A TIME DOMAIN STUDY OF SPEECH SOUNDS

by

John Crable Wauer

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major Subject: Electrical Engineering

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Head of Major Department

Signature was redacted for privacy.

Dean of Graduate College

Iowa State University
Of Science and Technology
Ames, Iowa

1963

TABLE OF CONTENTS

INTRODUCTION

Speech contains certain salient elements that allows man to extract sematic information. Speech also contains a great deal of irrelevant structure which may be disregarded if the information bearing elements are known. One approach to understanding the information bearing elements of speech as well as the irrelevant structure is to investigate the method of speech production. This approach is based on the assumption that if the coding scheme can be understood, a good decoding scheme will be simply the inverse. With this philosophy, the theory of speech production may be reviewed.

Theory of Speech Production

In order to produce a sound the lungs generate a smooth flow of air. The air is modulated either by the vibration of the vocal cords producing a pulsating air flow or by constrictions such as the lips or tongue producing a turbulent air flow. The sounds produced by vibration of the vocal cords such as the vowel sound in the word "heat" are called voiced as contrasted with the turbulent sounds such as /s/ in the word "see".

The pulsating air flow produced by the vocal cords has a large harmonic content. This signal is then filtered by the resonant cavities of the oral and nasal passages. These cavities are tuned by the tongue, teeth, and lips. The

resonant frequencies of the vocal tract are called the formant frequencies (47).

One possible set of properties that might be used to completely describe voiced speech sounds is the volume velocity waveform at the slit in the vocal cords or glottis and the transfer impedance[1] of the vocal tract. Since the slit in the vocal cords or the glottis is much smaller than the vocal cavities, the glottis is a high acoustic impedance source. It may be thought of as a volume velocity source.

Miller (41) has studied the wave form of the vocal cord volume velocity by building a network whose transfer function is the inverse of the vocal tract transfer impedance. It was found that the volume velocity wave form of the glottis was a series of periodic pulses approximately triangular in shape. This finding is in agreement with earlier observations of high speed motion pictures of the vocal cord action. Most people have a considerable period of closure during each pitch period. When a person changes the pitch of his voice he tends to keep the pulse shape and pulse width constant, only changing the period of closure. The main excitation of the formants occurs primarily in phase with the rapid closure of the vocal cords. The Fourier transform of the glottal wave may be approximated by

---

[1]Acoustical impedance is defined as the ratio of pressure to volume velocity.

$$F\left(\frac{2n\pi}{T}\right) = \left[\frac{\sin\left(\frac{\pi n t_1}{T}\right)}{\frac{\pi n t_1}{T}}\right]^2$$

where t_1 is the glottal pulse width and T is the pitch period.

Attempts have been made to measure the bandwidths of the formant frequencies. Several investigators (15) have attempted to fit the frequency spectrum of a voiced sound with the resonance curve of a simple tuned circuit. However, due to the uncertainty of the glottal source frequency spectrum and due to the coarse harmonic structure of the frequency spectrum, the results have had a large variance. Most of these bandwidth measurements fell in the range from 50 to 120 cps.

House and Stevens (26) set off a spark inside the mouth of a subject as he held his articulars in a given vowel position. From this impulsive response the formant bandwidth can be determined. They obtained bandwidths of 54, 65, and 70 cps for the first three formant frequencies respectively. This work has been criticized (15) because the bandwidths of the analyzing filters were not properly taken into account. This would imply that the bandwidths measured experimentally were low.

Mathews, Miller, and David (39) have computed the Fourier transform for one pitch period of several voiced sounds. A Laplace transform function fitting the frequency

data was computed. This function is the product of the Laplace transform of the glottal pulse and the transfer impedance function of the vocal tract. They have shown that the transform of the glottal pulse has an infinite number of zeros near the imaginary axis in the s-plane and no poles. For large frequencies these zeros will occur at frequency intervals equal to the reciprocal of the glottal pulse width. The glottal pulse width varies from 1/3 to 1/2 of the pitch period.

Weibel (59, 60) has shown that the transfer impedance function of the vocal tract has only one zero at s = 0. This allows the pole-zero plot of the speech sounds to be separated into the zeros due to the glottal pulses and the zero and poles due to the vocal tract.

The mean bandwidth computed by Mathews, Miller, and David using this method was 164, cps for zeros and 73, 204, 270, and 308 cps for the first four poles respectively. Dunn (15) has suggested that the large discrepancy between these results and the results of other methods was due to numerical problems in curve fitting to find the transfer impedance function.

Pinson (48) has approached the problem of describing voiced sounds slightly differently. He attempts to fit the speech waveform in the time domain with a set of exponentially damped sinusoids. The fitting is done only over that segment

of the pitch period when the glottis is closed. The mean bandwidth for the first two formants are found to be 60 and 56 cps respectively.

Manley (37) has attempted to represent the speech waveforms by orthogonalized exponentially damped sinusoids on a pitch synchronous basis. He found that a small number of terms would represent the signal quite well because of the similarity between the speech waveforms and the exponentially damped sinusoids.

The analyses of Mathews, Miller, and David, Pinson, and Manley were performed on the digital computer. Another method of analysis is to use a spectrograph (31). A spectrograph plots a short term frequency spectrum against time. Intensity of each frequency component is represented by the darkness of the spectrogram. The spectrum is found by averaging the outputs of the band pass filters. The response time of these filters is approximately equal to the reciprocal of the filter bandwidth. Narrow band spectrographs have bandwidths of 45 cps. This is sufficiently narrow to show the harmonic structure of the voiced sounds. Wide band spectrographs have bandwidths of 300 cps. These have a sufficiently fast response time to show a small amount of the spectral variations between pitch periods.

The actual spectrograph uses a fixed band-pass filter, and heterodynes the signal up to the frequency of the filter.

By making several passes of the data through the device, incrementing the frequency of the local oscillator each time, the short term frequency spectrum can be generated.

Speech can be reconstructed from the short term spectrum by exciting a bank of filters with some driving function that has a flat frequency spectrum. The gain of each filter is adjusted to give an output amplitude equal to that defined by the voice spectrum. The Bell Telephone Laboratories (12) have developed such a device, called the vocoder. The synthesis part of the vocoder uses two excitation sources. For voiced sounds the system is excited by a series of pulses periodic in the pitch frequency. These are then sent through an equalizing network to give a flat frequency spectrum. The pitch frequency is found by frequency-detecting the low-frequency component of the original speech signal. For unvoiced sounds, a noise generator is used to excite speech synthesizer. The voiced-unvoiced decision is made by looking at the energy above and below 900 cps. If the predominant energy is above 900 cps the sound is unvoiced.

The filters on both the analysis and the synthesis sections of the vocoder are usually spaced according to the Koenig aural scale (30). The Koenig scale approximates the frequency discriminating properties of the ear. For example if one tries to find a tone that appears to have twice the pitch of some reference tone, the frequency of the unknown

tone will probably not be twice the frequency of the reference tone. A subjective pitch scale can be made so that on this scale the unknown tone will have twice the pitch of the reference tone. The Koenig scale spaces the frequencies in such a way that it approximates the frequency spacing of the subjective pitch scale. The Koenig scale has a linear spacing of frequencies below 1000 cps and a logarithmic spacing above 1000 cps.

The conventional vocoder produces a somewhat distorted version of speech. Two of the reasons for this are that the pitch frequency detector of the conventional vocoder does not follow the pitch very well because of its long averaging time and the simple voicing-unvoicing measurement of the conventional vocoder is not always reliable. The voice excited vocoder (6) overcomes these difficulties to a large extent. Instead of making a voiced-unvoiced decision and measuring the pitch frequency, the voice excited vocoder transmits a base band signal from the original speech waveform. The base band signal consists of the lowest 700 cps of the speech waveform. In the synthesizer the base band is sent through a nonlinear element to regenerate the higher frequencies. After equalizing to flatten the signal's frequency spectrum, this signal is used to excite the vocoder for both voiced and unvoiced sounds.

The voice excited vocoder produces a more natural

sounding speech than the conventional vocoders. The conventional vocoders have been said to put "marbles in the talker's mouth", eliminate a talker's individuality so all talkers sound alike (6).

An obvious application of the vocoder is as a coding device to reduce the required channel capacity of a communication system. The signals specifying the operation of the speech synthesizer vary at a syllabic rate. This rate is less than 30 cps. The bandwidth requirements of a voice excited vocoder are somewhat higher than for a conventional vocoder. Bandwidth reductions of up to 4 or 5 have been realized for the voice excited vocoder as compared to 15 for the conventional vocoder.

An equally important application of the vocoder principle is to study the properties of speech. Haskins Laboratories (8, 33) has developed a device to convert spectrograms into sound. The device consists of 50 beams of light modulated at the first 50 harmonics of a 120 cps fundamental. Each light shines on the spectrogram at a spot corresponding to the modulating frequency of the light beam. Light is reflected from the spectrogram at the points corresponding to an intense signal and is detected by photo cells. Hand painted spectrograms can be converted to speech to test a theory of speech. They have shown that the important properties of voiced speech are the formant

frequencies. The formant bandwidths and formant amplitudes are relatively unimportant to the intelligibility of speech. These factors do affect the quality of speech.

Weibel (59) and many others have built analogs of the vocal tracts. These consist of filters tuned to the formant frequencies. The filter bandwidths are adjusted to the proper values. These investigators have found that the excitation waveform and the bandwidth of the filters were not critical in the operation of the analog.

Peterson and Barney (46) recorded ten simple vowel sounds placed in monosyllabic words beginning with /h/ and ending with /d/. The words were "heed", "hid", "head", "had", "hod", "hawed", "hood", "who'd", "hud", and "heard". Each word was spoken twice by 76 speakers giving a total of 1520 words. Seventy observers attempted to recognize the words when they were played back. There was some confusion because some speakers did not differentiate between "hod" and "hawed". These same people did not differentiate between the words when listening to the words. When the first formant frequency of each vowel is plotted against its second formant frequency, points representing the same vowels are clustered in the same region. However the point sets corresponding to each vowel are not disjoint even when words that had received poor scores in the listening tests are discarded. If data from only one speaker is used the points are quite well grouped

into disjoint sets.

The vowel sound in the word "heard" has a third formant frequency that is markedly lower than for any other vowel tested. This allows the vowel to be easily separated from the nine other vowels. This vowel also had a high articulation score on the listening tests. An articulation score is the number of times in per cent that the vowel was correctly recognized.

Foulkes (20) studied the pitch frequency and the first three formant frequencies of Peterson and Barley's speech data. He used the data from only nine vowel sounds since the word "heard" can be separated as mentioned above. He succeeded in mapping these four properties into a two dimensional space such that regions corresponding to the nine vowel sounds were disjoint. The region containing each vowel sound was also connected and had linear boundaries. This illustrates the sufficiency of the formant frequency description at least for these ten vowel sounds.

When the articulars of the vocal tract move from the position of a consonant to the position of a vowel or from a vowel to a consonant, the formant frequencies tend to follow the movements of the articulars producing formant transitions. The formant transitions are believed to offer clues for the identification of the consonant. Potter, Kopp, and Green (50) have proposed the "hub theory" to attempt to show a

pattern in these transitions. They have assumed that for a given consonant the second formant frequency will always start at the same place and go to the second formant position of the following vowel. The hub is the second formant position of a consonant or vowel when that sound is spoken by itself.

Delattre, Liberman, and Cooper (8) have proposed a similar theory called the "locus theory". The locus is the point on the frequency scale at which a transition begins or to which it may be assumed to point. The locus theory is based on the study of synthetic speech generated from hand painted spectrograms while the hub theory is based on the study of actual speech spectrograms. The locus theory was first applied to the voiced stop consonants. These sounds are produced by stopping the air at some point in the vocal tract. When the air is released the transient sound results. The voiced stop consonants are /b/, /d/, and /g/. The unvoiced stop consonants are /p/, /t/, and /k/.

Delattre, Liberman, and Cooper found that they could produce voiced stop consonants by converting hand painted spectrograms with the proper formant transitions to speech. They were able to locate the second formant locus of these sounds. The first formant locus appears to be equal to or lower than the second harmonic of the pitch frequency for all stop consonants. Later the work was extended to the third formant locus (23). However the results were inconclusive.

Hoffman studied the second and third formant transitions and also the frequency position of an initial short burst of noise. These studies were performed on the voiced stop consonants produced from hand painted spectrograms. The three clues studied appeared to be independent of each other since the optimum position of one clue did not vary when the other clues were varied.

Halle, Hughes, and Radley (21) have measured the frequency spectrum of the initial 20 milliseconds of the stop consonants. Their results seem to indicate that the stop consonant affects the steady state formant position of the adjacent vowel and also that the vowel affects the locus frequency of the consonants.

Lehiste and Peterson (36) have investigated the formant transitions of actual speech for most of the consonants. They have concluded that for two sounds in sequence each constitutes part of the environment of the other. It may be necessary to specify the transitions for each consonant-vowel pair separately. Data is given on the frequency ranges of second formant transitions.

Lehiste and Peterson also studied the characteristics of vowels that have time varying formants. One class called the glides consist of a short steady state formant position and a longer slow glide. These vowels are found in the words "fate" "lope", and "hurt". A second class of complex vowel phonemes

are the diphthongs. The diphthongs consist of two steady state formant positions. The transitions between formant positions are longer than the target positions. The diphthongs are the vowels found in the words "boy", "hide", "hued", and "howed".

The nasal phonemes are produced by closing or partially closing the oral cavities with the tongue or lips and opening the nasal cavity. The sound is modified by the nasal resonances. The nasal consonants are /m/, /n/, and /ng/ as in sing (47). Nakata (42) and House (25) have studied the nasal consonants by making electrical analogs of vocal tract and nasal cavities. The acoustical system is excited by the glottis and the sound is output at the nose. The transfer impedance function is similar to that for vowel production in that it contains several conjugate pair poles. In addition there are some conjugate pole-zero pairs due to the loose coupling of the oral cavities into the system. Nakata's analog consisted of several cascaded tank circuits tuned to the appropriate resonances. The first formant resonance had a bandwidth of 300 cps as contrasted with 30 to 100 cps for vowel bandwidths. This bandwidth was found to be the major factor in differentiating a voiced stop from a nasal consonant. House used a lumped parameter approximation to a distributed parameter analog. The lumped parameter analog consists of 35 cascaded inductance-capacitance pi sections

each of which represents one half cm of the vocal tract. Listening tests on the synthetically generated nasals resulted in better scores than test scores on actual spoken nasals. It was postulated that this was because the second formant was better defined on the synthetically generated sounds than on the naturally generated ones.

