SPEECH BANDWIDTH COMPRESSION

by

LEON ALVIN HOLLOWAY

B. S., Kansas State University, 1963

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1964

Approved by:

Harklyn S. Hayes

Major Professor

TABLE OF CONTENTS

INTRODUCTION 1

TIME AND FREQUENCY COMPRESSION METHODS 9

Scan Vocoder 9

Vobanc 13

Codimex System 17

Correlation Vocoders 22

Sampling Techniques for Bandwidth Compression 29

CONTINUOUS ANALYSIS-SYNTHESIS METHODS 35

Channel Vocoders 35

Voice Excited Vocoders 37

Resonance Vocoders 41

The Harmoniphone 47

A Chebyshev-Type Vocoder 49

DISCRETE SOUND ANALYSIS-SYNTHESIS METHODS 56

Phonetic Pattern Recognition Vocoder 56

SOUND GROUP ANALYSIS-SYNTHESIS METHODS 63

CONCLUSIONS 66

ACKNOWLEDGMENT 68

BIBLIOGRAPHY 69

INTRODUCTION

The human speech often contains sufficient information to identify the speaker and his emotional status from its "sound". Included in the actual speech signal are all the harmonics and overtones that identify the speech as real (human). This extra information is desirable in normal conversation, but it constitutes a waste of communication capacity in the case of speech transmission over long distances. The question as to what part of the speech spectrum is not essential to its intelligibility remains unanswered.

Speech may be described as a modulation process in which at least two modulated carriers contain the information. The sound of the vocal cords, which represents a periodic series of pulses, and the sound of exhaled air, which represents a signal with a nearly constant spectrum, constitute the two carriers. The sounds of the vocal cords and exhaled air will be referred to as "voiced" and "unvoiced" respectively (see Fig. 1). The voiced carrier is both amplitude and frequency modulated in order to produce loudness and pitch changes. Both carriers undergo frequency-noise modulation in the cavity of the mouth, and thus generate "formants". The frequency and amplitude of a gross concentration of energy in the spectrum of a speech sound is defined as a formant (Flanagan, 1956). The amplitude modulation of both carriers produces consonants. The frequency of vibration of the vocal cords is defined as the pitch frequency.

The concentration of energy in the speech spectrum does not change rapidly with time. An inference about redundancy in the speech sound can be made from the fact that the ear can identify a sound on receipt of only a portion of the total energy. Therefore a sampled speech signal contains essentially all the intelligibility of the original signal. Since a speech signal may be described in either the time or frequency domain by the Fourier integral pair,

$$s(t) = \int_{-\infty}^{\infty} S(f)\ e^{j2\pi ft} df \qquad (1)$$

and

$$S(f) = \int_{-\infty}^{\infty} s(t)\ e^{-j2\pi ft} dt, \qquad (2)$$

either or both time and frequency sampling may be employed.

Flanagan (1956) shows that knowledge of the first three formants is sufficient to specify most voiced and unvoiced English vowels and consonants, and that most of the significant information in speech is contained in the frequencies below 3,000 cps. From this it is inferred that the speech spectrum is not efficiently used. Thus, continuous identification of sounds is unnecessary and, therefore, the entire spectrum need not be transmitted. The significance of the reduced information can be seen from Shannon's equation for channel capacity,

$$C = W \log_2 \ (1 + \frac{S}{N}) \qquad (3)$$

where C, channel capacity in bits per second, is directly proportional to channel bandwidth, W is the channel bandwidth in cycles per second, and S/N is the signal-to-noise power ratio.

