MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

62 G - 1

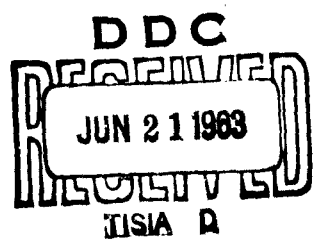
VOCODED SPEECH

Bernard Gold

7 May 1963

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the joint support of the U. S. Army, Navy and Air Force under Air Force Contract AF 19(628)-500.

LEXINGTON MASSACHUSETTS



ABSTRACT

This report constitutes an introduction to vocoders and, in particular to the vocoder built by Group 62. It is based on a lecture recently delivered by the author. An elementary discussion of speech fundamentals followed by a brief description of the different branches of speech research work somewhat prepares the reader for the later explanations of channel vocoders, voice-excited vocoders and, finally, the Group 62 vocoder.

Introduction

The vocoder, invented by Homer Dudley in about 1939, was the first system to achieve bandwidth compression of speech. Since then, many other schemes have been proposed and tried but none thus far has been as worthwhile as the vocoder.

Group 62 has been actively engaged in the design and construction of a vocoder. This report traces the history of Group 62's effort, its motivations, and some of the interesting problems encountered. In order for the reader to better grasp the flavor of these problems, it was thought desirable to include some preliminary discussion of the speech process and of vocoders in general.

Some Facts About Speech

The production, perception and organization of speech are complicated phenomena. In order to deal with these phenomena in a practical way, it is necessary to begin with a set of working assumptions which are approximations to the actual physical processes.

First we discuss speech production. We know that air supplied by the lungs causes the vocal cords to vibrate, that the throat, mouth and nose (the vocal tract) form a sort of resonance chamber, and that we have the ability to change the pitch of our voices by relaxing or tightening our vocal cords (or to continue the flow of air with no vocal cord vibration) and the ability to change the sounds we create by altering the shape of our vocal tract. This knowledge we have succeeded in condensing into a very useful electrical model, shown in Fig. 1.

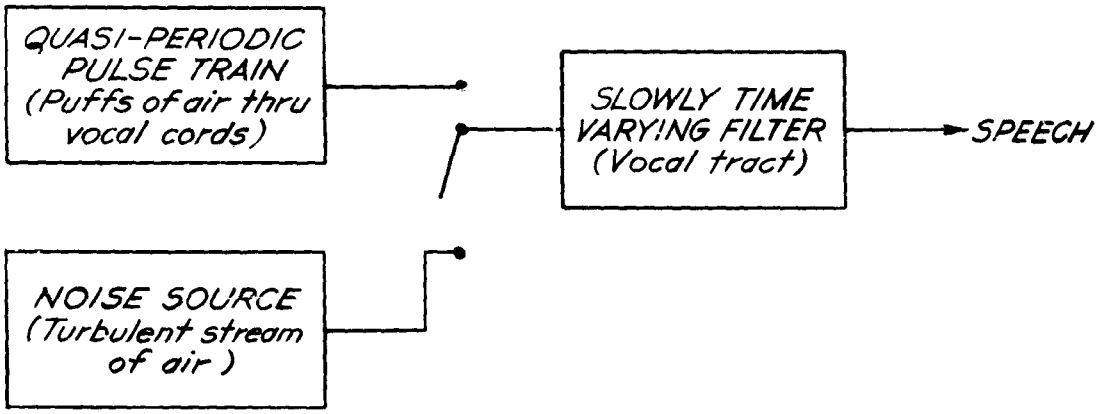


Fig. 1. Model of Speech Production

The model consists of two sources, only one of which is connected at any moment to the filter (vocal tract). The passage of either of these sources through the filter produces speech. The parameters of the filter determine the specific sound, the switch determines whether the sound is to be voiced or unvoiced, and the fundamental frequency of the pulse train determines the pitch. The parameters of the filter cannot vary too rapidly with time simply because we cannot physically change the shape of our vocal tracts too rapidly.

From this model, we deduce that, during the voiced portions of a spoken sentence, the speech is a quasi-periodic waveform wherein both the positions on the frequency scale and amplitudes and phases of the harmonics are slowly changing. By slowly, we mean that for most sounds, significant changes in the harmonic structure will not occur in fewer than 40 milliseconds. During the unvoiced portions of the sentence, the speech no longer has a periodic structure but resembles noise.

How are the perceived sounds organized in our brain? How is speech defined and categorized? We know that speech sounds can be organized into units; phonemes, syllables, words, sentences, etc. The point we want to stress is that our acceptances of speech as being organizable into these units is a linguistic assumption as to the nature of speech. It is thus an assumption as to the ability of the human brain to decipher speech sounds. Surely, any set of physical measurements of the same speech unit will never be exactly the same the next time it is spoken. Thus, acoustically, there are, strictly speaking, an infinite number of phonemes. However, we accept the assumption that there are only about 40 different phonemes in the English language. This assumption puts a very stiff constraint on the number of possible states of the model of Fig. 1. We will see later that the shape of the vocal tract further constrains the filter of Fig. 1 and allows us to conceive of surprisingly simple types of speech synthesizers.

For convenience, we will list most of the phonemes. This is definitely not an authorized list, which you can find in a dictionary. They are listed here for future reference.

Table I

1. Vowels:

   beet, bit, bet, bat, bot, but, boot

2. Semi-vowels:

   want, run, hill, you

3. Voiceless fricatives

   see, she, fee, thin

4. Voiceless plosives

   pin, kin, tin

5. Voiced fricatives

   zip, azure, give, then

6. Voiced plosives

   bin, gun, din

7. Aspiration

   heaven

8. Nasals

   nose, mouth, sing

Notice that only 29 phonemes are listed. There are others, but these will suffice for this discussion.