Another class of phonemes are the fricatives. The turbulence produced by the air passing through a small opening in the vocal tract is one sound source used in producing the fricatives. Fricative sounds may be voiced such as the first consonants in the vords "vest", "this", "zoo", and the middle consonant in the word "pleasure". The fricatives may also be unvoiced as in the first consonants of the words "fit", "hat", "thick", "see", and "shed" (47).

Heinz and Stevens (24) have made a theoretical study of the voiceless fricative consonants. On the basis of experimental studies they suggest that an equivalent circuit of the vocal tract would consist of a constant pressure noise source located somewhere within the vocal tract at the point of constriction. The poles of the transfer function would be the same as for a vowel produced with the same vocal tract configuration. The poles may be more heavily damped because of losses due to turbulence and losses due to the glottis being open. Zeros are located at the poles of the source impedance. Since the vocal tract posterior to the

constriction is shorter than the total vocal tract length the zeros of the transfer function are spaced further apart along the $j\omega$ axis of the s-plane than the poles. The poles and zeros tend to cancel for low frequencies but not at the higher frequencies. The voiceless fricative sounds were synthesized using a circuit containing a conjugate-pair of zeros and a conjugate-pair of poles. The circuit was excited with white noise. The synthetic spectra could be adjusted to fit the data for fricative consonants of Huges and Halle (29). Hughes and Halle measured the frequency spectra of fricative consonants in the context of words read by a large number of speakers, both male and female. The spectra was measured on a wave analyzer for a 50 millisecond segment of the fricative sound.

Listening tests were performed on the fricative consonant synthesizer using fricative-vowel syllables. It was found that the voiceless sounds /f/ and /th/ could not be differentiated without the vowel formant transition clues. The bandwidth of the poles and zeros could be varied over a 2 to 1 ratio without noticably affecting the intelligibility. The fricative sounds were all intelligible.

Hughes and Halle have also studied the voiced fricative consonants. The voiced fricatives usually but not always have a strong frequency component below 700 cps while the unvoiced fricatives never do. Above 700 cps the corresponding

pairs of voiced and unvoiced fricative consonants (/f/ and /v/; /s/ and /z/; /th/ both voiced and unvoiced) have similar spectra.

Schauer (52a) has studied the information structure contained in the frequencies from 0 to 15 cps for the spoken digits zero through nine. It was found that the frequency signal does contain some information about the spoken word.

Speech Processing Devices

There has been considerable interest in applying the principles previously discussed toward actual speech processing devices. These devices generally fall into two categories. One category is the coding devices that recode speech to minimize the required channel capacity to transmit the speech. Reductions of up to 60:1 have been claimed. The second category is speech recognition. This might be considered a special case of the coding problem in which the speech is coded in terms of the phonemes or words. In both of these problems one must find a set of properties that will adequately describe the speech.

David (5) has investigated the efficiency of representing speech with a speech processing device such as the vocoder. Three parameters can be used to describe such a device. They are the fidelity or quality of the reconstructed speech, the minimum bandwidth and signal to noise ratio required in a channel transmitting the reduced signals while maintaining a

given fidelity. A quantitative measure of fidelity is the articulation index which is the percent of words or sounds that are correctly understood by an observer listening to the device under test. Efficiency is defined as tne entropy rate of the reduced signal divided by the channel capacity used to transmit the reduced signal without coding in a noise free channel. A 16 channel vocoder with filters spaced according to a Koenig aural scale and a 90 percent consonant articulation is 22 times as efficient as a 3600 cps direct transmission speech channel. The bandwidth reduction of this channel is 15 to 1 which would indicate the vocoder channel can tolerate a smaller signal to noise ratio than the direct channel for the same articulation.

Koshikawa and Sugimoto (32) studied the information rate in the pitch signal for Japanese speech. They assumed that the pitch was frequency modulated by Gaussian noise. The bandwidth required to transmit 99 percent of the total energy in a signal proportional to the pitch frequency was found. The bandwidth is $5.16\,\sigma$ where σ is the variance of the Gaussian noise spectrum. This assumption checked very well with experimental data. Isolated vowel bandwidths for male speakers vary from 30 to 70 cps and from 40 to 90 cps for female speakers. Required bandwidths for continuous messages are slightly higher. Most vocoders used a 50 cps bandwidth for the pitch signal.

A method of reducing the bandwidth required to transmit speech even further than the vocoder does is to transmit only the formant frequencies instead of the entire frequency spectrum. Flanagan (17) has investigated the channel capacity and bandwidth necessary to transmit formant information of speech. The required bandwidths were found by tracing the formants on spectrograms and computing the Fourier transform of these signals. The upper limit of the precision necessary to transmit these signals is determined by the ability to just discriminate differences in formant frequencies. This has been found to be from ± 3 to ± 5 percent of the formant frequency. A ± 2 percent accuracy criterion was chosen which resulted in mean bandwidths of 7.1, 6.7, and 5.3 cps for the first three formants respectively. A 40 db signal to noise ratio results in a total error for each of the first three formants that is less than the just discriminable difference in formant frequency at least 65 percent of the time. This results in a channel capacity of 170 bits per second with a 20 cps bandwidth. The total system would also require channels for pitch, voicing excitation amplitude, and noise excitation amplitude. Flanagan claims that the entire system will result in a compression of 60 to 1 from a conventional telephone channel.

Flanagan (16, 18) has proposed two methods of tracking the formants. One method samples the short term frequency

spectrum to obtain repetitive time functions that represent the spectrum. In other words a time variable is substituted for the frequency coordinate of the spectrum. The frequency corresponding to the maxima of the time functions are found by differentiating with the appropriate circuitry. The disadvantage of this approach is that the system is easily confused by small local maxima in the spectrum. This can be partially overcome by using wide band filters to generate the spectrum so that the spectral details are masked.

A second method of formant tracking assumes that each of the formants occupy frequency ranges that do not overlap. The frequency corresponding to the maxima in each region is the formant frequency. The two frequency boundaries for the first three formants were 800 and 2280 cps. For male voices the major problem occurs at the boundary between the second and third formants since these formants often cross the 2280 cps boundary.

Tests with actual speech indicate that the second method tracks the formants more reliably. Other similar schemes have also been reported (27).

A set of properties that are invariant under changes in speakers and other modifications that would not affect the intelligibility of speech are sought in speech recognition work. Forgie and Forgie (19) have developed a computer program to recognize isolated words of the form /b/-vowel-/t/.

Recognition was made on the basis of the pitch frequency, the first three formant frequencies, and general shape information of the frequency spectrum. A sequential decision logic was used. The recognition score for 21 speakers both male and female, was 88 percent.

Foulkes (20), as has been mentioned previously, mapped the pitch frequency and first three formant frequencies onto a two dimensional space such that the phonemes would be separated with linear boundaries.

Welch and Wimpress (61) and Smith and Klem (53) have approached the speech recognition problem by using statistical decision theory. Welch and Wimpress used the pitch, first three formants and the formant amplitudes as their property space. The space was partitioned using quadratic surfaces and using hyperplanes. Both methods gave approximately the same result of 5.5 percent error on the speech data of Peterson and Barney (46).

Smith and Klem used the entire frequency spectrum sampled at 35 frequencies as their property space. The space was also partitioned using quadratic surfaces and using hyperplanes. The linear method gave a 13 percent error on Forgie and Forgie's (19) data. The quadratic method gave a 6 percent error.

Uhr and Vossler (56, 57) have applied a visual pattern recognition computer program to the problem of recognizing

spoken words by inspecting their spectrograms. The visual pattern recognition computer program generates and modifies its own property measurements in a way to improve its operation. This device was applied to the first five digits, zero through four, as spoken by several different speakers. A 100 percent recognition was attained for both the set of training data and the set of unknown test data after four passes of the training data.

Stevens and his coworkers (2, 22, 55) have discussed a model for speech recognition called analysis-by-synthesis. The frequency spectrum of speech is first converted to the articulatory domain. On an acoustical level this might consist of the formant for vowels and the formant loci for consonants. On an anatomical level this might be parameters describing the actual vocal tract configuration. A model then synthesizes a speech spectra from the articulatory data. A comparator then compares the two spectra and computes an error under some criterion. The articulatory data is then modified in a way to minimize the error. Experimental work has been attempted using the system to find the pole locations of vowels (2).

Several other recognition schemes have been proposed (11, 13, 28, 43, 44, 62). The basic principles used here are similar to recognition schemes previously discussed.

Denes and Mathews (9) have attempted to recognize spoken

digits by cross correlating the speech spectra with a set of reference spectra. The reference spectra are averages of several utterances of each of the words to be recognized. The reference spectra are also normalized on total energy. The best match is selected by finding the reference pattern which gives the greatest correlation coefficient with the pattern to be recognized. The time dimension of the spectra were scaled so that all patterns had the same length. When the reference patterns for the digits zero through nine were formed by five speakers and the device was presented utterances by the same five speakers, a 6 percent error rate was observed.

Dersch (10) has approached the speech recognition problem from the time domain. This is done by noting the time sequence of fricative and voiced sounds and noting whether the fricative sounds are "strong" or "weak". A sufficient logic can then be derived to recognize the digits zero through nine.

It has been observed that peak clipping the speech waveform does not remove the intelligibility of the speech although it does noticably affect the quality (34, 35). This is true particularly if the speech is differentiated or given a high frequency pre-emphasis of 20 db per decade before clipping. If the speech is differentiated and is infinitely clipped so that the speech waveform is represented by a train

of rectangular pulses, the intelligibility is greater than 90 percent for isolated words. This equivalent to 100 percent intelligibility for sentences. The intelligibility of clipped speech in noise is approximately the same as the intelligibility of unclipped speech with the same signal to noise ratio (49). The intelligibility of clipped speech showed a slight improvement for noise above 250 cps as compared to unclipped speech.

Spogen et al. (54) have investigated the intelligibility of speech sampled at the maximum and minimum points on the speech waveform. The signal generated is a box car waveform with the level of the signal at any time equal to the level of the speech waveform at its previous maximum or minimum. The signal had a higher intelligibility than clipped speech.

Several people have studied the effect of interrupting speech on the intelligibility of speech (1, 40). Miller and Licklider (40) tested the intelligibility of words using interruption frequencies of 0.1 to 10,000 cps and a percent on-time or percent of time the speech is not interrupted of 6 to 75 percent. Between 10 and 100 cps with 50 percent on-time the intelligibility is only slightly below the uninterrupted value.

Ahmend and Fatehchand (1) tested the effect of sample duration on the intelligibility of speech sounds. The intelligibility of various intervals of both initial and

final segments of consonants were tested.

Dukes (14) suggests that the high intelligibility of clipped speech is due to the similarity in the frequency spectrum of clipped and unclipped speech. He considers two types of signals, random signals which might be analogous to the fricative consonants and partially constrained signals which might be analogous to the voiced sounds. A theoretical study was made of the statistics of the signal necessary to have these properties.

Sakai and Inoue (51) propose a recognition scheme that measures the interval between zero crossings of the speech waveform. The time intervals are quantized into 14 ranges and the density of zero crossing intervals within each range is measured. This yields a spectral like plot. Chang et al. (3) plots the interval between zero crossing against time to obtain a display very similar to Sakai and Inoue's. Both of these devices only look at the zero crossing intervals at a syllabic rate.

Davis, Biddulph, and Balashek (7) have split the speech into two bands above and below 900 cps before clipping the speech. According to the data of Peterson and Barley this will separate the first two speech formants. The rate of zero crossings of signals in these two bands are plotted against each other on an oscilloscope screen. An attempt was made to recognize the spoken digits on the basis of these patterns.

Peterson (45), Marcou and Daguet (38), and Cherry and Phillips (4) have looked at the instantaneous frequency of the single-side-band modulated signal. The single-side-band frequency spectrum consists of the audio modulating signal spectrum translated up to the carrier frequency of the single-side-band signal. Both Peterson and Cherry used the average value of the instantaneous frequency in an attempt to measure the strongest formant frequency. Peterson compared the instantaneous frequency average value to the formant frequencies obtained from a spectrographic analysis of the same speech sounds. The instantaneous frequency average value deviated noticably from the expected formant frequency value obtained from a spectrographic analysis.

Marcou and Daguet derived a signal from single-side-band modulated speech by amplitude-limiting the single-side-band signal. The resulting signal was then single-side-band demodulated. The demodulated signal is only a function of the instantaneous frequency. This signal was found to be completely intelligible. A spectrographic analysis was performed on several examples of normal and processed speech vowel sounds. There was a marked similarity in the two spectra for each vowel sound.

METHOD OF APPROACH

The overwhelming majority of work done on the analysis of speech has been performed in the frequency domain, as can be seen by the discussion of the state of the art of speech analysis given in the introduction of this report. The initial step in most of these studies is to find the short term frequency spectrum of the signal. This usually involves averaging the signal over approximately 1 millisecond. Various schemes are then derived to extract the relevant information from the short term frequency spectrum.

A comparatively small effort has been given to analyzing the speech waveforms in the time domain. This paper studies some of the properties of the speech waveform in the time domain. The rate of zero crossings of the speech waveforms are investigated. The instantaneous frequency of the speech waveform as derived from a single-side-band modulated signal is also investigated. An attempt is made to relate the properties of the quantities studied to the theory of speech production.

Pitch synchronous interrupted speech is studied to see if there is any relationship between the intelligibility of the speech and the part of the pitch period that is interrupted. The variation in intelligibility with duration of interruption is also studied.

EQUIPMENT

A system was constructed that would produce a voltage proportional to the period between zero crossings. The system was later modified so that the output voltage approximated the Koenig frequency scale for a sine wave input. The circuit for this system is shown in Figure 1. The frequency detector approximates the Koenig frequency scale.