The channel capacity required to transmit the formant information of speech (Flanagan, 1956) can be determined from the truncated Fourier series representation of the formant signals

$$f(t) = \sum_{n=0}^{N} (a_n \cos nwt + b_n \sin nwt) \qquad (4)$$

where W = $2\pi/T$, T is the duration of the sample in seconds, and a_n and b_n are normally distributed random variables with zero mean. If the channel possesses negligible phase distortion, the bandwidth necessary to transmit f(t) with a prescribed accuracy may be computed from the number of terms in the series. Tables I and II show the results of Flanagan's experiment. Shannon's equation also shows that for a given information content, bandwidth reduction can be achieved with an increase in the signal-to-noise ratio, which has strong limitations.

TABLE I. Experimental results of the bandwidth required to transmit the first three formants for various samples and speakers (Flanagan, 1956).

Sample	Speaker	Sample Duration	F_1	F_2	F_3
Joe took father's shoe bench out.	A.H	1.7	7.0	4.6	4.6
Joe took father's shoe bench out.	R.B.	1.5	8.1	7.4	4.7
She was waiting at my lawn.	A.H	1.4	8.0	8.7	5.1
She was waiting at my lawn.	R.B	1.7	7.0	6.4	4.6
The birch canoe slid on the smooth planks.	A.H	1.7	7.0	7.0	4.1
The birch canoe slid on the smooth planks.	R.B	1.7	6.4	5.3	4.6
No, not this time.	W.L	1.5	7.4	8.0	7.4
What did you say before that?	W.L	1.8	6.3	6.3	6.8
MEAN	---	1.6	7.1	6.7	5.3

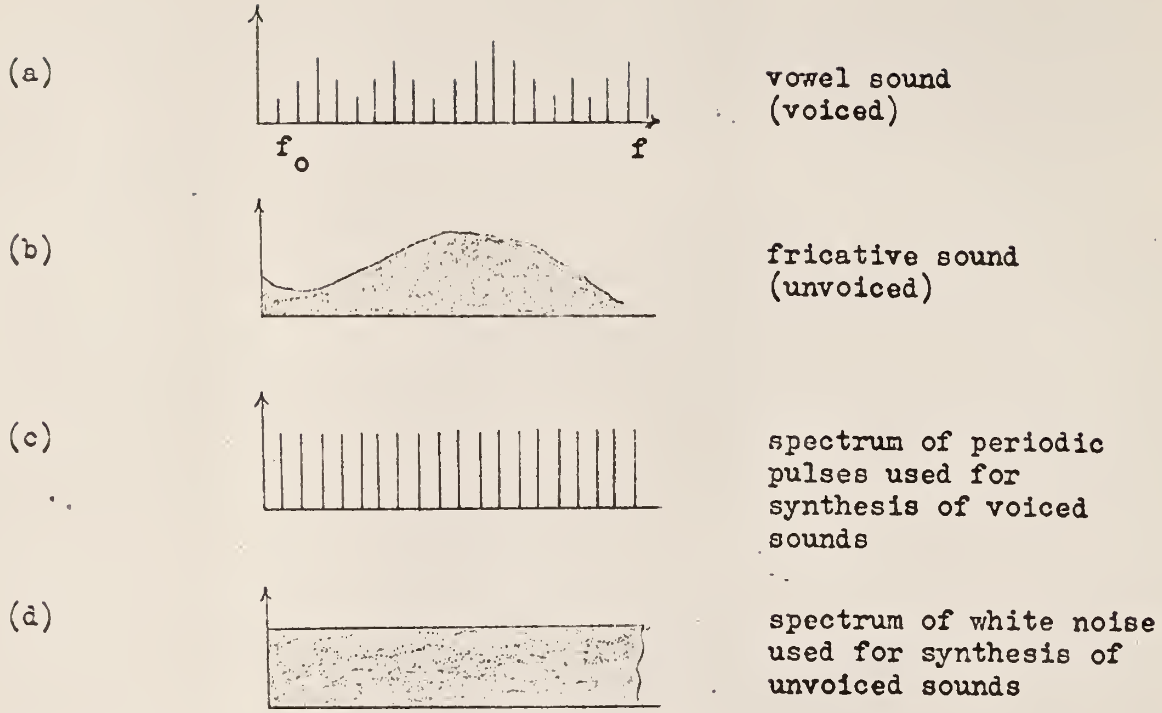
(a)
f_o
f
vowel sound (voiced)
(b)
fricative sound (unvoiced)
(c)
spectrum of periodic pulses used for synthesis of voiced sounds
(d)
spectrum of white noise used for synthesis of unvoiced sounds

Fig. 1. Fourier spectra (Schroeder, 1959).