How is speech perceived? More specifically, how does the human ear process the speech before presentation to the brain? There is one empirical fact out of this enormously complicated subject which is of prime use to the speech researcher. The fact is this: the ear is much less sensitive to the phase spectrum of the speech than it is to the amplitude spectrum. This is true to such an extent that almost all the work done in the speech field uses Ohm's law (the insensitivity of the ear to phase) as a working assumption. The graphic illustration of Ohm's law is the perception of the two waveforms shown in Fig. 2 as the same sound. The wave of Fig. 2b can have the same amplitude spectrum as that

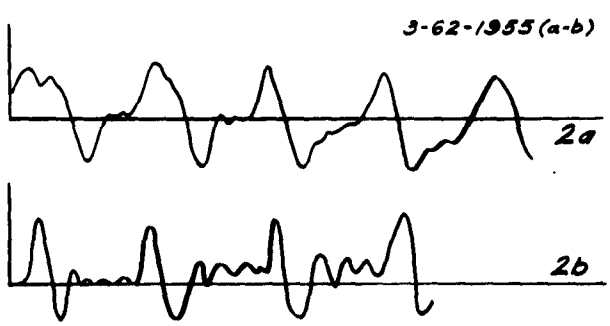


Fig. 2. Illustration of Waveform Changes Caused by Phase Shift

of Fig. 2a; its peakedness is caused by the fact that the phases of the harmonic components are more nearly alike. These two waves are very nearly indistinguishable by ear.

This insensitivity of the ear to phase is quite reasonable if one considers that speech sounds almost always reach the hearer's ears after multiple reverberations, so that the phase of these sounds depends greatly on the physical surroundings.

The above three assumptions can be distilled into a single definition of speech in spectral terms. At any instant of time, the immediate past history of a voiced portion of speech can be described by means of an amplitude spectrum, such as shown in Fig. 3. An unvoiced portion of speech can be described by an amplitude spectrum such as shown in Fig. 4.

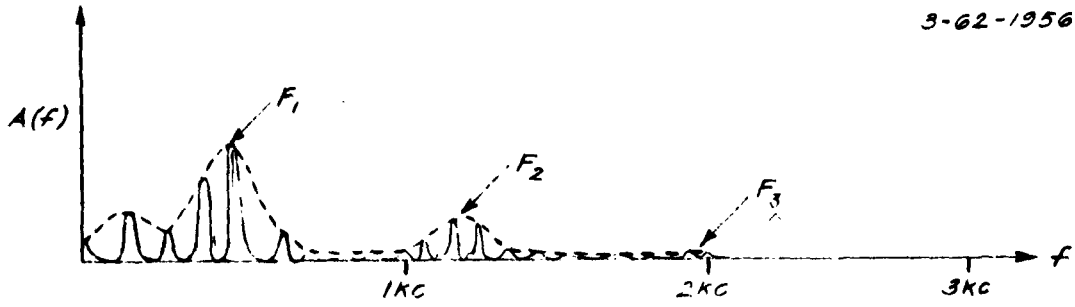


Fig. 3. Spectrum of Voiced Speech

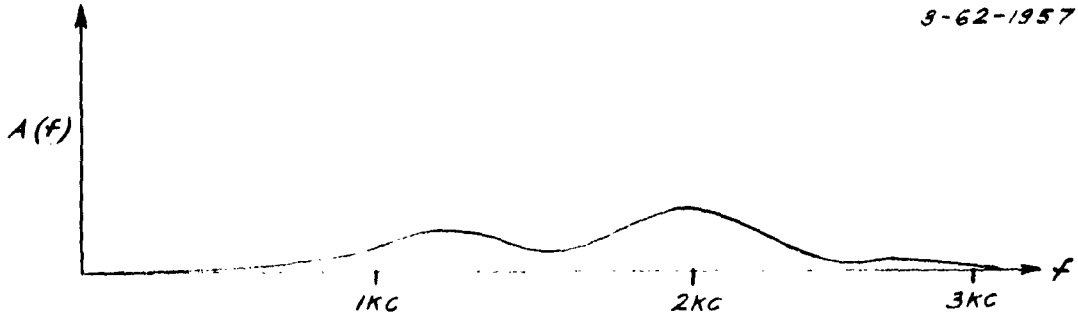


Fig. 4. Spectrum of Voiceless Speech

In Fig. 3, the solid line is the actual spectrum and the dotted line is the spectral envelope. The spacing between adjacent peaks of the spectrum is the fundamental frequency, or pitch; the spectral envelope is, roughly speaking, the spectrum of the filter of Fig. 1 and thus determines the phoneme. In Fig. 4, the excitation is noise; thus the spectrum is continuous.

Referring to Table 1, we can say that the vowels, semi-vowels and nasals are described by Fig. 3; the aspiration, voiceless plosives and voiceless fricatives are described via Fig. 4. The voiced fricatives and voiced plosives force us to modify our assumption somewhat. These sounds contain both voiced and frictional energy. It is as if, in the model of Fig. 1, the switch were to be replaced by two variable attenuaters following each of the two sources, both attenuaters feeding a common summing circuit whose output supplied energy to the vocal tract.