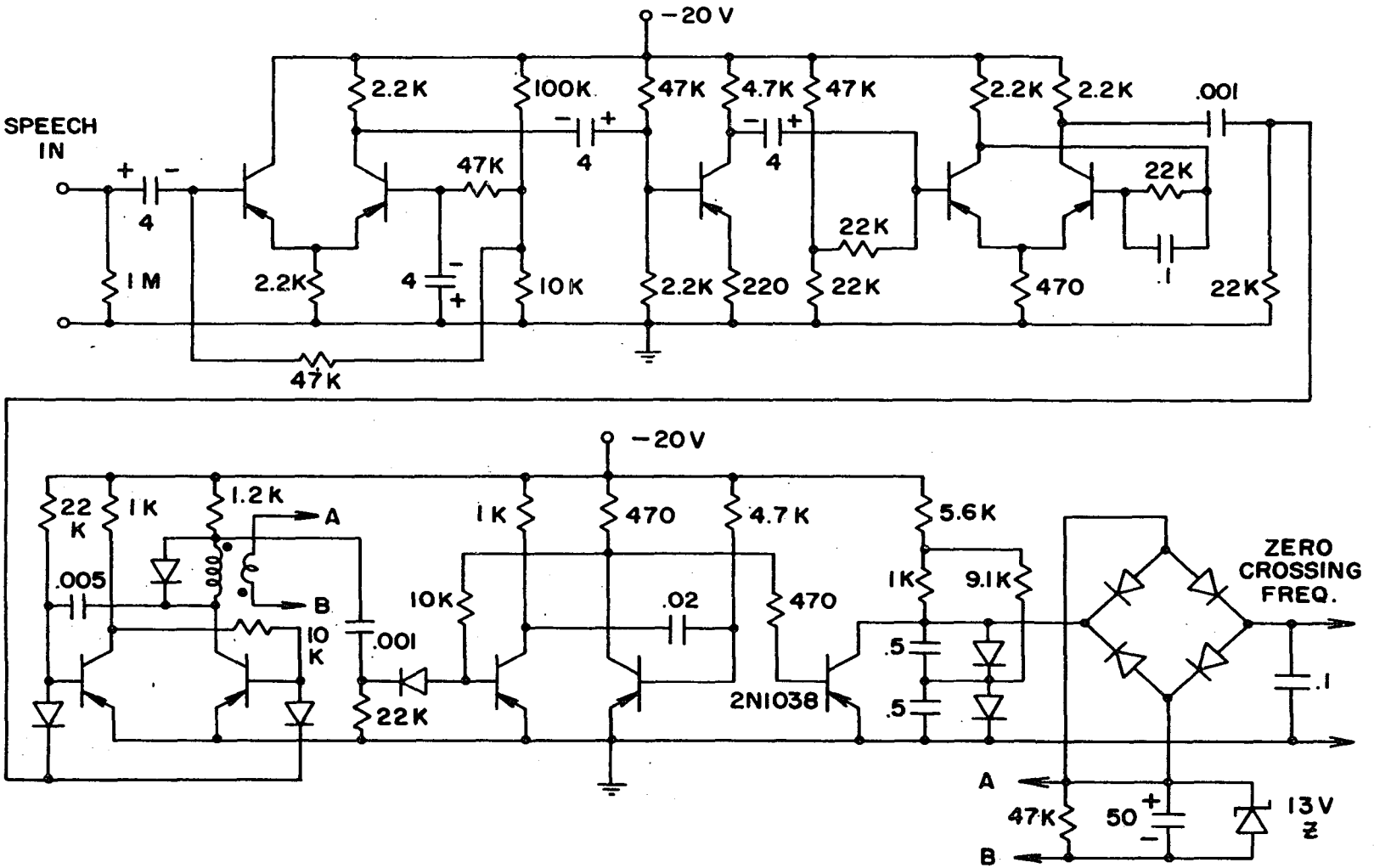
This signal is limited and then sent to a Schmitt trigger. This gives an infinitely peak clipped signal. The clipped signal then triggers a monostable multivibrator with a pulse width of 0.1 milliseconds. A delayed pulse is obtained by triggering another monostable multivibrator with the trailing edge of the first monostable multivibrator pulse. The delayed pulse sets a sweep generator to zero. As the sweep generator produces a voltage proportional to time, the peak voltage will be proportional to the period between zero crossings.

The sweep waveform could also be the sum of two decaying exponentials. When this is expressed in terms of the frequency of zero crossings, which is the reciprocal of the zero crossing period, the sweep waveform is

$$v = v_o \left(e^{-\frac{630}{f}} + 0.152\, e^{-\frac{90}{f}}\right). \qquad 1)$$

This expression very closely approximates the Koenig scale

Figure 1. Zero-crossing frequency circuit. All transistors are 2N581 except as noted. All capacitance is given in microfarads. All resistance is given in ohms unless otherwise stated.

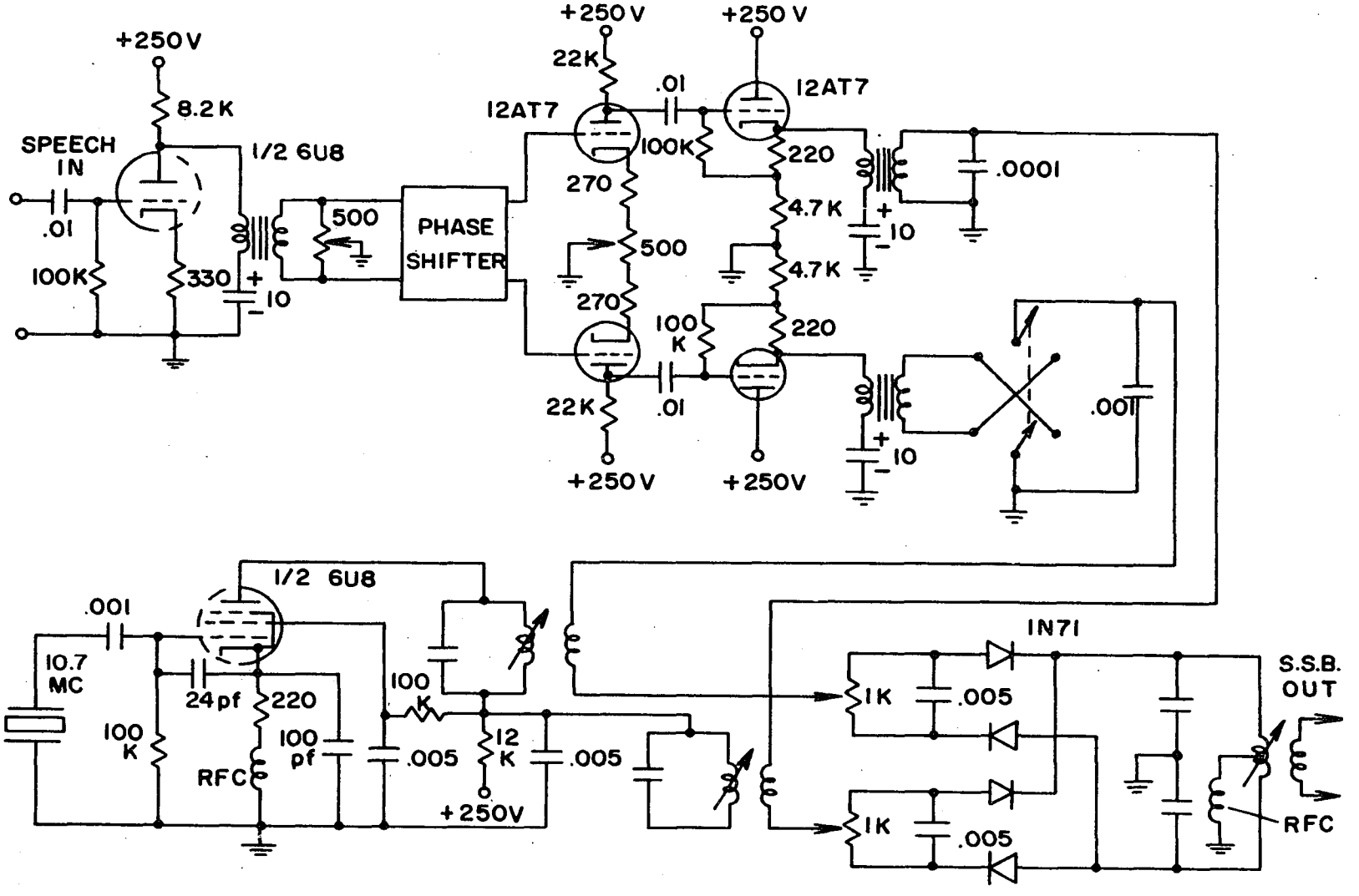

-20 V
SPEECH IN
2.2K
100K
47K
4.7K
47K
2.2K
2.2K
.001
47K
4
4
22K
1 M
2.2K
4
10K
2.2K
220
22K
470
.1
22K
47K
-20V
22 K
1K
1.2K
A
B
.005
10 K
.001
22K
1K
470
4.7K
5.6K
10K
.02
470
1K
9.1K
.5
.5
2N1038
ZERO CROSSING FREQ.
.1
A
47K
50
13V
B

which is linear with frequency below 1000 cps and logarithmic with frequency above 1000 cps. The numerical values in Equation 1 were obtained by empirical curve fitting.

The sweep generator voltage is sampled immediately prior to resetting to zero. The sampled voltage is then held until the next zero crossing giving a box car waveform. The voltage sampling is done with a diode gate that is actuated by the undelayed monostable multivibrator pulse.

A single-side-band suppressed-carrier modulator was constructed to process the speech. A single-side-band modulator linearly translates the audio frequency spectrum up to the carrier frequency. The modulator used produced two balanced modulated signals. The signals were phase-shifted in such a way that when the two balanced modulated signals were added, one side band of the signals added constructively and the other side band of the signals cancelled. As shown in the appendix, the carrier and audio signals for one balanced modulator must each be shifted 90 degrees relative to the carrier and audio signals for the other balanced modulator for proper side band cancellation. The single-side-band modulator circuit is shown in Figure 2 (58). The carrier frequency is 10.7 mc. A commercial audio phase shifter was used. The phase shifter is accurate to $\pm$ 1.5 degrees from 300 cps to 3 kc. Better than 40 db sideband and carrier suppression was obtained over the frequency range

Figure 2. Single-side-band modulator circuit. All capacitance is given in microfarads. All tank circuits are tuned to 10.7 mc. The phase shifter is a B and W model 350-2Q4. All resistance is given in ohms unless otherwise stated.


+250V
8.2K
SPEECH IN
.01
100K
1/2 6U8
330
10
500
PHASE SHIFTER
+250V
+250V
22K
.01
12AT7
12AT7
100K
220
270
500
4.7K
4.7K
.0001
10
270
100 K
220
22K
.01
10
.001
+250V
+250V
1/2 6U8
.001
10.7 MC
24pf
220
100 K
100 pf
RFC
.005
100 K
12 K
.005
+250V
IN71
1K
.005
1K
.005
S.S.B. OUT
RFC

of 300 cps to 3 kc.

A product detector was used to demodulate the single-side-band signal. This device beats a carrier frequency signal with the single-side-band signal and extracts the difference frequency. The circuit is shown in Figure 3.

A single-side-band signal may be thought of having an amplitude modulated component and a frequency modulated component. The signal can be expressed as

$$S(t) = a(t) \cos \left[\phi(t) + \omega_c t\right] \qquad 2)$$

where S(t) is the speech waveform, a(t) is the amplitude modulation, $\phi(t)$ is the phase modulation, and ω_c is the carrier frequency. The relationships of a(t) and $\phi(t)$ to S(t) are not unique. However, the expressions take an intuitive meaning if one thinks of a(t) as the envelope of the single-side-band signal and $\phi'(t)$ as the rate of zero crossings of the single-side-band signal. By use of trigonometric identities it can be shown that

$$a(t) = (S^2 + Q^2)^{1/2} \qquad 3)$$

and

$$\phi(t) = \cos^{-1} \frac{S}{(S^2 + Q^2)^{1/2}} \qquad 4)$$

where Q(t) is the signal obtained by phase shifting all the frequency components of the original signal, S(t), by -90 degrees (4).

The measuring of zero crossing rate of the audio signal

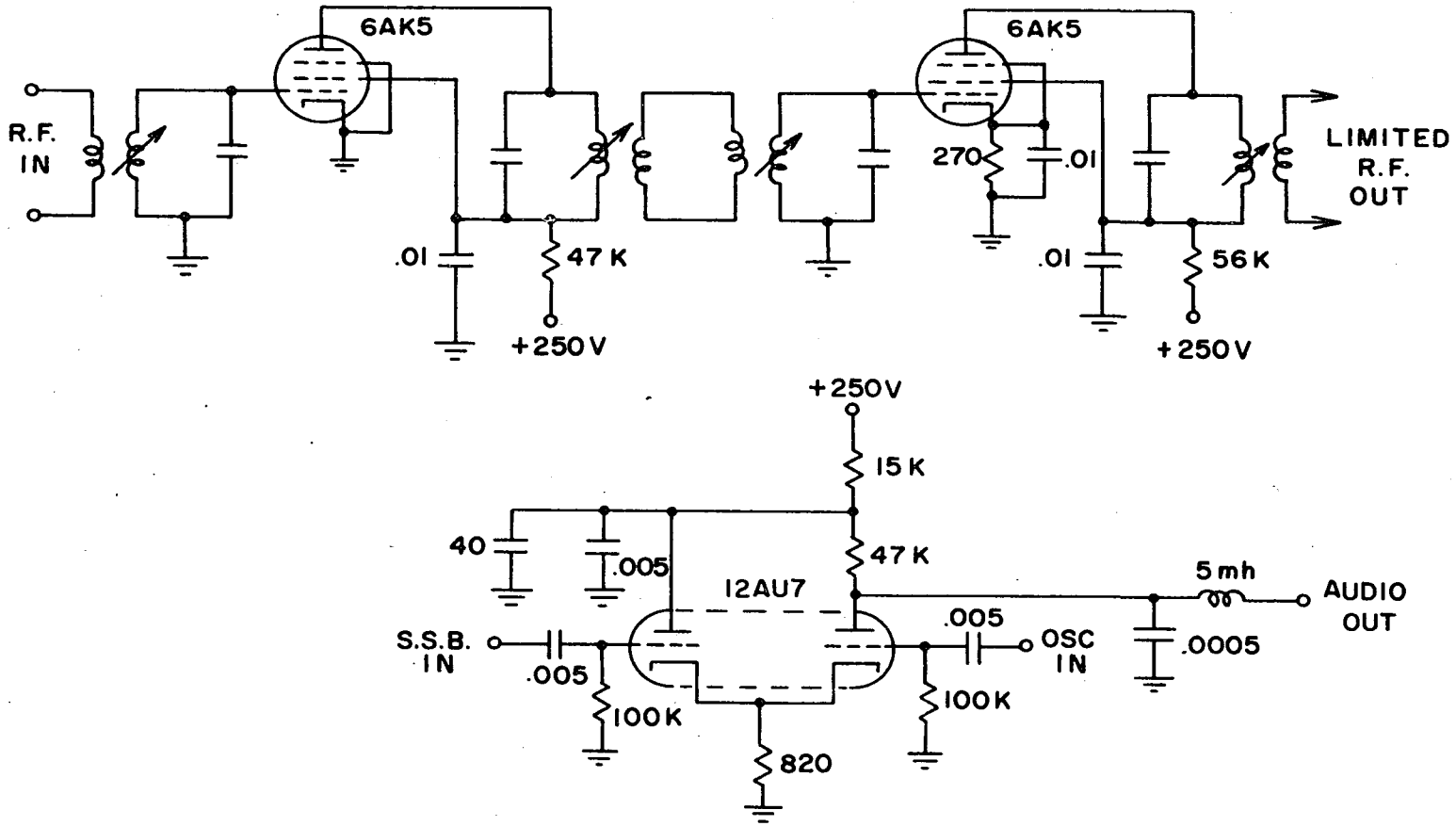


Figure 3. Limiter and product detector. All capacitance is given in microfarads. All tank circuits are tuned to 10.7 mc.

can be thought of as sampling the instantaneous frequency, $\emptyset'(t)$, at an audio rate. Since $\emptyset'(t)$ must contain higher frequencies it can be described better by sampling at a faster rate. Single-side-band modulating allows one to sample at a faster rate without changing the frequency spectrum of $\emptyset'(t)$. If a zero crossing occurs at time t_1 then another zero crossing must occur when the instantaneous phase changes by $\pm 2\pi$ or returns to the phase angle at time t_1. In other words

$$\int_{t_1}^{t_1+1/f} \emptyset'(t)dt = \pm 2\pi, 0 \qquad 5)$$

where f is the frequency of zero crossings and is the largest finite positive number satisfying Equation 5. The instantaneous frequency is then a generalization of the frequency of zero crossings. Although this is not the only generalization possible hopefully it is at least intuitively satisfying.