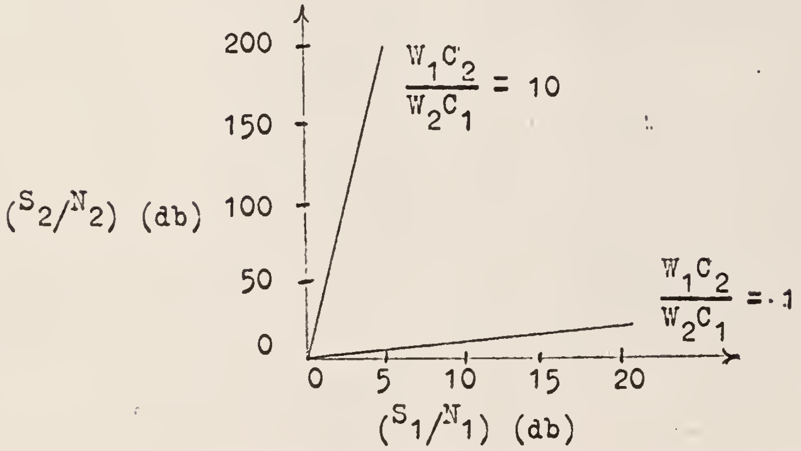

Fig. 2. Compressed speech channel signal-to-noise ratio as a function of original channel signal-to-noise ratio (Campanella, 1958).

TABLE II. Channel capacity necessary to transmit the formant signals for an overall S/N of 40 db (Flanagan, 1956).

Formant	W (cps)	S/N (db)	C (Bits/sec)
F_1	7.1	33	78
F_2	6.7	24	54
F_3	5.3	20	35

All bandwidth compression techniques attempt to eliminate at least part of the insignificant information in speech. Time and frequency compression methods are possible because most vowel sounds have a duration in excess of that required for the ear to identify the sound. Also, a typical vowel sound contains many repetitions of a basic vowel waveform. Furthermore, the information identifying the speaker and the speaker's emotional status is not essential, and may be eliminated.

Continuous analysis-synthesis methods exploit the redundancy in speech as well as the fact that the speech does not occupy all of the spectrum space all of the time. Discrete sound analysis-synthesis methods also eliminate redundancy and inefficient use of the spectrum as well as identity and emotional status information.

Sound group analysis-synthesis methods eliminate all insignificant information and transmit only a code to "call out" an entire word or phase held in storage in the synthesizer. This results in extremely low information rates. Table III shows some comparative channel capacities necessary for speech transmission by various means.

TABLE III. Comparative channel capacities necessary for speech transmission (Slaymaker, 1959).

Coding Method	Necessary Channel Capacity (Bits/sec)
Digitized speech waveform	30,000
Phonetic pattern coded speech	60
Word coded speech (at 120 words/min)	
Vocabulary of 2 words	2
Vocabulary of 8,000 words	26
Vocoder	2,000
Teletype (120 words/min)	75

It is possible to relate the signal-to-noise ratio in the compressed speech channel to the bandwidth reduction factor, the information reduction factor, and the signal-to-noise ratio in the original speech channel by the use of Shannon's equation for channel capacity. For example, let the subscript 1 in Equation (3) refer to the non-compressed channel and the subscript 2 refer to the compressed channel.