One question that could legitimately be raised concerns the use of the word spectrum. In transform methods of systems analysis, the word spectrum has an unequivocal mathematical definition and, in particular, is strictly defined only for waveforms lasting an infinite time. Since we are dealing here with notions such as the spectrum of a phoneme, which lasts, perhaps, 40 msec, we are on quite shaky ground. The most satisfying working definition of a speech spectrum appears to be this: imagine a bank of band-pass filters covering the audio band. Each filter is about 25 cycles wide. The spacing between adjacent filters is as close as you like. The closer together are your filters, the more continuous will be your definition of spectrum. Each filter output is rectified and then passed through a low pass filter, also about 25 cycles wide. At any instant of time, the set of low pass filter outputs constitutes a spectral measurement.

Subdivisions of Speech Research

The title of this report is vocoded speech. How do vocoders fit into the larger picture of speech research and development work?

Speech is one of the recognized topics in the broader field of acoustics. The Journal of Acoustical Society of America (JASA) is, in this country, the leading journal of acoustics and most of the speech work in the U.S.A. gets published or presented via J.A.S.A. This speech work is roughly divided into several categories:

1. Speech analysis and automatic recognition
2. Synthesis
3. Psychoacoustics
4. Speech communications and band compression

It is item 4 that we are mainly talking about but the fundamentals we have already discussed bear on all 4 items, which, are moreover, tightly connected to one another. It is worth a small digression to discuss items 1, 2 and 3 before getting into the vocoder problem.

Speech analysis is concerned primarily with the relation of the particular phoneme to its spectral description. The most important concept in speech analysis has been that of the formants, or resonances of the vocal tract. These (F_1, F_2 and F_3) are indicated in Fig. 3. According to this concept, the position, in frequency, of the formants of a given vowel are the primary cues for the identification of that vowel. Now, vowels are extremely important sounds, which, in many ways dominate the other classes of phonemes. Nearly every phoneme in connected speech is either a vowel or is adjacent to a vowel. Furthermore, the phonemes in connected speech actually run into one another. The vocal tract shape obviously does not change discontinuously. Thus, it turns out that analysis of the formant motion

during a vowel can lead to cues for identification of the phoneme adjacent to that vowel. The formant description of speech is therefore fairly general and that, combined with its great simplicity, has made the formant concept a cornerstone of much of the work in all the branches of speech. It has surely been the starting point for all work on automatic speech recognition and for much of the work on speech synthesis and has strongly affected the band compression field via the so called "formant tracking vocoders". (More about them later.)

A word about automatic speech recognition: no one has yet succeeded in building a reliable and practical speech recognizer. Speech is a complicated and variable phenomenon and in order to make any headway, one is forced into quite sophisticated statistical techniques. This leads to a very long data processing problem, since each of the 40 sounds must be analyzed in great detail and under many different conditions (different speakers, different contexts, etc.). Thus, the field is still in a very "researchy" stage and one hopes that soon some very clearcut practical results will be attained so that greater impetus is given to the research work. Probably the most notable work being done now is that under the direction of James Forgie of Lincoln Laboratory.

Perhaps the most naturally interesting of all the speech fields is that of synthesis. In fact, it is probably true that all of the speech work presently being done is a descendant of man's natural curiosity about "talking machines". References to this curiosity dating back many hundreds of years can be found. Dudley, who invented the vocoder, derived his main inspiration from his work on the Voder, which was a talking machine displayed at the 1939 New York Worlds' Fair. A vocoder, as we shall see later, is a special kind of speech analyzer-synthesizer.

Despite its long history, synthesis as a practical art does not yet exist. Many synthesizers have been built. Those that try to synthesize speech beginning from a set of a priori rules can be made to sound intelligible but not natural. Just as we lack the knowledge to write down specific rules for a recognition, so do we lack the knowledge to write down specific synthesis rules which will result in speech with a natural quality.

Psychoacoustic testing pertains to the speech perception capabilities of humans. For example, the human ear can detect very small differences in the frequencies of two sine waves. Measurements have been made which describe how sensitive the ear is to spectral fluctuations, how good a speech recognizer a human is when no context information is available etc. The importance and generality of this type of work is two fold: one, it gives us useful information about human abilities and two, it establishes a basis for all speech research wherein the human is the ultimate receiver (such as the field of speech communication).

The Vocoder

A good deal of work in the speech communications field has centered around the band compression problem. The reason is simple. Thus far, all practical long distance communications channels are sufficiently limited in channel capacity, so that severe limitations are placed on the number of speech messages that can be simultaneously transmitted. Whether or not this situation will be modified by the advent of satellite communications links is not clear. One thing is clear; interest in the band compression field has noticeably increased during the past few

years. It is also true that this work has focused on the development of two particular schemes 1) the channel vocoder and 2) the formant tracker. We shall describe these now and also say a few words about other band compression schemes.

The classical vocoder of Dudley consists of a spectrum analyzer, a buzz-hiss detector, a pitch detector and a synthesizer. A sketch is shown in Fig. 5.