The system used to extract the amplitude and instantaneous frequency is shown in Figure 4. The amplitude variations were obtained by amplitude detecting the single-side-band signal on a Collins R 388/URR communications receiver. The single-side-band signal was limited to obtain a constant amplitude signal over a 40 db range of input amplitudes. The product detector will yield cos $\emptyset(t)$ when the input is the limited single-side-band signal. This signal is completely intelligible when converted to sound.

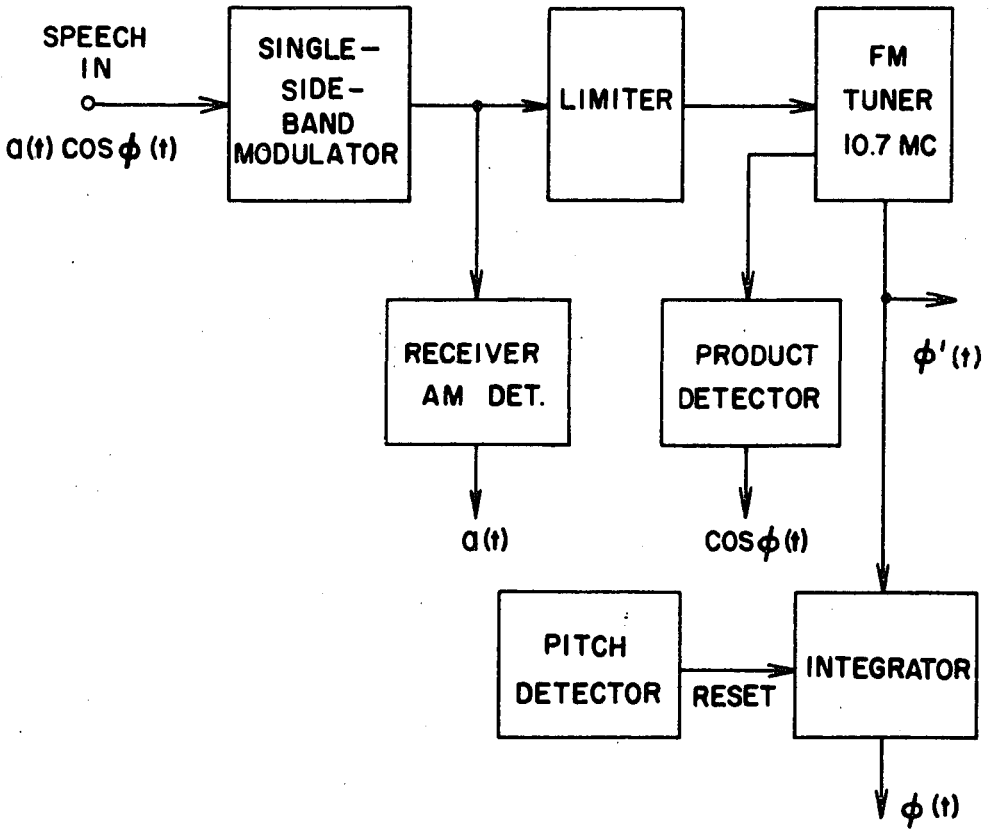


Figure 4. Instantaneous frequency system.

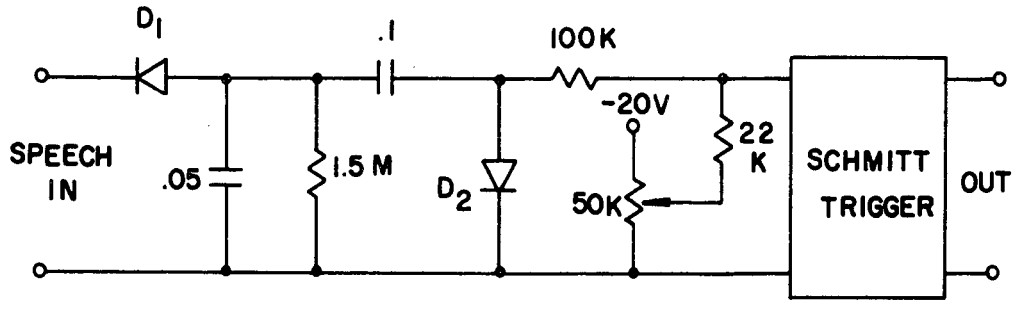


Figure 5. Pitch detector circuit.

The limited signal was frequency-detected, using the 10.7 mc I. F. strip of a Heathkit Model FM-4 FM tuner. The output of the FM tuner is $\emptyset'(t)$. This can then be integrated to give $\emptyset(t)$. The instantaneous frequency has an average value so $\emptyset(t)$ will increase with time.

The instantaneous phase is reset to zero at the beginning of each pitch period of voiced sounds by the pitch detector. This time corresponds approximately to the instant at which the glottis closes and is characterized by a burst of energy in the speech waveform. The circuit used to detect the beginning of the pitch period is shown in Figure 5. Diode D_1 will conduct only during the peaks of the input signal, since capacitor C_1 has a long discharge-time-constant. The time constant of C_1 can be adjusted so that diode D_1 only conducts once every pitch period. The pulse from D_1 is differentiated. Diode D_2 clamps the DC level. The Schmitt trigger has a sufficient amount of hysteresis to prevent it from triggering on small pulses that may occur during the pitch period. With proper adjustment of the DC triggering level a very stable operation is obtained.

The integrator and associated circuitry to re-zero the integral once every pitch period is shown in Figure 6. The integrator uses a Philbrick model UPA-2 operational amplifier. The input to the integrator, $\emptyset'(t)$, must be a negative voltage for proper operation of the rezeroing circuit.

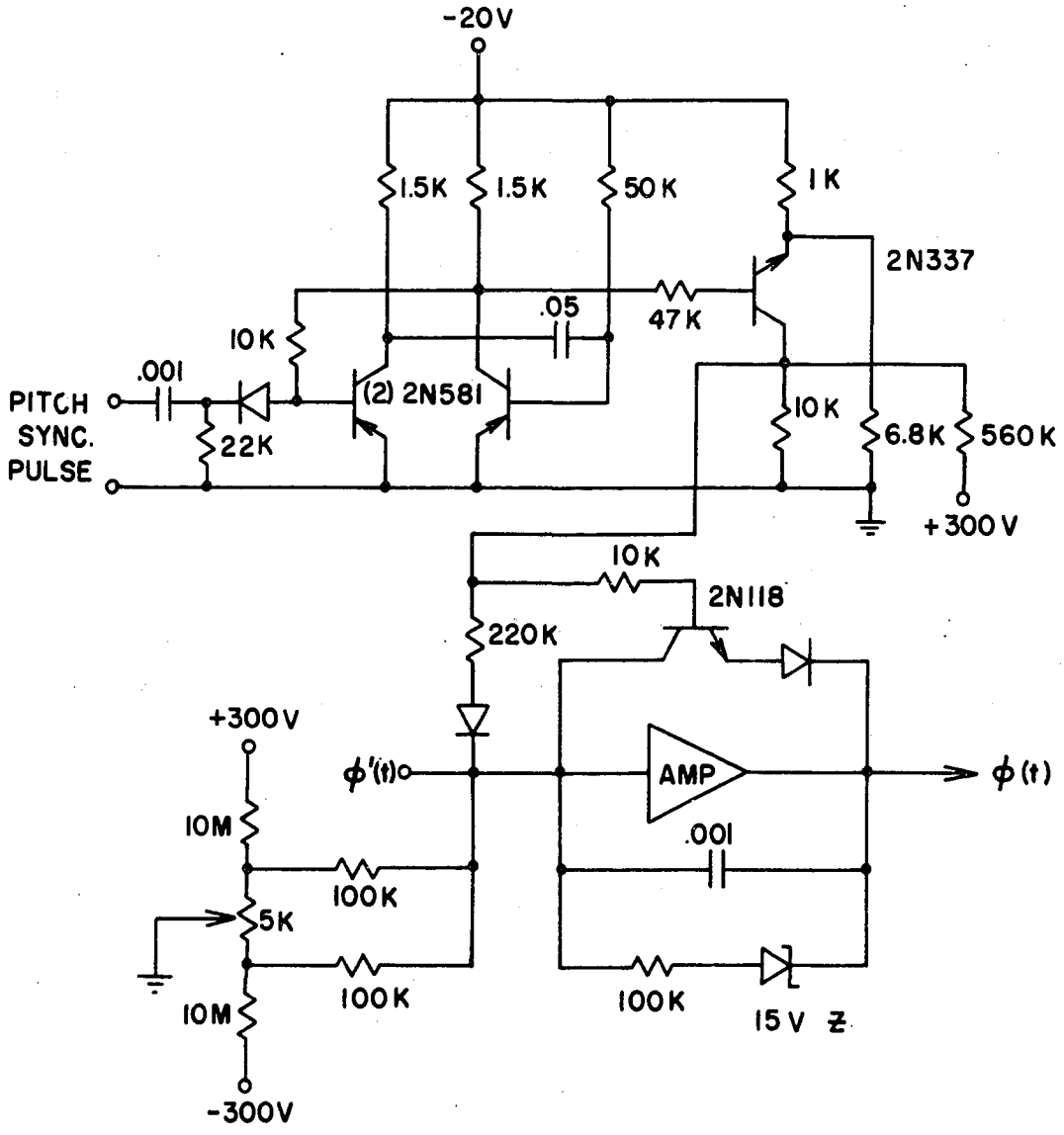


Figure 6. Pitch synchronous integrator circuit. All capacitance is given in microfarads.

The effects of periodically interrupting the processed speech at a pitch synchronous rate was also studied. Figure 7 shows the system used to generate interrupted speech. The pulse widths of the two monostable multivibrators can be adjusted to give any gatting pulse desired. The gate is the same diode gate shown in Figure 1. The output voltage of the gate is zero during the off time.

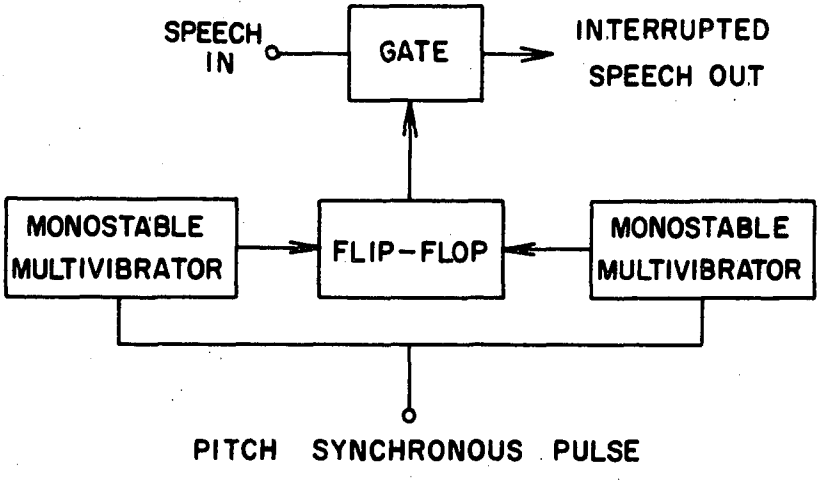


Figure 7. Interrupted speech system.

EXPERIMENTAL RESULTS AND DISCUSSION

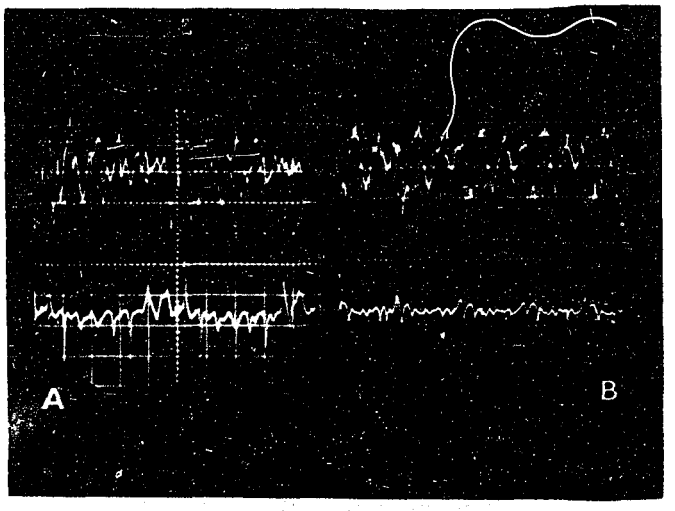
Studies were made of a wide variety of speech sounds. The speech sounds were processed with the equipment described in the preceding section. The processed speech was then displayed on an oscilloscope in various ways in the hopes of finding some unique characteristics of the speech sounds. Approximately two pitch periods of the male voiced sounds were displayed.

