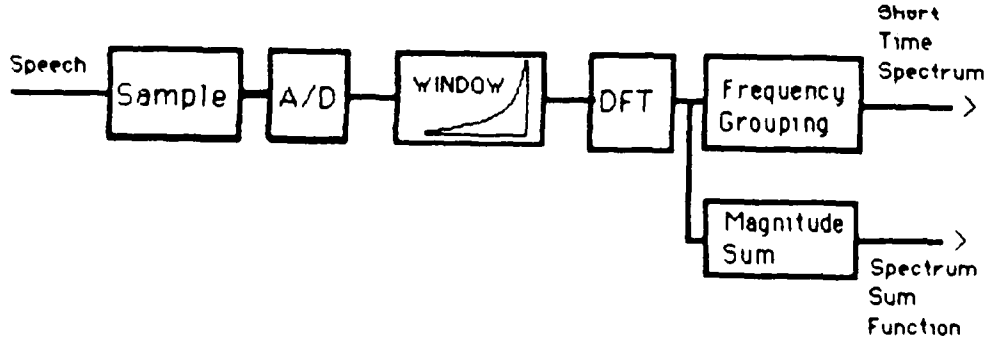


Figure 8. The block diagram of a signal processing system based on the auditory system.

An option that could reduce the number of bits needed to represent a spectral frame would be to log compress the spectral amplitudes. This mimics the saturation effect of the ear, and may not substantially reduce the usefulness of the spectra for the purpose of speech recognition.

At this point, the 10 ms section of speech has been fully processed, and may be stored or passed on to a higher level process.

The spectral processing is repeated by stepping the the window by 13 samples (about 1 ms). The time samples in the new window overlap those in the old one by about 9 ms. However, this fine-grained stepping is important for the preservation of the time information through the spectral processing.

The effect of processing a segment of speech by exponentially windowing and then transforming the data is shown in Figure 9. The utterance was a portion of the phrase "fleecy clouds in an azure sky." The bottom scale is a file index corresponding to the spectrum number, where the spectra are spaced by approximately 1 ms. The vertical axis corresponds to frequency on a linear scale, with the bottom being 0 Hz and the top

about 6 kHz.

The line above the spectrogram is a phonetic transcription of the utterance. This was put into the system by an operator reading the display and locating phonetic symbols at appropriate time points. (The alphabet has been modified somewhat to match the capabilities of the workstation.)

The top trace in the figure is the sum of the spectral values for each magnitude spectrum. That is, for each spectrogram that is computed, the sum of the magnitudes of the frequency components is computed. This sum will fluctuate as the instantaneous signal level fluctuates. It will be shown presently that the spectral sum fluctuates at the pitch frequency. The top trace exhibits a high frequency fluctuation, which is at the pitch frequency, superimposed on a much lower rate fluctuation. The low rate fluctuation is more or less at the phoneme rate. Unvoiced sections show exhibit a more noiselike variation. The spectral sum drops to a very low value during silent periods. Thus, it appears that the spectral sum will be useful in deriving prosodic information as well as a pitch measure. These measures would be useful in any kind of parameterization

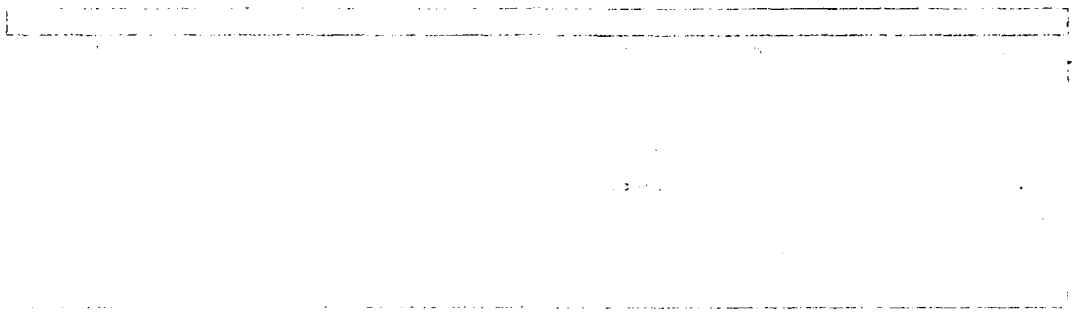


Figure 9. Short-time spectrum and spectral sum functions generated by the signal processing system of Figure 8 for a portion of the phrase "fleecy clouds in an azure sky".