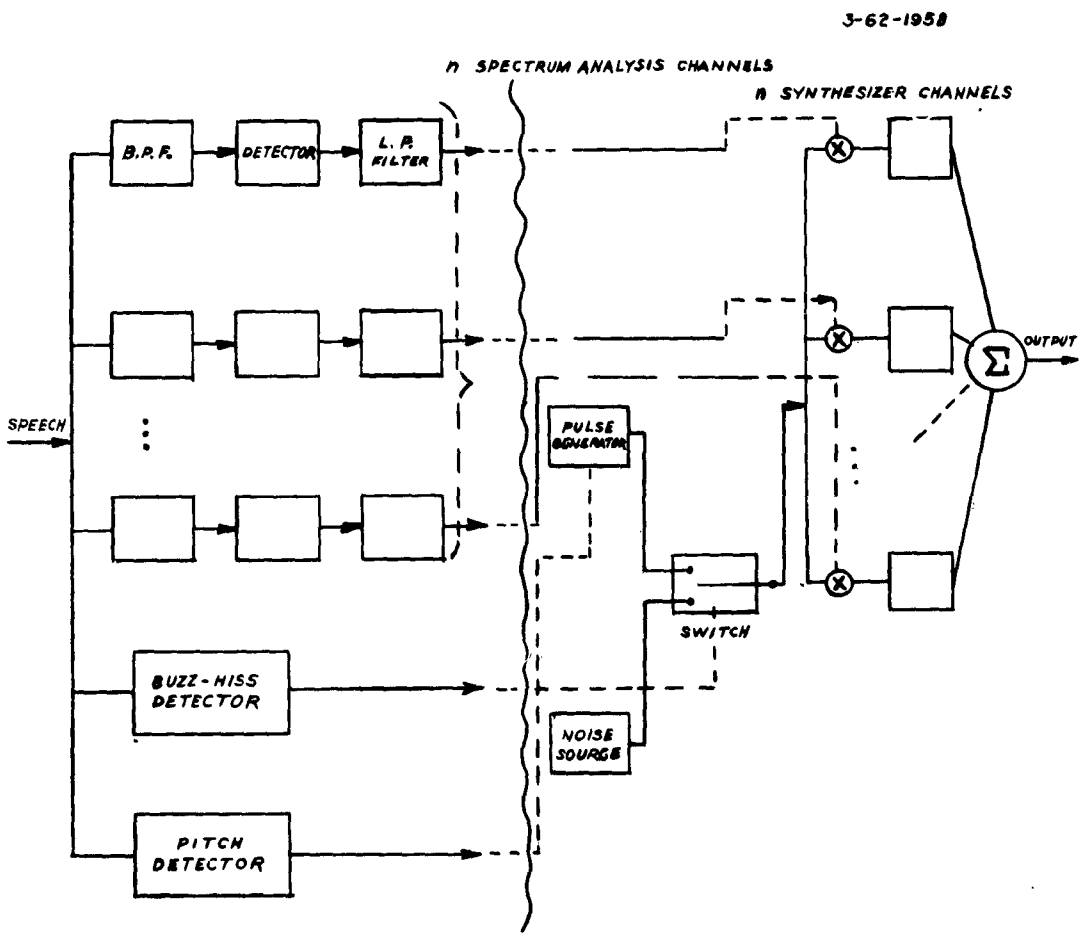


Fig. 5. Classical Channel Vocoder

The spectrum analyzer consists of n channels; each channel consists of a differently tuned band pass filter with its associated detector and low-pass filter; the band pass filters are contiguously spaced to cover the audio spectrum. Vocoders have been built with as many as forty and as few as twelve spectrum channels. The output of the buzz-hiss detector is a binary decision signal describing whether the excitation into the vocal tract is voiced or unvoiced. The pitch detector produces a voltage which should be directly proportional to the fundamental frequency of the voice at the analyzer, the buzz-hiss switch determines whether the input is to consist of quasi-periodic pulses or random noise. The pulse generator is imagined to be voltage controlled with frequency proportional to the voltage generated by the pitch detector.

Notice how the vocoder synthesizer relates to the model of Fig. 1. The source, or excitation is either a buzz(voicing) or a hiss (non-voicing). Table 1 shows that the vowels, semi-vowels and nasals are buzz, the voiceless fricatives, voiceless plosives and the aspirations are hiss. The slowly varying vocal tract is simulated by the bank of synthesizer filters and modulators. If, to take an ideal example, the vocal tract exhibited resonances at 400 and 800 cycles, the output would show roughly the same resonances. The fine structure of the spectrum shown in Fig. 3 is reproduced via the pitch controlled pulse generator.

Notice also that the vocoder has completely thrown away all phase information. Neither the pitch detector nor the spectrum analyzer preserve phase. Thus, if all else were neglected, a vocoder is, in principle, a system which compresses speech bandwidth by 2:1 by throwing away half of the information in the speech wave. According to our basic assumptions, the thrown-away half is possibly that half which is almost ignored by the ear.

But a vocoder compresses speech bandwidth by more than 2:1. The key to this fact lies in the low pass filters in the analyzer. It has been empirically determined that the energy content of any single band pass filter in the analyzer does not change at a rate faster than that corresponding to a 25 cps bandwidth. Thus, a 25 cps low pass filter is sufficiently wide to pass all spectral energy fluctuations in the speech. If we assume that 20 channels are sufficient, the bandwidth needed to transmit all spectral information is 500 cps. In a classical vocoder such as we are describing, pitch information is sent over a 25 cps bandwidth and the buzz-hiss information rate is negligible.

Incidentally, the number 25 cps for the low pass filter agrees with an intuition that the spectrum should change at something like a phonemic rate. Phonemes are rarely of shorter duration than 40 msec.

From the above description, it should not be implied that the vocoder is a well understood and smoothly functioning mechanism. The description was purposely simplified for the sake of directness. In fact, a great many seemingly simple questions about vocoders have not yet been conclusively answered. Here are a few of them:

1. How many channels should one use in the analyzer and synthesizer?
2. What should be the sharpness of cutoff of the filters?
3. What should be the bandwidth of the bandpass filters?
4. How accurate must be the pitch detector and buzz-hiss detector?
5. Why do vocoders sound like "kazoos", like "talking down a drain pipe": why do they have "electrical accents"?
6. What kind of spectral distortion can be expected from a given vocoder?

Many more questions can be, and have been, raised, particularly with regard to digitized vocoders. Today, most vocoder designs still depend more on personal taste and on system constraints than on any well accepted set of principles.