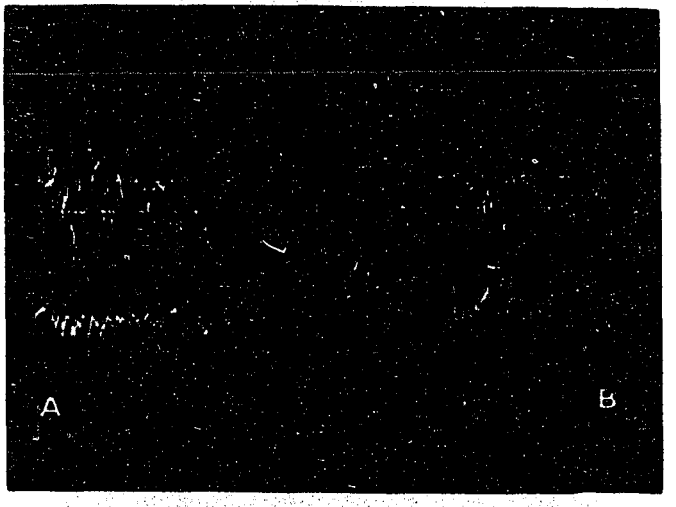
Time Domain Waveforms of the Vowels

Figure 8 shows the speech waveform, upper trace, and the instantaneous frequency, lower trace, for 2 speakers uttering the word "head". In these recordings the vowel sound was unnaturally prolonged to give a display that could be observed on the oscilloscope. There is considerable evidence in the literature that natural speech sounds never persist long enough to reach a steady state periodic waveform. However since the prolonged monothong speech sounds are completely intelligible all the necessary information must be contained in the steady state waveform. Figure 8a is a male voice (D.C.). Approximately two full pitch periods are displayed. The time calibration on the horizontal scale is 2 milliseconds per large division. Figure 8b is a female voice (S.D.). Approximately five pitch periods are displayed for the same sweep rate as in Figure 8a. Zero instantaneous

Figure 8. Speech waveform, top trace, and instantaneous frequency, bottom trace, of the vowel sound in the word "head", a) as spoken by D. C. (male), and b) S. D. (female).

Figure 9. Speech waveform, top trace, and instantaneous frequency, bottom trace, of several vowel sounds as spoken by J. W. (male). Vowel sound in the words, a) "hid", b) "heed", c) "had", and d) "head".


A
B


A
B

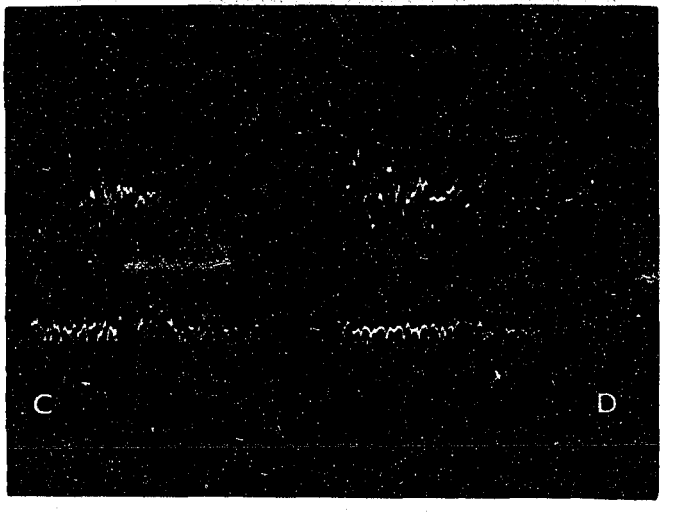
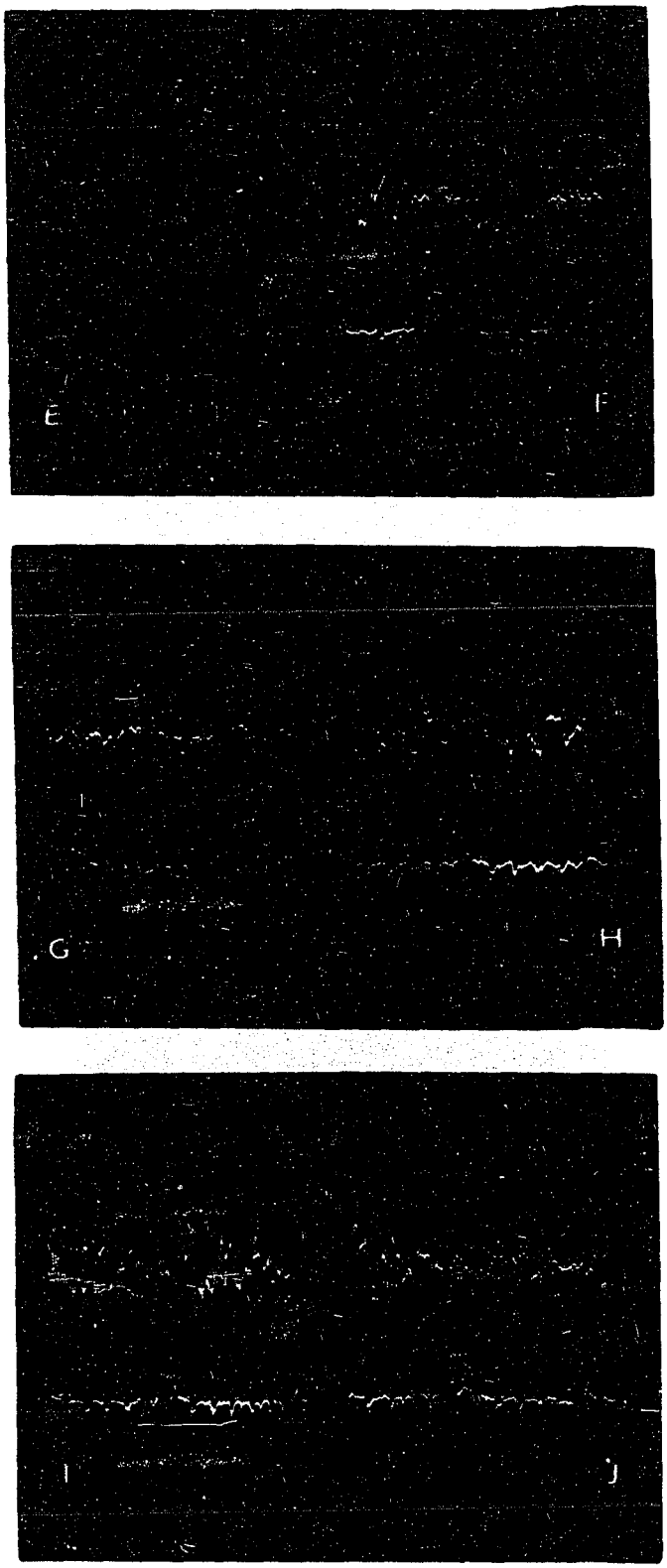

C
D

Figure 9. (Continued) Vowel sounds in the words e) "hawed", f) "hod", g) "hood", h) "who'd", i) "hud", and j) "heard".


E
F
G
H
I
J

frequency corresponds to the third horizontal line from the bottom of the picture. The approximate instantaneous frequency calibration is 2.25 kc per large division on the vertical scale. Positive frequency is up. This instantaneous frequency calibration will be the same for all plots of instantaneous frequency to be discussed.

The exposure times of the photographs were all 0.04 second. The sweep time for all cases was 20 milliseconds. Allowing time for the next trigger pulse to occur, the oscilloscope pictures should display approximately 1 1/3 to 1 1/2 traces. The oscilloscope trigger pulses were not synchronized with the camera shutter.

Figure 9 shows the speech wave form and instantaneous frequency of ten vowel sounds as spoken by the author. The vowel sounds are those in the words "hid", "heed", "had", "head", "hawed", "hod", "hood", "who'd", "hud", and "heard" and are found in parts a to j of Figure 9 respectively. These are the same words that are studied by Peterson and Barney (46).

The upper trace in each figure is the speech waveform, the lower trace is the corresponding instantaneous frequency. All of the vowel sounds except the vowel in "heard" are monothongs. That is, they are characterized by one steady state vocal tract position. The vowel in the word "heard" however is a glide. The articulars slowly glide to a steady

state position. The waveform displayed in Figure 9j was produced by the terminal position of the vocal tract for the word "heard".

It will be noted that the instantaneous frequency waveform is quite complex and there does not appear to be any obvious distinguishing characteristics for the 10 vowel sounds. Each of the 10 vowel sounds have quite distinctive sounds with the possible exception of the vowels in the words "hod" and "hawed". Peterson and Barley reported that their listeners had considerable trouble distinguishing between the two words.

A better insight into these problems can be obtained by considering the instantaneous frequency of the sum of two harmonically related sinusoidal signals. In speech this would correspond to the two lowest formants assuming that the formants have a very narrow bandwidth.

Cherry and Phillips (4) have computed the instantaneous frequency of a two-frequency signal. The signal can be written

$$S(t) = \cos\ \omega_1 t + A \cos\ \omega_2 T \qquad 6)$$

The instantaneous frequency is found by differentiating the instantaneous phase which is given by Equation 4. Differentiating Equation 4 gives

$$\phi'(t) = \frac{S\,Q' - S'\,Q}{S^2 + Q^2} \qquad 7)$$

where the prime denotes differentiation with respect to time. The S(t) is given by Equation 6 and the Q(t) is given also by Equation 6 with the $\cos \omega t$ terms replaced with $\sin \omega t$ terms. The two component instantaneous frequency then becomes

$$\phi'(t) = \frac{\omega_1 + A^2 \omega_2 + A(\omega_1 + \omega_2)\cos(\omega_2 - \omega_1)t}{1 + A^2 + 2A\cos(\omega_2 - \omega_1)t} . \qquad 8)$$

The average value of $\phi'(t)$ will be ω_1 for $|A| < 1$ and will be ω_2 for $|A| > 1$. For $|A| = 1$ the average value will be $\omega_1 + \omega_2/2$. The instantaneous frequency will be periodic in $\omega_2 - \omega_1$ as can be seen in Equation 8.

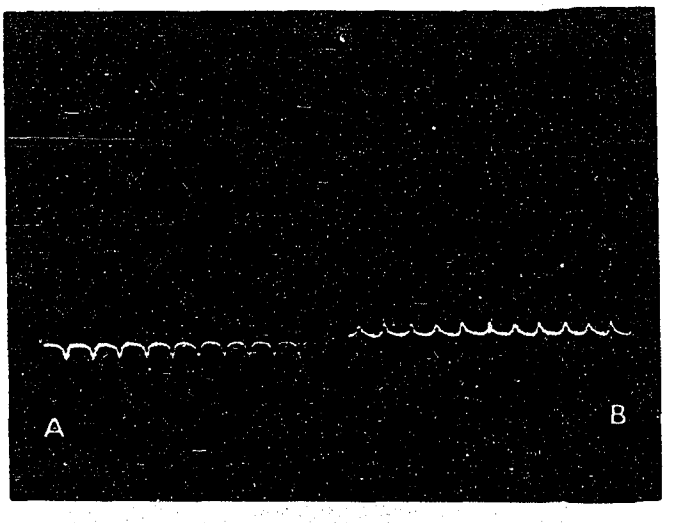
Figure 10 shows the instantaneous frequency of a signal defined by Equation 6. The parameters ω_1 and ω_2 are $(400)2\pi$ radians per second and $(1000)2\pi$ radians per second respectively. The two frequencies are not synchronized. In Figure 10a the parameter A is 0.5 and in Figure 10b it is 2. Note the periodic structure with a fundamental frequency of 600 cps. Also the average value is equal to the frequency of the stronger component. The instantaneous frequencies of separate 400 cps and 1000 cps signals are superimposed on these pictures. They are the lower and upper heavy horizontal lines respectively. Zero frequency is the third horizontal grid line from the bottom of the picture.

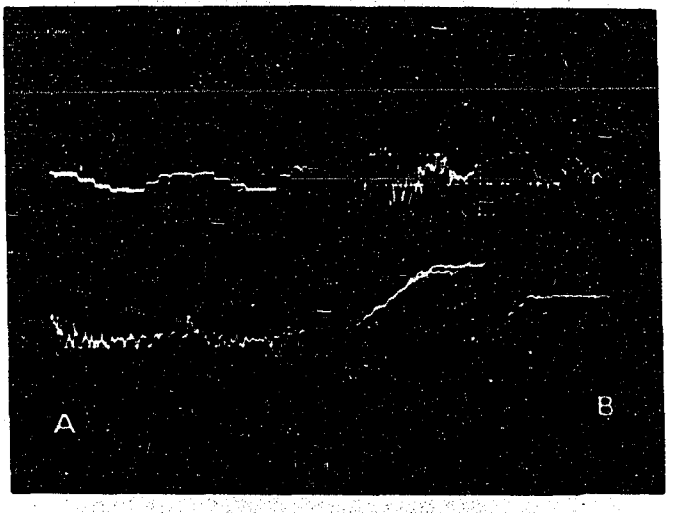
One should note the striking similarity in the instantaneous frequency waveforms of Figure 10 and Figures 8 and 9. This might suggest that one could determine the

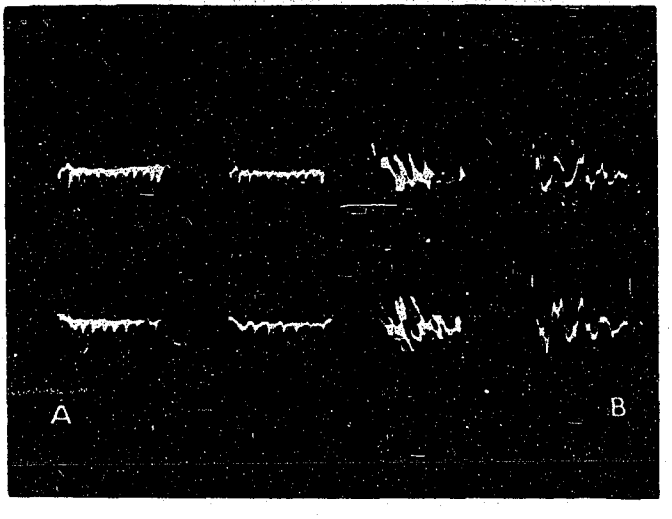
Figure 10. Instantaneous frequency of the sum of a 400 and 1000 cps signal for an amplitude ratio, A, of a) 0.5 and b) 2.