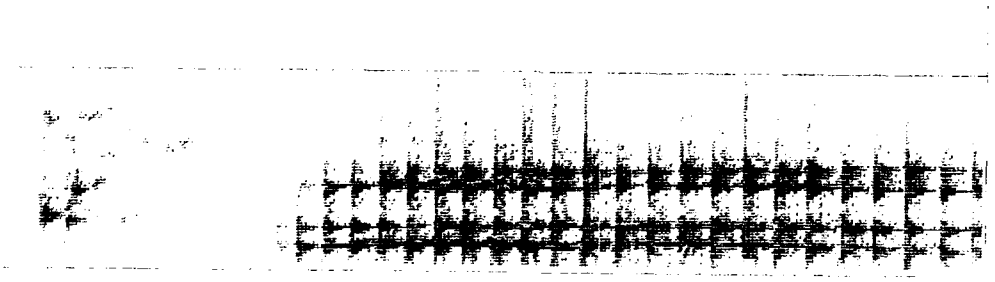


Figure 10. An expanded version of the same functions shown in Figure 9.

of the speech, whether for recognition or for vocoder applications.

The speech spectrogram of Figure 9 is very similar to spectrograms that are routinely generated in speech analysis. The heavy areas are the formants, showing the expected variations across the utterance. It is not difficult to pick out stops, fricatives, voiced and unvoiced sounds. The major difference between this spectrogram and others that are commonly used is the graininess. Of course, that could be reduced by appropriate smoothing.

A section of the same spectrogram is expanded and shown in Figure 10. The portion that is shown is the first part of the word "clouds". Note that the time index covers the region between 380 and 610 on the spectrogram of Figure 9.

A notable feature of Figure 10 is the vertical banding at approximately every 8 ms. The heavy vertical bands occur once each pitch period, and coincide with the maximum points of the spectral sum function shown above the spectrum. It appears that there is a strong representation of the pitch in the spectrogram, and that, with proper processing, it should be possible to

extract a satisfactory representation.

The spectral sum function can be further enhanced by signal processing to provide a signal that is convenient for pitch extraction. This technique has been used by Seneff (1985) with excellent results, and can be expected to have similar performance in this system.

A further notable feature of the voiced part of the spectrogram is the periodically changing widths of the formants. Each formant is widest at the time the spectral sum is the greatest. In the example shown in Figure10, the formants merge at the points of maximum width. The formants become narrower at what appears to be an exponential rate until, just before the next pitch pulse, they reach their minimum widths. The minimum width appears to be only about 20% of the maximum width. The fact that the formants are most obvious between pitch pulses is consistent with the results reported by Kates (1983).

The points of minimum formant width provide maximum definition of the formant frequencies. This resolution is substantially better than that which is ordinarily obtained by spectrogram analysis. The reason for the

improvement is to be found in the windowing that is applied before the short-time spectra are computed. To see how this comes about, it is necessary to examine the windowed functions in the time domain.

A 20 ms portion of the time waveform that begins with STS index 464 is shown in Figure 11. This waveform corresponds to the same utterance that was used to create Figures 9 and 10. It represents approximately three pitch periods of the voiced portion of the word "clouds". The large time devisions fall every 2 ms, with the finest divisions representing 0.2 ms.

The time function that would be produced by multiplying the speech waveform by an exponential window depends upon the time constant and location of the window. For this speech sample, the signal level decays exponentially with a time constant that is in the range of 4 ms, which matches the time constant that was used for the analysis window. Is it a surprise that the time constant that is found in the auditory system is matched to the speech signal? The auditory experimentation was done on cats, while the speech sample was obviously from a person. However, cats and people communicate in the same frequency band.