The formant tracker is a band compression device which exploits, even more than the vocoder, the redundancies in the speech. For example, during a spoken vowel, the assumption is made that a cascade of three resonant circuits is a sufficient simulation of the vocal tract. Thus, in many formant trackers, only six parameters are transmitted, i. e., the three formants and their amplitudes. At the synthesizer, these parameters slowly vary the resonance frequencies and gains of the resonant circuits. Many different types of formant trackers have been built; Fig. 6 shows a basic sketch of one.

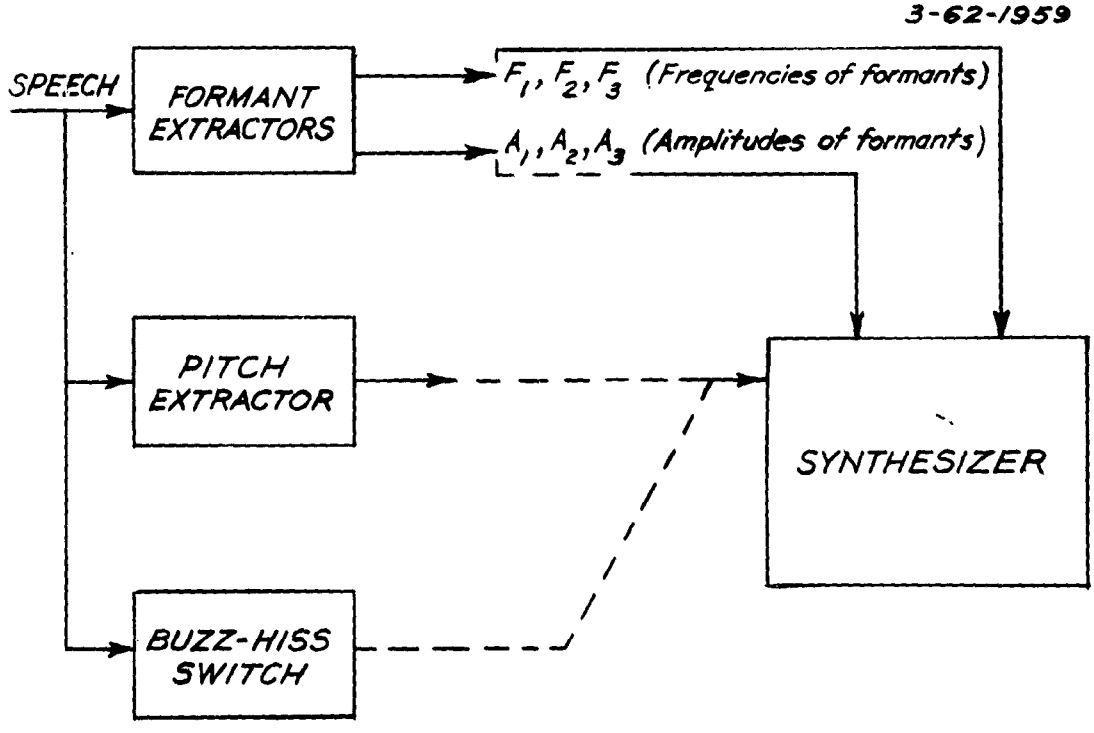


Fig. 6. Formant Tracking Vocoder

The derivation of excitation (pitch and buzz-hiss) is the same for both formant trackers and vocoders; they differ in the method of extracting spectral information.

Many digitized formant trackers transmit 1200 bits per sec.; many digitized vocoders transmit 2400 bits per sec.; this gives an idea of their relative band compression capabilities.

Formant trackers sound appreciably worse than vocoders; the reason is simple; speech spectra cannot always be represented by three resonances. During the semi-vowel and nasals, the first few harmonics are often the strongest and no definite resonances of the spectral envelope exists. Voiced plosives and voiced fricatives may also lack clear resonances. To make matters worse, vowel spectral envelopes may not exhibit three obvious peaks; for high pitched voices this problem is further complicated by the fact that sometimes no harmonic of the fundamental is present at a formant and thus there is no spectral peak to measure.

This is not meant to disparage formant trackers but, rather, to point out that those that have been built are too simple to handle the variety of speech sounds. Recently, quite sophisticated synthesizers have been built. One by Gunnar Fant of Sweden, uses 12 control parameters (including special networks for nasals) and can produce quite decent artificial speech. We can hope, therefore, that improved synthesis techniques will lead to better band compression systems.

There are a number of other band compression schemes; one is called the autocorrelation vocoder and the other is a frequency division scheme; a third is called pattern matching. Let us pass over them here. They are not yet practical (or even semi-practical) systems. Finally, we should mention the possibility of compression by recognition and synthesis.

This work is strictly in the research stage but it is worth discussing because it is, in principle, the "ultimate" band compressor. Imagine a device at the transmitter which can recognize words. Imagine also that the vocabulary is limited to 8192 words. Upon recognition of the spoken word, the transmitter needs to send only 13 bits. Assuming 5 words a second, the bit rate becomes 65 bits/second. Of course, no speaker identification is possible with such a scheme. However, there has recently been a surge of interest on this question and hopefully, we will soon have some idea of the information needed sufficiently to specify the speaker.

The Voice Excited Vocoder

About five years ago, speech research workers at the Bell Telephone Laboratories invented a new vocoder. The two important facts about this invention are, 1) a great improvement in quality was achieved, b) appreciably less band compression was obtained. Thus, the Bell people demonstrated that the vocoder principle was indeed sound in that a vocoder could be built which was of high enough quality to satisfy most telephone subscribers. Still unresolved, however, was the basic question: "what is the ultimate band compression capability of vocoders"?