Figure 11. a) Frequency of zero crossings, top trace, and instantaneous frequency, bottom trace, for the vowel sound in the word "head".
b) Speech waveform, top trace, and instantaneous phase, bottom trace, for the vowel sound in the word "heed". Spoken by J. W.

Figure 12. a) Instantaneous frequency plotted against the instantaneous phase, left trace, and plotted against time, right trace, for the vowel sounds in the words "heed", top trace, and "hood", bottom trace. b) Speech waveform plotted against instantaneous phase, left trace, and plotted against time, right trace, for the vowel sounds in the words "head", top trace, and "hood", bottom trace. Spoken by J. W.


A
B


A
B


A
B

formant frequencies of speech by measuring the average value of the instantaneous frequency and measuring the periodic frequency of the instantaneous frequency waveform. It should also be noticed from Figure 10 that when the low frequency component is the strongest the scallops in the waveform tend to point down as in Figure 10a. When the high frequency component is strongest the scallops in the waveform tend to point up as in Figure 10b. This information could then be used to determine whether the instantaneous frequency average value is equal to the first or second formant frequency.

There are problems involved in using this approach to study the speech waveform. First, voiced speech contains more than two formant frequencies. Although the higher formant frequencies are weaker, they will corrupt the instantaneous frequency waveform. Perhaps more important the formant frequencies have bandwidths of 70 to 100 cps. An average male voice has a pitch frequency of about 120 cps. This means that the harmonics of the glottal pulses close to the formant frequency will have an appreciable amplitude.

The group of harmonic frequencies close to the formant frequency can be thought of as a carrier with two side bands. The carrier frequency being the harmonic frequency with the largest amplitude. The one formant signal can be expressed as

$$S_f(t) = \sum_{n=-N_1}^{N_2} A_n \cos\left[(\omega_c + n\omega_p)t + \theta_n \right] \qquad 9)$$

where ω_c is the carrier frequency, ω_p is the pitch frequency, and A_n and θ_n are the amplitude and phase angle of the n^{th} harmonic above ω_c respectively. The expression is summed over the range of frequencies contributing to the formant under consideration. Equation 9 may be rewritten as

$$S_f(t) = \left[\left(\sum_{n=-N_1}^{N_2} An \cos(n\omega_p t + \theta_n) \right)^2 + \left(\sum_{n=-N_1}^{N_2} A_n \sin(n\omega_p t + \theta_n) \right)^2 \right]^{1/2} \cos\left[\omega_c t + a(t) \right] \qquad 10)$$

where

$$a(t) = \tan^{-1}\left[\frac{\sum_{n=-N_1}^{N_2} A_n \sin(n\omega_p t + \theta_n)}{\sum_{n=-N_1}^{N_2} A_n \cos(n\omega_p t + \theta_n)} \right] \qquad 11)$$

The single formant signal can now be seen to consist of a carrier frequency with amplitude and phase modulation. If we assume only three adjacent harmonics are important, the envelope of the signal is periodic in the pitch frequency. Assuming further that the pitch frequency is 1 1/2 times the formant bandwidth the signal will have more than 60 percent amplitude modulation. Addition of more harmonics should not change the fundamental shape of the envelope.

Consider the case of only two formant frequencies with each formant frequency amplitude modulated at the pitch

frequency rate. If the envelopes of these two carrier frequencies are not in phase, one formant frequency may have the largest amplitude over part of the pitch period and the other formant frequency will have the largest amplitude over the rest of the pitch period. This will cause the average value of the instantaneous frequency to jump from one formant frequency to the other. This phenomenon can be seen in all of the pictures of Figure 9. Over most of the pitch period the scallops in the instantaneous frequency waveform point down which indicates that the first formant is the strongest frequency. However just before the glottis closes starting a new pitch period, the instantaneous frequency waveform jumps to a higher value indicating that here the higher frequencies predominate.

The vowel sound in the word "heed" has a particularly strong second formant as can be seen in Figure 9b. The highly unstable appearance of the instantaneous frequency in this picture is because the first two formants have approximately equal average amplitudes.

The instantaneous phase, which is the integral of the instantaneous frequency, shows this bistable effect particularly well. Here the high frequencies are suppressed at the rate of 20 db per decade. Figure 11b shows the instantaneous phase for the vowel sound in the word "heed". The abrupt change in slope indicates the shift in the

average value of instantaneous frequency. The instantaneous phase is re-zeroed once every pitch period.

A female voice with a pitch frequency of 250 cps has its harmonic frequencies widely separated with respect to the formant bandwidth. Consequently the amplitude modulation of the formant frequency by the formant bandwidth will be appreciably smaller than that for a male voice. Figure 8b shows the instantaneous frequency of a female voice for the vowel sound in "head". The pitch frequency is about 250 cps. Note that there is no sign of instability in the average value of the instantaneous frequency. As a comparison Figure 8a shows the same vowel sound for a male voice. Here the positive spikes indicate shifts in the instantaneous frequency average value. The formant frequencies of the female voice of Figure 8b do have some amplitude modulation. This causes a corresponding amplitude variation in the scalloped waveform of the instantaneous frequency. This amplitude variation is periodic in the pitch frequency.

Peterson (45) found discrepancies between the formant frequency as measured by the average value of the instantaneous frequency and as measured by the spectrograph. He averaged the instantaneous frequency by passing it through a 30 cps low pass filter. This smooths out any shifts in the short time average value. It may be seen from the preceeding discussion that the resultant average value for

a two formant signal will be somewhere between the two formant frequencies.

Another way to relate the instantaneous frequency waveform to the known properties of speech is with the pole-zero functions of the vocal tract and the glottis. The transfer impedance function of the vocal tract consists entirely of poles except for a pole-zero pair representing the radiation impedance of the mouth. The position of the poles of the transfer function are determined by the position of the articulars. The articulars cannot change at a rate faster than 30 cps so the pole positions will be very stable for a given voiced sound. The Laplace transform of one glottal pulse will consist of an infinite number of zeros and no poles. The waveform of the glottal pulses has been observed to vary considerably from pulse to pulse (41). The variation in the pulse waveform means that the zeros of the glottal pulse will change position for each pulse.

The Laplace transform of the sound pressure caused by the n^{th} glottal pulse is the product of the zeros of the glottal pulse, the poles of the transfer function, and the pole-zero pair of the radiation impedance function. The waveform in the time domain will consist of damped sinusoids. The damping factor and frequency will be determined by the poles and will not be affected by the glottal pulse waveform. The glottal pulse waveform will affect the residues of the

poles only. The sound pressure waveform can be written

$$S_n(t) = \sum_{i=0}^{I} A_{in}\, e^{-\sigma_i t} \cos(\omega_i t + \theta_{in}) \qquad 12)$$

where ω_i and σ_i/π are the i^{th} formant frequency and bandwidth respectively. Parameters A_{in} and θ_{in} are the amplitude and phase of the i^{th} formant produced by the n^{th} glottal pulse.

The total sound pressure waveform during the N^{th} pitch period can be written

$$S(t) = \sum_{n=0}^{N} \sum_{i=0}^{I} A_{in}\, e^{-\sigma_i(t+\tau_n)} \cdot\cos\left[\omega_i(t+\tau_n) + \theta_{in}\right] \qquad 13)$$

where $\tau_n - \tau_{n+1}$ is equal to the n^{th} pitch period and $\tau_N = 0$. The N^{th} glottal pulse then starts at time equal zero. Equation 13 can be written as

$$S(t) = \sum_{i=0}^{I} \left[\left(\sum_{n=0}^{N} A_{in}\, e^{-\sigma_i \tau_n} \cos(\omega_i \tau_n + \theta_{in}) \right)^2 + \left(\sum_{n=0}^{N} A_{in}\, e^{-\sigma_i \tau_n} \sin(\omega_i \tau_n + \theta_{in}) \right)^2 \right]^{1/2} \cdot e^{-\sigma_i t} \cos(\omega_i t + \alpha_{iN}) \qquad 14)$$

for $\quad 0 < t < \left|\tau_{N+1}\right|$

where

$$\alpha_{iN} = \tan^{-1}\left[\frac{\sum_{n=0}^{N} A_{in}\, e^{-\sigma_i \tau_n} \sin(\omega_i \tau_n + \theta_{in})}{\sum_{n=0}^{N} A_{in}\, e^{-\sigma_i \tau_n} \cos(\omega_i \tau_n + \theta_{in})}\right] \qquad 15)$$

Equation 14 is the actual speech waveform from t = 0 to $t = |\tau_{N+1}|$. At τ_{N+1} another glottal pulse occurs and N must be replaced by N+1 in Equations 14 and 15.

Since A_{in} and θ_{in} are not constant with n the amplitude and phase of each damped sinusoid in Equation 14 will be different for each pitch period. This variation can be seen in the top traces of Figure 9 which are the speech waveforms. Approximately half the waveform in each picture has been traced twice. The variation between the two traces is particularly noticable during the glottal pulse i.e. when the glottis is open. This occurs just prior to the large pulse of energy in the speech waveform. In the female voice the glottal pulse rate is approximately doubled but the time constants of the formant bandwidths are not appreciably changed. This would increase the exponential weighting factors of Equations 14 and 15. The amplitude and phase of the damped sinusoids of Equation 14 will depend more on the past history and should therefore be more stable than for a male voice. The female voice of Figure 8b does indeed have a more stable waveform than the male voice of Figure 8a. Only slight variations in the peaks of the waveform can be

seen.

The instantaneous frequency will also display an instability due to the shifting position of the glottal pulse zeros. The amplitude and phase of each of the frequencies will change for each pitch period. This will cause a phase shift in the amplitude modulated envelope of each formant frequency. The jumps in the instantaneous frequency average value will occur at different points in the pitch period depending on the phase and amplitude of the formant envelope. These variations will also effect the instantaneous frequency waveform. The bottom traces of Figure 9 show the variations in the instantaneous frequency between two pitch periods.

These variations can be seen clearly on the part of the picture that has a double trace. On some pitch periods the average value does not jump but the amplitudes of the waveforms all vary considerably.

The female voice suppresses the variations in the instantaneous frequency due to the variation in glottal pulses. The reason for this stability is the same as that given earlier for the stability in the speech waveform, i.e. the female voice harmonics are spaced further apart than the male voice harmonics. The instantaneous frequency of the female voice shown in the lower trace of Figure 8b has a high correlation from one pitch period to the next as would be expected.

It is possible that the speech waveform variations due to variations in the glottal pulse waveform might be suppressed by autocorrelating the speech waveform. This will remove the statistical variations in the waveform. The speech waveform of Equation 14 may be generalized to include all values of time. This gives

$$S(t) = \sum_{n=-\infty}^{\infty} \sum_{i=0}^{I} B_{in}\, e^{-\sigma_i(t-\tau_n)} \cos\left(\omega_i(t-\tau_n)+\alpha_{in}\right) \left(U(t-\tau_n) - U(t-\tau_{n+1})\right) \tag{16}$$

where B_{in} is the amplitude expression given in Equation 14 and $U(t)$ is a unit step function. To simplify the problem one may assume that the pitch period is constant so that $\tau_n = n\Delta\tau$ where $\Delta\tau$ is the pitch period and assume that all the B_{in} and α_{in} for $n = 0, \pm 1, \pm 2, \ldots$ are statistically independent. The autocorrelation function is

$$\phi(t) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} S(x)\, S(x+t)\, dx \tag{17}$$

The random variable

$$\int_{n\Delta\tau}^{(n+1)\Delta\tau} S(x)\, S(x+t)\, dx \tag{18}$$

is statistically stationary with respect to n. Therefore Equation 17 may be replaced with the ensemble average of 18. The autocorrelation function now becomes

$$\phi(t) = \sum_{j=0}^{I} \sum_{i=0}^{I} e^{-\sigma_j(t-m\Delta\tau)} \int_0^{\Delta} e^{-(\sigma_i+\sigma_j)x} \cdot \overbrace{\left[A_{in}\cos(\omega_i x+\theta_{in})\right]} \cdot \overbrace{\left[A_{jn}\cos(\omega_j x+\omega_j(t-m\Delta\tau)+\theta_{jn})\right]} \, dx \qquad 19a)$$

for $m\Delta\tau < t < (m+1)\Delta\tau$, $m \geq 1$ and

$$\phi(t) = \sum_{j=0}^{I} \sum_{i=0}^{I} e^{-\sigma_j t} \int_0^{\Delta\tau-t} e^{-(\sigma_i+\sigma_j)x} \cdot \overbrace{\left[A_{in}A_{jn}\cos(\omega_i x+\theta_{in})\cos(\omega_j x+\omega_j t+\theta_{jn})\right]} \, dx$$

$$+ \int_{\Delta\tau-t}^{\Delta} e^{-(\sigma_i+\sigma_j)x} \overbrace{\left[A_{in}\cos(\omega_i x+\theta_{in})\right]} \cdot \overbrace{\left[A_{jn}\cos(\omega_j x+\omega_j t+\theta_{jn})\right]} \, dx \qquad 19b)$$

for $0 < t < \Delta\tau$ where the wavy line indicates the ensemble average on n.

The autocorrelation function can be seen from Equation 19 to be periodic for $t > \Delta\tau$. It will reduce to the form

$$\phi(t) = \sum_{n=-\infty}^{\infty} \sum_{i=0}^{I} \bar{B}_i e^{-\sigma_i(t-n\Delta\tau)} \cos\left(\omega_i(t-n\Delta\tau)+\bar{\theta}_i\right) \cdot \left(U(t-n\Delta\tau)-U(t-(n+1)\Delta\tau)\right) \qquad 20)$$

for $t > \Delta\tau$

where $\bar{B}_i$ and $\bar{\theta}_i$ are independent of n and are determined by Equation 19a. The autocorrelation function for $t > \Delta\tau$ has the same form as the original signal given by Equation 16.

However the statistical variations have been removed.

A possible disadvantage of using this approach is it will take several pitch periods to obtain a good approximation to the ensemble average of Equation 18. Due to the transient nature of speech there may not be enough periods available.