Then, the signal-to-noise ratio is given by

$$S_2/N_2 = (1 + S_1/N_1)^{(W_1/W_2)(C_2/C_1)} - 1 \qquad (5)$$

and for $S_2/N_2 \gg 1$ and $S_1/N_1 \gg 1$ Equation (5) becomes

$$S_2/N_2 = (S_1/N_1)^{(W_1/W_2)(C_2/C_1)}. \qquad (6)$$

In Fig. 2, the compressed speech channel signal-to-noise ratio is plotted as a function of the signal-to-noise ratio of the non-compressed speech channel for valves of the exponent $(W_1/W_2)(C_2/C_1)$

of 1 and 10. The unity exponent corresponds to the case when the information rate is reduced by the same factor as the channel bandwidth. Thus, for comparable performance, the signal-to-noise ratio will be the same in a compressed channel as in a noncompressed channel. Also, since the bandwidth required to transmit the compressed channel signal is reduced by the factor (W_1/W_2), the white noise energy picked up in the channel is reduced by the same factor, and the immunity of the compressed speech channel to noise interference is improved by 10 log (W_1/W_2) db.

The exponent $(W_1/W_2)(C_2/C_1)$ may take on values greater than one when the channel capacity is not reduced as much as the bandwidth is compressed. In this case, the signal-to-noise ratio in the compressed channel will always be higher than that in the noncompressed channel for comparable performance. In order to obtain comparable performance from a compressed channel and a noncompressed channel, with $(S_1/N_1) = 20$ db, $(W_1/W_2) = 10$, and $(C_2/C_1) = 1$, a signal-to-noise ratio of 200 db would be required in the former.

The preceding discussion points out that the effectiveness of a bandwidth compression system cannot be measured by the bandwidth reduction factor alone; the influence of information reduction must also be taken into account in terms of the signal-to-noise ratio, or the information rate that must exist in the compressed speech channel to obtain speech reproduction with a reasonable signal-to-noise ratio.

Two major advantages are gained by bandwidth compression. The first is a more efficient use of the communication space. Generally, telephone channels have a 3,000 cps bandwidth. This wide bandwidth greatly limits the number of channels possible in the alloted communication space. If the bandwidth can be reduced, say by a factor of 10, then the number of possible channels can also be increased by a factor of 10. The second advantage is found in increased noise immunity. The noise in communication channels is usually approximated by white noise, that is, noise with a constant spectral density. Thus, the total noise energy in a channel is directly proportional to the channel bandwidth, and the noise immunity improvement is directly proportional to the bandwidth reduction factor. This is especially desirable in long, noisy communication links.

TIME AND FREQUENCY COMPRESSION METHODS

Scan Vocoder

The Scan Vocoder (voice-coder) is one of the early time compression systems. A diagram of the Scan Vocoder is shown in Fig. 3 (a). In this system, the transmission of the speech signal spectrum-envelope requires frequency analysis. This analysis is performed by a set of magnetostriction filters covering the frequency range from 130 to 133 kc.

The output voltages of the analyzer filter set are rectified, and stored in capacitors. These voltages are then scanned by a rotating switch. The amplitudes correspond to the envelope of the voltage labeled as (3) in Fig. 3 (b). All future references in this section will be understood to be referred to Fig. 3 (b). The sampled output is then smoothed to obtain the envelope as indicated by waveform (4). The cut-off frequency of the smoothing filter is approximately 200 cps which is the bandwidth needed for transmission of the envelope. If a switch with low shunt capacitance is used, the high frequency filter outputs may be connected directly to the switch contacts and rectification may be accomplished with a single rectifier located between the switch arm and the low-pass filter.