The difference between the voice excited vocoder (V. E. V.) and Dudley's channel vocoder was in the method of generating excitation. If, in Fig. 5, the pitch detector and buzz-hiss detector are replaced by a band-pass filter (from about 300 cps to about 1000 cps) called a base-band filter, we have the analyzer of a V. E. V. The V. E. V. synthesizer is sufficiently different from that of Fig. 5. that we will show it separately (Fig. 7).

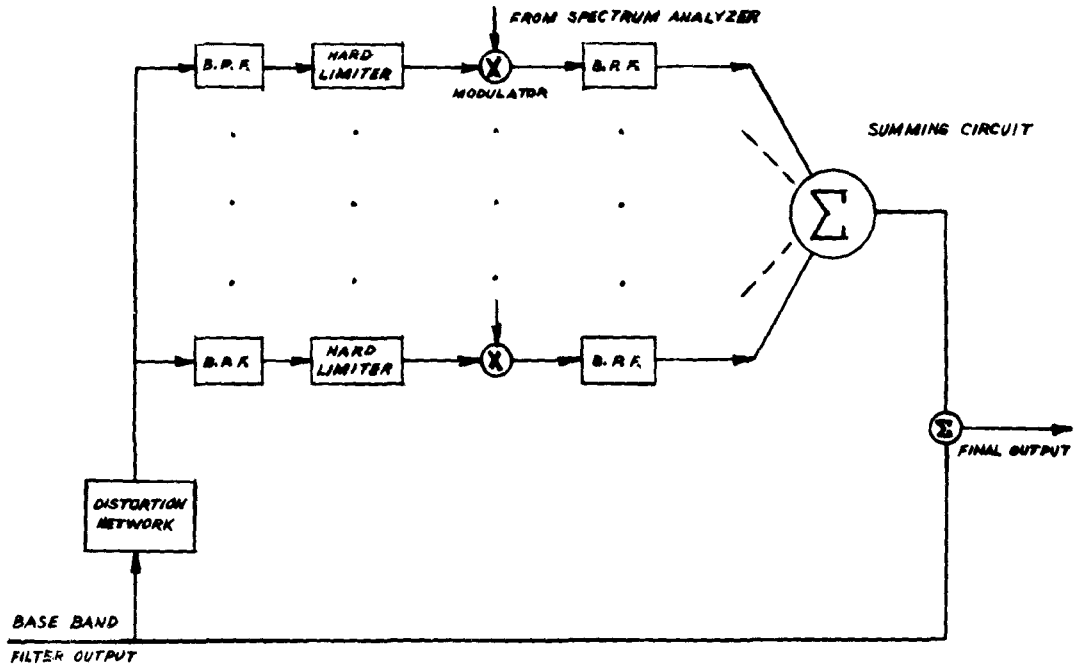


Fig. 7. V. E. V. Synthesizer

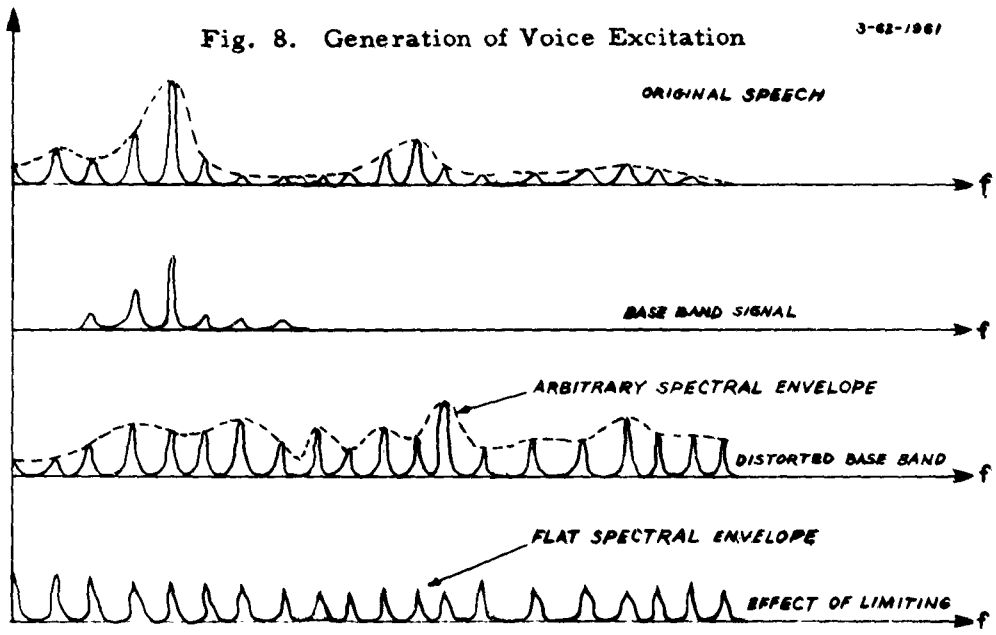
The band pass filters cover the band from 1000 cps up. Since the base band signal is an undistorted, filtered version of the original speech, there is no need for synthesis in the band 300 - 1000 cps.

We think that the reason for the high quality of a V. E. V. lies in its ability accurately to simulate the pitch and the buzz-hiss switching of the speech. This can be explained by reference to Fig. 3.