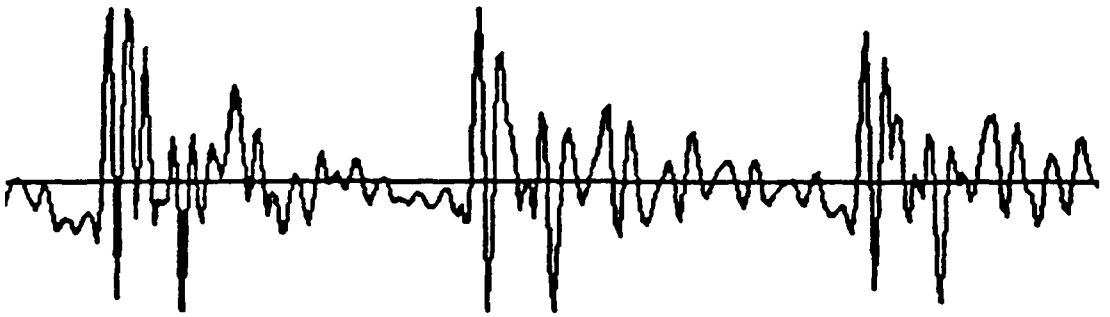


Figure 11. A 20 ms portion of the voice waveform used to generate the functions shown in Figures 9 and 10.

The time function that would be generated for different window positions is sketched in Figure 12. The function in Figure 12a resembles a single pulse because the window position is such that the leading edge of the window just overlays the first cycle of the speech waveform. The tail of the window matches the tail of the previous speech waveform section, and therefore the product in that portion is very small. As

Figure 12a.

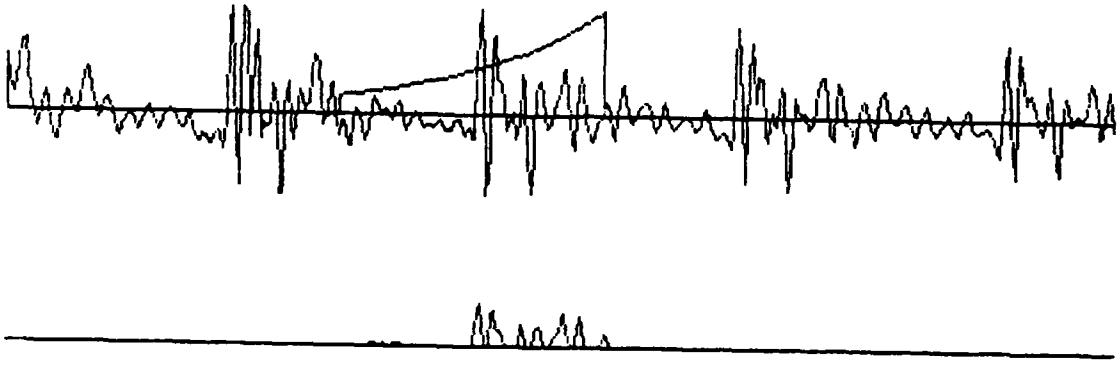


Figure 12b.

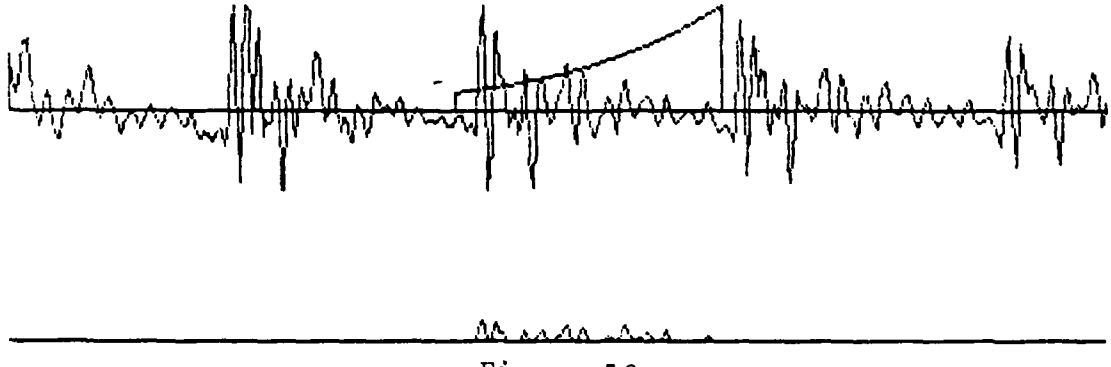


Figure 12c.

Figure 12. The effect of the exponential window on the voice waveform and short-time spectrum. Note the change in frequency domain resolution with window position.

the window is stepped to the right in 1 ms intervals it includes more of the large excursions at the beginning of the speech waveform. The product waveform therefore exhibits more cycles, although they decrease in size with each step of the window. In seven 1 ms steps, the window location will have moved through a full pitch period and will select the first pulse of the next pitch period. The is the waveform shown in Figure 12c.