Schroeder and Atal (52b) have proposed a short term autocorrelation function and a corresponding short term power spectra. The short term autocorrelation function is

$$\phi(\tau, t) = \int_{-\infty}^{t} S(x)\, S(x - |\tau|)\, r_{\phi}(t-x)\, dx \qquad 21)$$

where $r_{\phi}(t)$ is a weighting function, t is the present time and τ is the time interval shifted. For a good choice of the weighting function, the autocorrelation function should be approximately periodic on the τ axis.

The autocorrelation approach was not pursued experimentally in the investigation reported here, primarily because an autocorrelator was not available nor was an analog to digital converter which would allow the correlation to be performed on the digital computer.

The variation of the glottal pulse waveforms and the finite formant bandwidths will also affect the short term spectral analysis on speech sounds. Equations 14 and 15 describe a voiced speech sound. One may assume for the time being that the formant frequencies, ω_i, are well separated as compared to the formant bandwidths, σ_i/π. The amplitude

frequency spectra will be the sum of the individual formant spectra. The short term individual formant spectra will approximately be the Fourier transform of $e^{-\sigma_i t} \cos \omega_i t$. The over-all amplitude of the individual formant spectra will depend on the position of the glottal pulse zeros. Since it was assumed that the formant spectra did not appreciably overlap, variations in the amplitude formant spectra will not effect the shape of adjacent formant spectra. Variations in the glottal pulse zeros will not effect the frequencies of local maxima or the bandwidth of the local maxima. If two formant frequencies are very close, the skirt of one spectra will overlap on the other spectra. Variations in the amplitude of the individual spectra will then effect the frequencies of local maxima. For this effect to be large the difference in the formant frequencies would have to be less than the formant bandwidth.

The observations on the instantaneous frequency can now be extended to the frequency of zero crossings. As was mentioned in the last chapter a zero crossing will occur every time the instantaneous frequency changes by 2π starting from a zero crossing point. A zero crossing will also occur when the instantaneous phase returns to its value at the last zero crossing without changing by 2π. The frequency of zero crossings can then be written as

$$f_z = \frac{1}{2\pi}\left|\frac{1}{\Delta t}\int_{t_1}^{t_1+\Delta t} \phi'(t)dt\right| \qquad 22)$$

for the case in which the instantaneous phase changes by 2π. The parameters t_1 and $t_1+\Delta t$ are the times of two adjacent zero crossings. The frequency of zero crossings here is the magnitude of the average value of $\phi'(t)$ averaged over the time interval Δt. The frequency of zero-crossings has no relationship to the average value of the instantaneous frequency when

$$\int_{t_1}^{t_1+\Delta t} \phi'(t)dt = 0 \qquad 23)$$

and $f_z = 1/\Delta t$. The instantaneous frequency averaged over a time interval Δt can be either negative or zero since $\phi'(t)$ can be negative as well as positive. For a signal made up of two frequencies the instantaneous frequency will have large negative peaks if the lower frequency component has a slightly larger amplitude than the higher frequency component. The instantaneous frequency will have negative peaks for an amplitude ratio less than 3. The peaks will increase as the amplitude ratio approaches one. The frequency of zero crossings may not have the same value as the instantaneous frequency averaged over that same time interval.

The short-term average value of the zero crossing frequency therefore cannot be interpreted as the formant

frequency. However, the instabilities of the instantaneous frequency will be reflected in the zero crossing frequency. Figure 11a compares the frequency of zero crossings, top trace, to the instantaneous frequency, bottom trace, for the vowel sound in the word "head" as spoken by J. W. The zero crossing frequency increases when the instantaneous frequency has large negative pulses. The average value of the instantaneous frequency during the time interval between adjacent zero crossings may then be zero or -2π. If the average value is zero the frequency of zero-crossing is the reciprocal of the period between zero crossings. If this period is less than the first formant period the frequency of zero crossings will increase.

An attempt was made to plot the zero crossing frequency of ordinary speech against the zero crossing frequency of differentiated speech. It was hoped that this would display a characteristic waveform for each pitch period. However, in light of the above discussion, it is apparent that this would not display a stable pattern. By differentiating speech, one enhances the higher frequencies. This compounds the stability problems by making the first two formant frequencies approximately equal in amplitude. The average value of the instantaneous frequency and hence the zero crossing frequency will then frequently jump back and forth between its two levels. This is caused by the formant

bandwidth. The time of each jump in the pitch period and the time interval between jumps will vary considerably for each pitch period. This is caused by the variation in the glottal pulse waveform. These instabilities will mask any characteristic waveform that might be present in the frequency of zero crossings.

The instantaneous phase can be used as a sweep voltage in order to observe the instantaneous frequency and the speech waveform. Figure 12a shows the instantaneous frequency on the vertical axis plotted against the instantaneous phase on the horizontal axis in the left trace and against time on the horizontal axis in the right trace. The top pair of traces is the instantaneous frequency for the vowel sound in "head" and the bottom pair of traces is the instantaneous frequency for the vowel sound in the word "hood". Each pair of traces was recorded simultaneously and so they correspond to identical instantaneous frequencies.

On the basis of the previous discussion it can be seen that the instability of the instantaneous frequency does not occur in the time axis of the waveform. The periodicity of the scallops is determined by the first two formant frequencies and are relatively independent of the formant bandwidth and variation in the glottal pulse waveform. Variations in the amplitude of the instantaneous frequency will destroy the strict periodicity. However the scalloped

effect will prevail with the proper regularity in the time axis. The instantaneous phase will destroy the regularity even if the average value of the instantaneous frequency remains constant. This can be seen in the left traces in Figure 12a.

Figure 12b shows the speech waveform plotted against the instantaneous phase on the left and against time on the right. The top pair of traces is the vowel sound in "head" and the bottom pair of traces is the vowel sound in "hood". The problems of this approach are similar to those discussed above for the instantaneous frequency.

First Formant Harmonic Number Determination of the Vowels

When cos $\emptyset(t)$ is plotted against $\emptyset$ the result should be a sinusoidal waveform. The sinusoidal waveform will go through a complete cycle every time $\emptyset$ increases by 2π. If $\emptyset'$ has a constant short term average value it will be an interger multiple, n, of the pitch frequency. The average slope of the instantaneous phase curve will then be n times the pitch frequency. The interger multiple, n, will equal the harmonic number of the instantaneous frequency average value. In one pitch period the instantaneous phase will change by $2n\pi$ and the waveform cos $\emptyset$ vs. $\emptyset$ will go through n cycles. This device presents an easy method of measuring the harmonic number of a speech formant. Care is needed to

maintain a constant instantaneous frequency average value. This may be done by filtering the unwanted formant frequencies.

Table 1 shows the first formant harmonic numbers for the ten vowel sounds shown in Figure 9. The harmonic numbers were measured using the method discussed above. If one lowers the pitch of the vowel sound the harmonic number will jump to the next highest number since the formant frequency remains relatively fixed.

Table 1. First formant harmonic number of several vowel sounds as spoken by J. W.

Vowel sound as found in the word	Harmonic number pitch frequency $\simeq$ 100 cps
Heed	2
Hid	3
Head	3
Had	4
Hod	4
Hawed	4
Hood	3
Who'd	2
Hud	4
Heard	3

Time Domain Waveforms of the Consonants

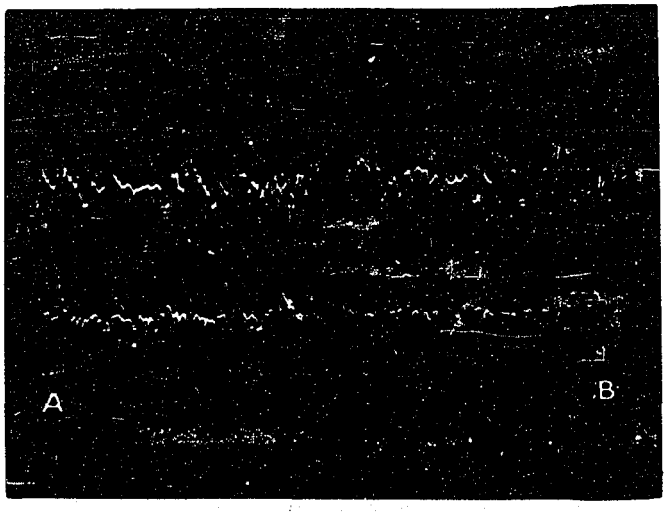
Figure 13 shows the speech waveform of two nasal sounds in the upper traces and the corresponding instantaneous frequencies in the lower traces. Figure 13 a shows the sound /m/ and Figure 13b shows the sound /n/. These sounds are generated in the same way that the vowel sounds are generated except that the nasal cavity is also used. The comments made on the speech waveform and instantaneous frequency of the vowel sounds also apply to the nasals.

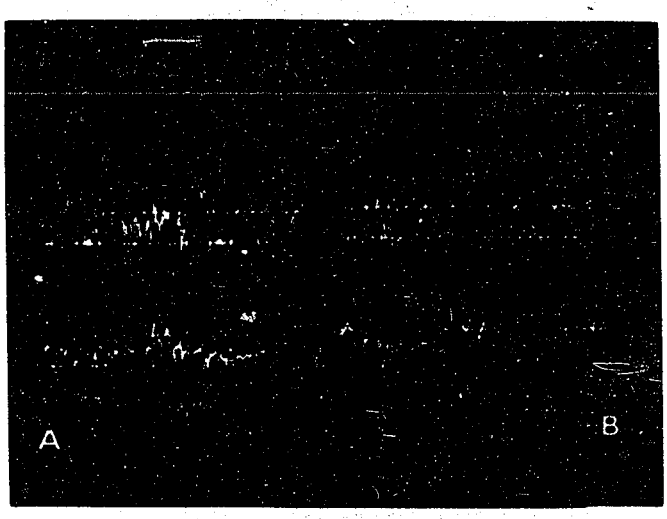
Figure 14 shows the speech waveform of four voiced fricatives in the upper traces of each picture and the corresponding instantaneous frequency in the lower traces. The fricative voiced sounds found in the words "this", "voice", "zoo", and "pleasure" were used in Figures 14a through d respectively. These sounds contain a large noise component in addition to the voiced quasi-periodic component. This can be seen in the section of each picture that contains a double trace.

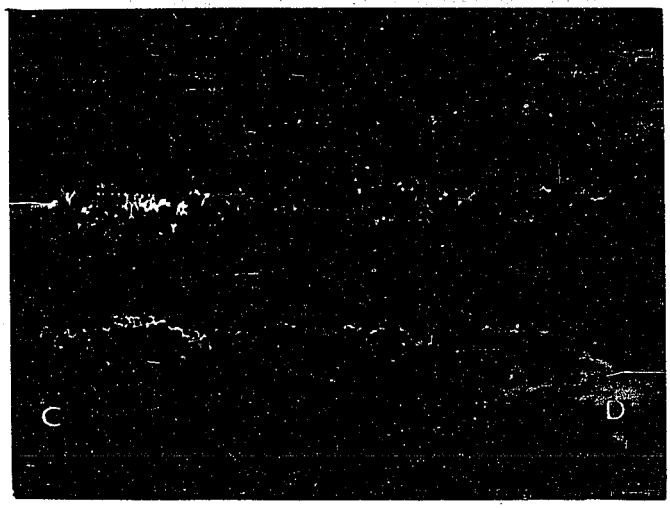
Nonperiodic sounds such as the voiceless fricatives and the stop consonants have also been studied using the methods described above. Since there is no periodicity the instantaneous frequency must be displayed at a syllabic rate rather than at a pitch synchronous rate as was done for the voiced sounds. The amplitude modulated component of the speech waveform was also investigated. In the case of

Figure 13. Speech waveform, top trace, and instantaneous frequency, bottom trace, of the nasal sounds a) /m/ and b) /n/. Spoken by J. W.

Figure 14. Speech waveform, top trace, and instantaneous frequency, bottom trace, of the voiced fricative sounds in the words a) "this", b) "voice", c) "zoo", and d) "pleasure".


A
B


A
B


C
D

voiced sounds the amplitude modulation was a relatively low-frequency signal and did not appear to have any distinguishing characteristics. The sounds produced by the amplitude modulation were unintelligible. The stop consonants have an initial plosive followed by a fricative sound. The low-frequency component of the stop consonant amplitude modulation follows the transient energy of the plosive sound. None of the amplitude modulation signals or the stop consonant and fricative instantaneous frequencies appeared to contain any readily available useful information. Therefore this avenue was not pursued further.

Pitch Synchronously Interrupted Speech

The effects of interrupted speech at a pitch synchronous rate were also investigated. The speech sounds could be interrupted during any part of the pitch period and the length of interruption could be any fraction of the pitch period. Informal listening tests were performed by recording interrupted consonant-vowel-consonant words. The words used were those used by Peterson and Barney (46) and also used in the instantaneous frequency studies of this report. The words are "heed", "hid", "had", "head", "hod", "hawed", "hood", "who'd", "hud", and "heard". The position of the interruption within the pitch period had no discernible effect on the intelligibility of the speech. It did not even appear to have any discernible effect on the quality of the

speech. This would imply that the frequency spectrum of small segments of speech is relatively invariant under the position of the segments within the pitch period. This assumes, of course, that the ear performs some sort of frequency analysis on the speech sounds.

The above finding supports the work of Pinson (48) who measured the formant bandwidth by curve fitting a segment of the speech waveform with damped sinusoids. The segment of the speech waveform was chosen as that part of the pitch period for which the glottis is closed. Since the ear cannot distinguish between two segments of the pitch period, the bandwidth during the glottis-open phase must be approximately equal to the bandwidth during the glottis-closed phase. Barring any numerical problems, Pinson's results should be comparable to the bandwidths obtained by measurements on the entire pitch period. These results should be comparable at least as compared with the ear's ability to detect small differences in bandwidth.

Miller and Licklider (40) found that with 50 percent on-time, the intelligibility of interrupted speech is only slightly below the uninterrupted value. The intelligibility deteriorates as the percent on-time is decreased. This same result was found for pitch-synchronous interruptions. The intelligibility was still quite good for 33 percent on-time but it was quite poor for 25 percent on-time.

CONCLUSIONS

Several time domain operations on the speech waveform have been studied. The speech waveform, the instantaneous frequency and the rate of zero crossings all appear to have prominent characteristics that are not primary in the identification of the sound.