A(f), we remember, is the spectrum at any moment of the voiced speech. When the speech is passed through a base band filter, the spectral envelope is changed but the harmonic structure is undisturbed. Of course, there will be no energy beyond 1000 cps. At the synthesizer, a simple distortion network

(such as a half wave rectifier) will reproduce the harmonic structure of the original speech. Finally, the combination of the leftmost bank of filters (of Fig. 7) and the limiters will yield an excitation which has the correct harmonic structure and whose spectral envelope is now flat. The sequence of spectral changes is illustrated in Fig. 8.

Fig. 8. Generation of Voice Excitation



We see that the V. E. V. has performed a synthesis very much in accord with the basic model of Fig. 1; the excitation contains only information of harmonic structure and cannot cause any distortion of the spectral envelope of the synthesized speech.

The action of the limiters allows the V. E. V. also automatically to reproduce buzz-hiss information. When voicing disappears, the base band signal becomes aperiodic or zero. In either case, as long as there is even a slight amount of noise available at the synthesizer input, the limiters will be captured by the noise and produce constant power at the modulators. Furthermore, buzz-hiss transitions are more gradual than in standard vocoders.

The transmission of 700 cps of speech more than doubles the bandwidth requirements of the V. E. V. over the channel vocoder. The obvious next question is; can one build a pitch detector and buzz-hiss detector of sufficient accuracy so as to compete in quality with the V. E. V. and to compete in band compression capability with the channel vocoder?

Summary of Group 62 Activity

The remainder of this report describes part of the activities of Group 62 in the field of vocoders during the past two and one-half years. No attempt will be made to describe in detail the vocoder that has been built; such a description will be given in a future paper. Here we will discuss original motivations, relate our work to previous work already described earlier in this report and survey what appear to be the most obvious accomplishments.

To begin with, speech research work was quite well established at the Lincoln Laboratory two and one-half years ago. Lerner had built a band compression system, utilizing a novel pitch detector. Frick had initiated and the Forgies had established an automatic speech recognition program. There was active work in speech under Stevens and Halle at M. I. T. and under Caldwell Smith and Wathen-Dunn at AFCRL. The TX-2 computer was probably the most powerful speech processing facility in existence. Thus, an individual interested in speech research would not have great difficulty in acquiring the tools and knowledge necessary to speech work.

The immediate motivation for an actual effort was the Group's interest in the vocoder as a vehicle for strategic communications. Our hope that something useful could be accomplished was based on two assumptions: (a) that the successful construction of a very accurate pitch detector would

lead to a vocoder of comparable quality to that of the voice-excited vocoder but using appreciable less bandwidth and (b) that the techniques of pattern recognition would constitute a new and useful application to the problem of pitch detection.

By the summer of 1961, a TX-2 program existed for the accurate extraction of pitch. Furthermore, the TX-2 derived pitch pulses had been recorded on the second track of audio tape which contained on the first track the original speech. This tape was played through the Bell Laboratories voice-excited vocoder (used, for our purposes, as a channel vocoder) and recordings made of the output speech.

The pitch extracting TX-2 program demonstrated that accurate pitch detection was possible. The recordings we made at BTL indicated that this accurately obtained pitch gave speech quality which was comparable to that of the V. E. V. As a result of this work, it was decided that Group 62 should try to build a complete digitized vocoder capable of giving relatively high quality speech over a 5 K bit channel. At present, this work is very near completion.*

There are three items worth discussing in more detail. The first is; what do we mean by "good quality" of a vocoder? The second is; how does our pitch detector differ from other pitch detectors? The third is; what were the most important facts we learned about vocoder principles?

It is obvious that a written description of vocoder sounds can do little to acquaint the reader with these sounds. It is strongly recommended

---

*The construction of the vocoder has been under J. Tierney's direction. J. Dumanian has done the design and construction of the necessary digital hardware.

that anyone who has more than an academic interest in vocoders listen to many vocoder outputs. Unfortunately, even listening tests are not yet sophisticated enough to assure that different vocoders can be properly compared. When listening to vocoders, beware of the following:

1. A voice (usually male) with very good diction and speaking with little intensity and intonation changes will make many vocoders sound very good.

2. Listening to a loudspeaker rather than good earphones decreases greatly the perceptiveness of the listener to differences between two vocoders.

3. Knowing the sentence beforehand makes you think the vocoder is better.

4. It is much more difficult to judge quality absolutely than relatively. Thus, to compare two vocoders, or one vocoder with two different parameters, the same sentence should be played sequentially to the listener. Even here, a bias appears; the sentence will often sound better the second time.

It seems clear that no rigorous rating of any vocoder can be made at present. (This problem belongs to the field of psycho-acoustics where there have recently been some advances). However, let us risk a comparison of Group 62's vocoder with others. Suppose that 1000 people were given sample outputs from (a) the best channel vocoder (b) the voice excited vocoder and (c) Group 62's vocoder, and asked the question, "Does (c) sound more like (a) or like (b)?" Our guess is that a large majority (80%) would say "b". If furthermore, 1000 people were asked to give their opinions of (b) vs. (c), we would guess that from 40% to 60% of them would choose (c). Stated much more crudely, we think the V. E. V. and the Group 62 vocoder are about the same. We feel even more strongly that the Group 62 vocoder is, in any case, much closer to the V. E. V. than is a channel vocoder.

The above opinion is only one man's and probably controversial. The main point to making it is to give the reader some insight as to where we think we stand.

Also, it is worth noting that all our comparisons were based on a high quality speech input. For speech input through a telephone handset with a carbon button, it has been established by tests conducted at the Bell Laboratories that the V. E. V. degrades the telephone speech very little, if at all. No such claim can be made for the Group 62 vocoder. Some preliminary tests indicate a definite degradation of carbon button speech; however, it is only fair to point out that very little effort on our part was made in this direction. Our assumption was that, for strategic communications, high quality inputs could be afforded.