The sequence of short-time spectra that correspond to the sequence of windowed time functions is shown in Figure 13. The spectrum labeled with the index 464 corresponds to a window location such as that shown in Figure 12a in which the first pulse of the pitch period is emphasized. The spectrum is displayed in 32 frequency bins, ranging from 0 Hz to about 6 kHz. It is notable that the spectrum has two broad humps with a rather smooth transition between them.

The spectra labeled 465, 466,... in Figure 13 represent the amplitude spectrum that would correspond to time waveforms such as Figure 12b, 12c,... The window is stepped by 1 ms for each frame.

The spectra in the sequence are seen to gradually change in such a way that the individual maxima become

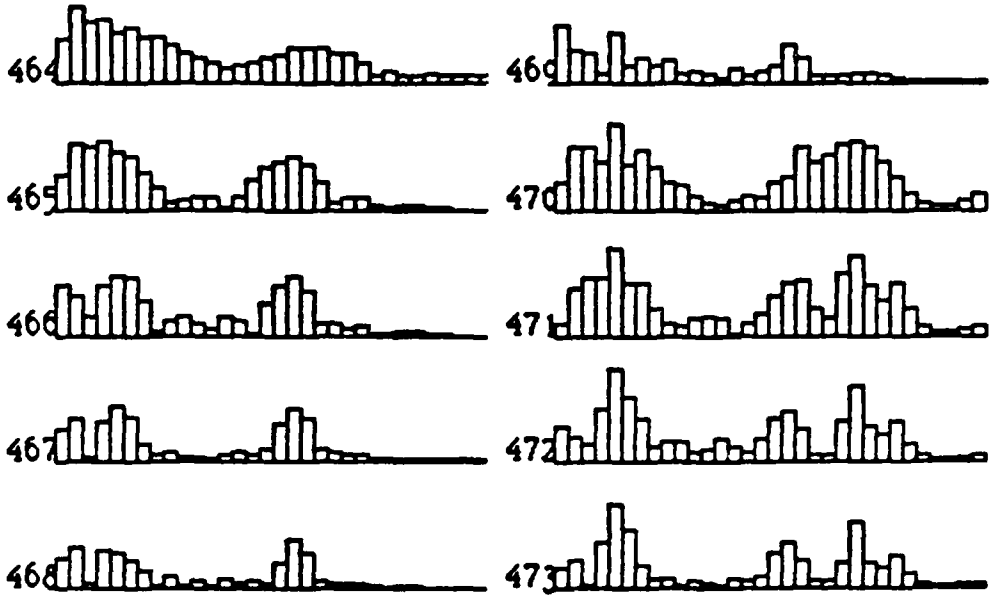


Figure 13. The sequence of short-time spectra that corresponds to the windowed time functions. For each analysis the window is advanced by 1 ms.

sharper, narrower, and more distinct. The spectral peaks are narrowest for indices 468 and 469, and then become rather broad again with spectrum 470. This spectrum corresponds to the time function at which the window has reached the beginning of the next pitch period.

The spectral peaks correspond to the formant frequencies of the utterance. The cyclic narrowing and widening of the formants is evident in this sequence. It can also be seen that the amplitudes of the spectra at the formant frequencies do not change much throughout the sequence. This observation, in particular, may be somewhat surprising since there is a radical change in the waveforms of the sequence.

The fact that the first spectrum is rather broad can be explained by noting that the corresponding time function is quite short. This short time function would contain a broad band of energy, much as would be expected of a pulse. Some evidence of the vocal resonances is present, as seen in the broad formants, but the resolution of the resonances is very weak.

Later time functions in the sequence contain more pulses. Because the exponential decay of the window

matches the time constant of the time waveform envelope, the pulses are of roughly the same amplitude. The later time sequences have longer signal sections in which there is a significant amount of energy, and therefore have the ability to provide more resolution in the frequency domain. This greater resolution is evident in the spectral sequence. Because there is greater definition in the frequency domain, relatively more energy must be contained in the formants. At the point of maximum resolution, most of the energy must be in the formants. However, the total energy will be about the same because the pulses in the signal become shorter as the number of pulses in the window increase. This balancing effect leads to spectral components at the formant frequencies which tend to have the same amplitude throughout the sequence.