The formant bandwidths change the instantaneous frequency average value and amplitude-modulate the instantaneous frequency waveform. The formant bandwidths amplitude-modulate their respective formant frequencies. The changing amplitudes of the formant frequencies cause the strongest formant frequency to switch back and forth between the first two formant frequencies. This in turn causes the average value of the instantaneous frequency to jump back and forth between the two formant frequencies. The jumps in the instantaneous frequency average value destroy the formant tracking properties of the instantaneous frequency long term average value. A possible solution is to use the proper pre-emphasis and filtering to suppress the undesired formant frequency. This would not solve the problem but would minimize it.

Variation in the glottal pulse waveform destroys the pitch periodic nature of the time domain waveforms. The instantaneous frequency and the zero crossing frequency waveforms are very sensitive to variations in the glottal

pulse waveform. The average value as well as the amplitude of the frequency waveforms will change radically with the glottal pulse shape.

Through the use of vocal tract analogs it has been shown that the formant bandwidth and glottal pulse waveform are relatively independent of speech intelligibility (59). The short term frequency spectrum on the other hand presents the formant frequencies as local spectral maxima. These maxima are independent of the bandwidth and the glottal pulse variations provided the formant frequencies are spaced further apart than their respective bandwidths.

The short term autocorrelation function appears to overcome the difficulties involved in the other time domain analyses presented in this paper. After the aperiodicities of the voiced speech waveform are removed by the autocorrelation function, the instantaneous frequency and zero crossing frequency analyses could again be attempted with perhaps better results. The autocorrelation will only smooth out the effects of the glottal pulse waveform variation. The effects of the finite formant bandwidths will still be present. Autocorrelation could also be used to some advantage on the voiced fricatives. The correlation process will remove the noise component from the speech sound assuming that the noise is uncorrelated. The resulting periodic component can then be analyzed. A description of the noise component must

also be found to give a complete description of the speech sound. Additional efforts on this particular approach may be warranted.

Informal listening tests performed on pitch-synchronously interrupted speech indicate that the phase of the interruption relative to the pitch period does not effect the word intelligibility. This indicates that any time domain analysis of the speech sound should be equally effective on any segment of the speech waveform. This should at least be true provided the segment of the speech waveform is greater than 50 percent of the pitch period. The instantaneous frequency would meet this requirement if it were not contaminated by the effects of the formant bandwidth.

The information bearing elements of speech are contained in the time domain analyses. However the information bearing elements do not appear, in the time domain waveforms studied, in forms that allow them to be easily separated from the irrelevant structures. Analysis of speech sounds in the frequency domain present the information bearing elements in a form that is largely independent of the irrelevant structure. The results of the investigation tend to show that a machine-recognition system based on the time domain analysis of speech sounds is probably impractical since the information is much more readily identifiable in the frequency domain.

BIBLIOGRAPHY

1. Ahmend, Rais and Fatehchand, Richard. Effect of sample duration on the articulation of sounds in normal and clipped speech. Acoustical Society of America J. 31: 1022-1029. 1959.

2. Bell, C. G., Fujisaki, J. M., Heinz, J. M., Stevens, K. N., and House, A. S. Reduction of speech spectra by analysis-by-synthesis techniques. Acoustical Society of America J. 33: 1725-1736. 1961.

3. Chang, S. H., Pihl, G. E., and Wiren, J. The intervalgram as a visual representation of speech sounds. Acoustical Society of America J. 23: 675-679. 1951.

4. Cherry, E. Colin and Phillips, V. J. Some possible uses of single side band signals in formant-tracking systems. Acoustical Society of America J. 33: 1067-1077. 1961.

5. David, E. E. Naturalness and distortion in speech-making devices. Acoustical Society of America J. 28: 586-589. 1956.

6. David, E. E., Schroeder, M. R., Logan, B. F., and Prestigiacomo, A. J. Voice excited vocoders for practical speech reduction. Institute of Radio Engineers Professional Group on Information Theory Trans. 8: 101-105. 1962.

7. Davis, K. H., Biddulph, R. G., and Balashek, S. Automatic recognition of spoken digits. Acoustical Society of America J. 24: 637-642. 1952.

8. Delattre, Pierre C., Liberman, Alvin M., and Cooper, Franklin S. Acoustic loci and transitional cues for consonants. Acoustical Society of America J. 27: 769-773. 1955.

9. Denes, P. and Mathews, M. V. Spoken digit recognition using time-frequency pattern matching. Acoustical Society of America J. 32: 1450-1458. 1960.

10. Dersch, William C. A decision logic for speech recognition. Bionics Symposium Proc. 1960: 287-306. 1960.

11. Dudley, Homer. Phonetic pattern recognition vocoder for narrow band speech transmission. Acoustical Society of America J. 30: 733-739. 1958.

12. Dudley, Homer. Remaking speech. Acoustical Society of America J. 11: 169-177. 1939.

13. Dudley, Homer and Balashek, S. Automatic recognition of phonetic patterns in speech. Acoustical Society of America J. 30: 721-732. 1958.

14. Dukes, J. M. C. The effect of severe amplitude limitation on certain types of random signal: a clue to the intelligibility of infinitely clipped speech. Institution of Electrical Engineers Proc. 103, part C: 88-97. 1955.

15. Dunn, H. K. Methods of measuring vowel formant bandwidths. Acoustical Society of America J. 33: 1737-1746. 1961.

16. Flanagan, James L. Automatic extraction of formant frequencies from continuous speech. Acoustical Society of America J. 28: 110-117. 1956.

17. Flanagan, James L. Bandwidth and channel capacity necessary to transmit the formant information of speech. Acoustical Society of America J. 28: 592-596. 1956.

18. Flanagan, James L. Evaluation of two formant-extraction devices. Acoustical Society of America J. 28: 118-125. 1956.

19. Forgie, James W. and Forgie, Carma D. Results obtained from a vowel recognition computer program. Acoustical Society of America J. 31: 1480-1489. 1959.

20. Foulkes, J. D. Computer identification of vowel types. Acoustical Society of America J. 33: 7-11. 1961.

21. Halle, M., Hughes, G. W., and Radley, J. P. A. Acoustic properties of stop consonants. Acoustical Society of America J. 29: 107-116. 1957.

22. Halle, M. and Stevens, K. Speech recognition: a model and a program for research. Institute of Radio Engineers Professional Group on Information Theory Trans. 8: 155-159. 1962.

23. Harris, Katherine S., Hoffman, Howard S., Liberman, Alvin M., Delattre, Pierre C., and Cooper, Franklin S. Effect of third-formant transitions on the perception of the voiced stop consonants. Acoustical Society of America J. 30: 122-126. 1958.

24. Heinz, John M. and Stevens, Kenneth N. On the properties of voiceless fricative consonants. Acoustical Society of America J. 33: 589-596. 1961.

25. House, A. S. Analog studies of nasal consonants. J. Speech and Hearing Disorders. 22: 190-204. 1957.

26. House, A. S. and Stevens, K. N. Estimation of formant bandwidths from measurements of transient response of the vocal tract. J. Speech and Hearing Research 1: 309-315. 1958.

27. Howard, Calvin R. Speech analysis-synthesis scheme using continuous parameters. J. Acoustical Society of America 28: 1091-1098. 1956.

28. Hughes, G. W. and Halle, M. On the recognition of speech by machine. International Conf. on Information Processing Proc. 1959: 252-256. 1959.

29. Hughes, George W. and Halle, Morris. Spectral properties of fricative consonants. Acoustical Society of America J. 28: 303-310. 1956.

30. Koenig, W. A new frequency scale for acoustic measurements. Bell Labs Record 27: 299-301. 1949.

31. Koenig, W., Dunn, H. K., and Lacy, L. Y. Sound spectrograph. Acoustical Society of America J. 18: 19-49. 1946.

32. Koshikawa, T. and Sugimoto, T. The information rate of the pitch signal in speech. Institute of Radio Engineers Professional Group on Information Theory Trans. 8: 92-100. 1962.

33. Liberman, A. M., Ingemann, Frances, Lisker, Leigh, Delattre, Pierre, and Cooper, F. S. Minimal rules for synthesizing speech. J. Acoustical Society of America 31: 1490-1499. 1959.

34. Licklider, J. C. R. The intelligibility of amplitude-dichotomized, time-quantized speech waves. Acoustical Society of America J. 22: 820-823. 1950.

35. Licklider, J. C. R. and Pollack, Irwin. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. Acoustical Society of America J. 20: 42-51. 1948.

36. Lehiste, Ilse and Peterson, Gordon E. Transitions, glides, and diphthongs. Acoustical Society of America J. 33: 268-277. 1961.

37. Manley, H. J. Spectral speech signal representations. Unpublished paper presented at Northeast Electronics Research and Engineering Meeting, Boston, Mass., Nov. 1962. Multilithed. Waltham, Mass., Sylvania Electric Products, Inc. ca. 1962.

38. Marcou, P. and Daguet, J. New methods of speech transmission. In Cherry, C., ed. Information theory. pp. 231-244. New York, N. Y., Academic Press, Inc. 1956.

39. Mathews, M. V., Miller, Joan E., and David, E. E., Jr. Pitch synchronous analysis of voiced sounds. Acoustical Society of America J. 33: 179-186. 1961.

40. Miller, George A. and Licklider, J. C. R. The intelligibility of interrupted speech. Acoustical Society of America J. 22: 167-173. 1950.

41. Miller, R. L. Nature of vocal cord wave. Acoustical Society of America J. 31: 667-677. 1959.

42. Nakata, Kazuo. Synthesis and perception of nasal consonants. Acoustical Society of America J. 31: 661-666. 1959.

43. Oleson, Harry F. and Belar, Herbert. Phonetic typewriter. Acoustical Society of America J. 28: 1072-1081. 1956.

44. Olson, Harry F. and Belar, Herbert. Phonetic typewriter. III. Acoustical Society of America J. 33: 1610-1616. 1961.

45. Peterson, E. Frequency detection and speech formants. Acoustical Society of America J. 23: 668-674. 1951.

46. Peterson, Gordon E. and Barney, Harold L. Control methods used in a study of the vowels. Acoustical Society of America J. 24: 175-184. 1952.

47. Pierce, John R. and David, Edward E., Jr. Man's world of sound. Garden City, N. Y., Doubleday and Company, Inc. 1958.

48. Pinson, Elliot N. Pitch synchronous time domain estimation of formant frequencies and bandwidths. Unpublished paper presented at Sixty-third Meeting Acoustical Society of America, New York, N. Y., May, 1962. Mimeo. Murray Hill, New Jersey, Bell Telephone Labs., Inc. ca. 1962.

49. Pollack, Irwin and Pickett, J. M. Intelligibility of peak clipped speech at high noise levels. Acoustical Society of America J. 31: 14-16. 1959.

50. Potter, Ralph K., Kopp, George A., and Green, Harriet C. Visible speech. New York, N. Y., D. Van Nostrand Co., Inc. 1947.

51. Sakai, T. and Inoue, S. New instruments and methods for speech analysis. Acoustical Society of America J. 32: 441-450. 1960.

52a. Schauer, Ralph Floyd. Very low frequency characteristics of speech. Unpublished Ph. D. thesis. Ames, Iowa, Library, Iowa State University of Science and Technology. 1960.

52b. Schroeder, M. R. and Atal, B. S. Generalized short time power spectra and autocorrelation functions. Acoustical Society of America J. 34: 1679-1683. 1962.

53. Smith, J. E. Keith and Klem, Laura. Vowel recognition using a multiple discriminate function. Acoustical Society of America J. 33: 358. 1961.

54. Spogen, L. R., Shaver, A. N., Baker, D. E., and Blom, B. V. Speech processing by the selective amplitude sampling systems. Acoustical Society of America J. 32: 1621-1625. 1960.

55. Stevens, Kenneth N. Toward a model for speech recognition. Acoustical Societyof America J. 32: 47-55. 1960.

56. Uhr, Leonard and Vossler, Charles. A pattern recognition program that generates, evaluates, and adjusts its own operators. Western Joint Computer Conf. Proc. 19: 555-569. 1961.

57. Uhr, Leonard and Vossler, Charles. Recognition of speech by a computer program that was written to stimulate a model for human visual pattern recognition. Acoustical Society of America J. 33: 1426. 1961.

58. Vitale, Anthony. Cheap and easy s.s.b. QST; devoted entirely to amature radio. 40: 16-20. 1956.

59. Weibel, E. S. Vowel synthesis by means of resonant circuits. Acoustical Society of America J. 27: 858-865. 1955.

60. Weibel, E. S. On Webster's horn equation. Acoustical Society of America J. 27: 726-727. 1955.

61. Welch, Peter D. and Wimpress, Richard S. Two multivariate statistical computer programs and their application to the vowel recognition problem. Acoustical Society of America J. 33: 426-434. 1961.

62. Wiren, Jacob and Stubbs, Harold L. Electronic binary selection system for phoneme classification. Acoustical Society of America J. 28: 1082-1092. 1956.

ACKNOWLEDGMENTS

The author wishes to express his sincere thanks to Dr. V. W. Bolie for his very helpful encouragement and advice.

APPENDIX

A single-side-band modulated signal may be generated from two balanced modulated signals. The modulating signal is

$$S(t) = \sum_{n=0}^{\infty} a_n \cos (n \omega_0 t + \theta_n) \qquad 24)$$

where the signal is periodic in ω_0. The single-side-band modulated signal is then

$$S_M(t) = \sum_{n=0}^{\infty} a_n \cos \left[(\omega_c \pm n \omega_0)t \pm \theta_n \right] \qquad 25)$$

where ω_c is the carrier frequency. The plus signs in the expression for the argument correspond to the upper side-band and the minus signs correspond to the lower side-band. Equation 25 may be written

$$\begin{aligned} S_M(t) &= \left[\sum_{n=0}^{\infty} a_n \cos (n \omega_0 t + \theta_n) \right] \cos \omega_c t \\ &\mp \left[\sum_{n=0}^{\infty} a_n \sin (n \omega_0 t + \theta_n) \right] \sin \omega_c t \end{aligned} \qquad 26)$$

with the use of a trigonometric identity. Equation 26 is the sum of two balanced modulated signals. Both the carrier and modulating signals of one of the balanced modulated signals are phase shifted -90 degrees from their normal positions. The minus sign in Equation 26 corresponds to the upper side-band and the plus sign corresponds to the lower side-band.