How does our pitch detector differ from other pitch detectors? The most fundamental difference lies in the fact that the pitch detector makes use of redundancy. With two exceptions (which we will discuss later), all pitch detectors have been built to make decisions based on a single measurement. Dudley, for instance, in his original vocoder, used a filter which sloped downward beginning at about 90 cps, so that for most voices, the higher harmonics of the speech would be attenuated relative to the fundamental. Pitch was extracted by generating pulses at the positive going zero crossings of the filter output. Others, such as Dolansky, have used modified versions of peak detectors. Lerner's* scheme uses a most ingenious technique for extracting pitch. It is well described in a Lincoln Laboratory Group Report so we will not go into it here. Many others have built pitch detectors. Using hindsight, one can easily catalog what is wrong with each and every one of the

*R. M. Lerner, "A Method of Speech Compression", G-Report 36-41, 24 August 1959.

above ideas. The important point is this: No one has yet succeeded in satisfactorily extracting pitch using one decision rule, no matter how sophisticated. The Group 62 pitch detector is based on the following idea: Let us build a few pitch detectors which run simultaneously and then let us see if we can, by combining the results from the individual detectors, arrive at a decision more accurate than can be had from each of the separate detectors.

Since the details of the Group 62 pitch extractor are available in the literature*, no further description will be given here. The TX-2 program that was mentioned earlier proved to be impractical for hardware realization, so a modified version was written, also on the TX-2. This modified version has been built by Group 62† and is now satisfactorily operating as part of our vocoder. It is a rather expensive gadget, using over 100 flip-flops, a small core memory to do the decision making and about 50 analog cards (mostly operational amplifiers), which comprise six individual pitch detectors.

Gill, of the Joint Speech Research Unit in England and Minsky of M. I. T., have both built pitch extractors based on the notion of more than one rule. Gill, instead of using different rules on a fixed portion of the audio band, used the same rule repeatedly on different parts of the audio band. Minsky used four different rules on a fixed band of speech. Although the individual pitch detectors in all three cases, Gill, Minsky and ours, differ, the important difference lies in the method of picking the "most likely" pitch. Ours is an appreciably more complicated method which makes more use of a larger amount of data. Whereas Gill and Minsky developed their ideas in the

---

*B. Gold, "Description of a Computer Program for Pitch Detection", 4th International Congress on Acoustics, Copenhagen, August, 1962.

†V. Sferrino has designed and built the real time pitch computer. E. Aho has designed and built the six analog pitch detectors which supply data to the pitch computer.

laboratory, building the equipment to test their ideas, the Group 62 pitch extractor was developed wholly on a powerful general purpose computer.

What did we, during our work, learn of vocoders? Before we go into this, permit us the luxury of a very slight polemic on the subject of vocoders. They are disarmingly simple looking gadgets and, many engineers have thought to themselves that they could easily build one. Unfortunately, too many engineers with little or no understanding of the fundamentals of speech have actually gone ahead and built a vocoder. The results have usually been horrible; success is almost impossible unless one builds up a good intuition as to how variations in the many design parameters affect the final output.

We were lucky to have, at the beginning, the advice of experts in the field from both Bell Laboratories and Lincoln Laboratory. We were readily convinced that we needed very stable band-pass filters, wide dynamic range detectors and modulators, low pass filters which rang very little, etc. But our most important discovery was this; our very accurate pitch detector did not greatly improve the vocoded speech quality unless the vocoder synthesizer included spectral flattening. This meant that our pitch extractor had to be used in conjunction with the configuration of Fig. 7 rather than with the synthesizer of Fig. 5. The reasons for this are discussed in a forthcoming letter to the editor in the Journal of the Acoustical Society of America, (B. Gold and J. Tierney).

Briefly, the variations in pitch cause a spectral fluctuation even if they are in the form of sharp pulses rather than voice excitation. This fluctuation propagates through the synthesizer, causing a spectral distortion of the final product. The point we want to stress is this; all channel vocoders were victims of this pitch induced spectral distortion. It follows that all channel vocoders could be improved by adopting the synthesizer configuration used

for the V. E. V., regardless of the pitch detection scheme used! Although it is still true that vocoder performance is proportional to pitch detector performance, our work has resulted in a pitch-independent proposal for improving vocoders. We say "proposal" since until many listening tests are made, there will be no real verification of our idea.

DISTRIBUTION LIST

Director's Office

W. B. Davenport, Jr.
W. H. Radford

Division 2

F. C. Frick
C. D. Forgie
J. W. Forgie
M. L. Groves
C. K. McElwain
J. L. Mitchell
A. I. Schulman
O. G. Selfridge
J. E. K. Smith
M. L. Stone
W. S. Torgerson
M. L. Wendell

Division 6

G. P. Dinneen
P. E. Green, Jr.
D. Karp
R. M. Lerner
W. E. Morrow, Jr.
R. T. Prosser
L. J. Ricardi
T. F. Rogers
H. Sherman
Group 62 Files (5)

AFL

N. Levine

Group 62

E. J. Aho
R. S. Berg
N. L. Daggett
R. E. Drapeau
J. A. Dumanian
R. M. Fano
R. L. Givan
B. Gold
J. N. Harris
B. Howland
D. Huffman
D. A. Hunt
B. H. Hutchinson, Jr.
K. L. Jordan
R. S. Kennedy
I. L. Lebow
R. W. MacKnight
A. C. Parker
C. M. Rader
P. Rosen
V. J. Sferrino
J. Tierney
N. C. Vlahakis
R. V. Wood
J. Wozencraft