CONCLUSIONS

A model for the auditory system was constructed on the basis of interpretation of experimental data. The model resembles the traditional filter bank in the sense that any given filter corresponds to location along the basilar membrane. The output of the processing prior to the stochastic pulse generators is a signal that is

similar to the PST histogram. The signal that is generated in that fashion can then control a set of parallel stochastic pulse generators for the purpose of transmission to the brain. Although the stochastic pulse generators are controlled by a deterministic signal, their output sequences would resemble the random signals on the auditory nerves.

This model provides insight into the mechanism by which speech information is encoded for transmission to the brain. This understanding can be exploited to develop a speech representation that is useful for applications such as the front end for speech understanding systems or the basic processing element of a vocoder analyzer.

The deterministic signal that is used in the model to control the stochastic pulse generators contains the speech information. The parallel set of stochastic signals represents an encoding of this information. Therefore, the deterministic signal can be used as a speech representation.

A method was developed to extract both detailed time information and detailed frequency information from the deterministic PST replica. This analytical proced-

ure involves exponential windowing, Fourier transformation and spectrum processing. The pitch can be easily observed in the voiced portions of the spectrogram, and can be extracted through the use of the spectral sum function. The formants can be located accurately by extracting the spectrum peak locations in the spectra that occur just before the pitch pulse.

It appears that this representation of speech will make it possible to make reliable voice/unvoice decisions. Moreover, it appears that it will provide critical time and frequency details that will be helpful in segmentation.

FUTURE WORK

The next phase of the speech recognition project will be the development of an approach to extract a parameter set that will permit phoneme identification and speech segmentation.

It is recognized that reliable phoneme identification, even by trained human spectrogram readers, depends upon the identification of words and phrases. It has been shown by Klatt[3] that use must be

made of the speech context if phonemes are to be identified in continuous speech. Therefore, no simple parameter extraction and pattern matching algorithm will be sufficient for phoneme recognition.

The goal of a parameter extraction algorithm should be to determine a set of possible phonemes, and, if possible, provide some ranking in terms of likelihood. This ranking cannot be in terms of probabilities, since the probabilities, to be meaningful, would have to be conditionally related to the context. The parameterization should be such that a set of phonemes containing the true phoneme is preserved. The selection of the true phoneme would then be made by higher level decoding. This is the essence of the fuzzy logic structure to decision making in such a context.

The parameterization study requires the development of speech analysis tools that make use of the auditory model representation of speech. A working relationship has been established between RIT and SR Systems, Inc., through which tools that have been developed by SR Systems would be made available to this study on an exclusive basis. These tools have been developed for the Sun Microsystems computer, and are being converted to operate on an Apple Macintosh computer.

REFERENCES

1. Kates, James M., "An Auditory Spectral Analysis Model Using the Chirp z-Transform", IEEE Transactions on Acoustics, Speech, and Signal Processing, v. 31, No. 1, February, 1983, pp. 148-156.

2. Keidel, Wolf D., S. Kallert and M. Korth, The Physiological Basis of Hearing, Thieme-Stratton, New York, 1983.

3. Klatt, Dennis H., "Word verification in a speech understanding system," in Speech Recognition, D. Raj Reddy, ed., Academic Press, New York, 1975.

4. Parkins, Charles W., Robert Houde and John Bancroft, "A Fiber Sum Modulation Code for a Cochlear Prosthesis", in COCHLEAR PROSTHESIS: An International Symposium, Annals of the New York Academy of Sciences, Vol. 405, New York, 1983, pp. 490-501.

5. Seneff, Stephanie, "Pitch and Spectral Analysis of Speech Based on an auditory Synchrony Model", Technical Report 504, Massachusetts Institute of Technology, Research Laboratory for Electronics, Cambridge, Massachusetts 02139, January, 1985.

This will be the focus of the effort for the next contract phase.

Concurrent with the development of the speech processing workstation, a literature search will be conducted to develop reasonable parametric sets for speech representation. This information will be used to structure the following research phase.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control, Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of competence is provided to ESD Program Offices (POs) and other ESD elements to perform effective acquisition of C^3I systems. The areas of technical competence include communications, command and control, battle management, information processing, surveillance sensors, intelligence data collection and handling, solid state sciences, electromagnetics, and propagation, and electronic, maintainability, and compatibility.

END

DTIC

6-86