The multivibrator and the hiss generator in the synthesizer are controlled by the pitch-frequency signal, so that the modulator input is either a line spectrum (see waveform B) or a noise spectrum (see waveform C). The upper sideband (130-133 kc) of

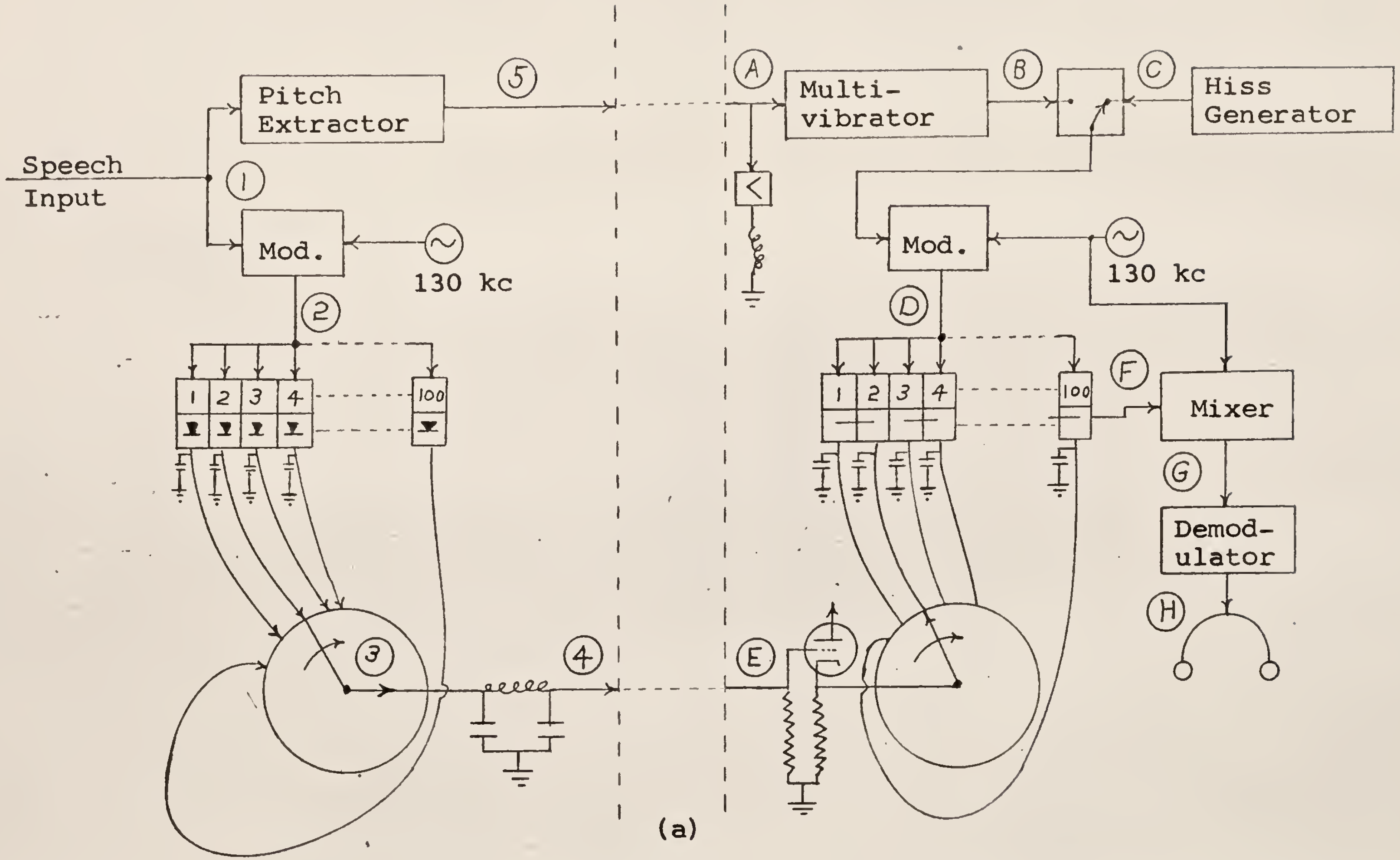

Fig. 3. The scan Vocoder and frequency spectra at various points in the system (Vilbig and Haase, 1956).

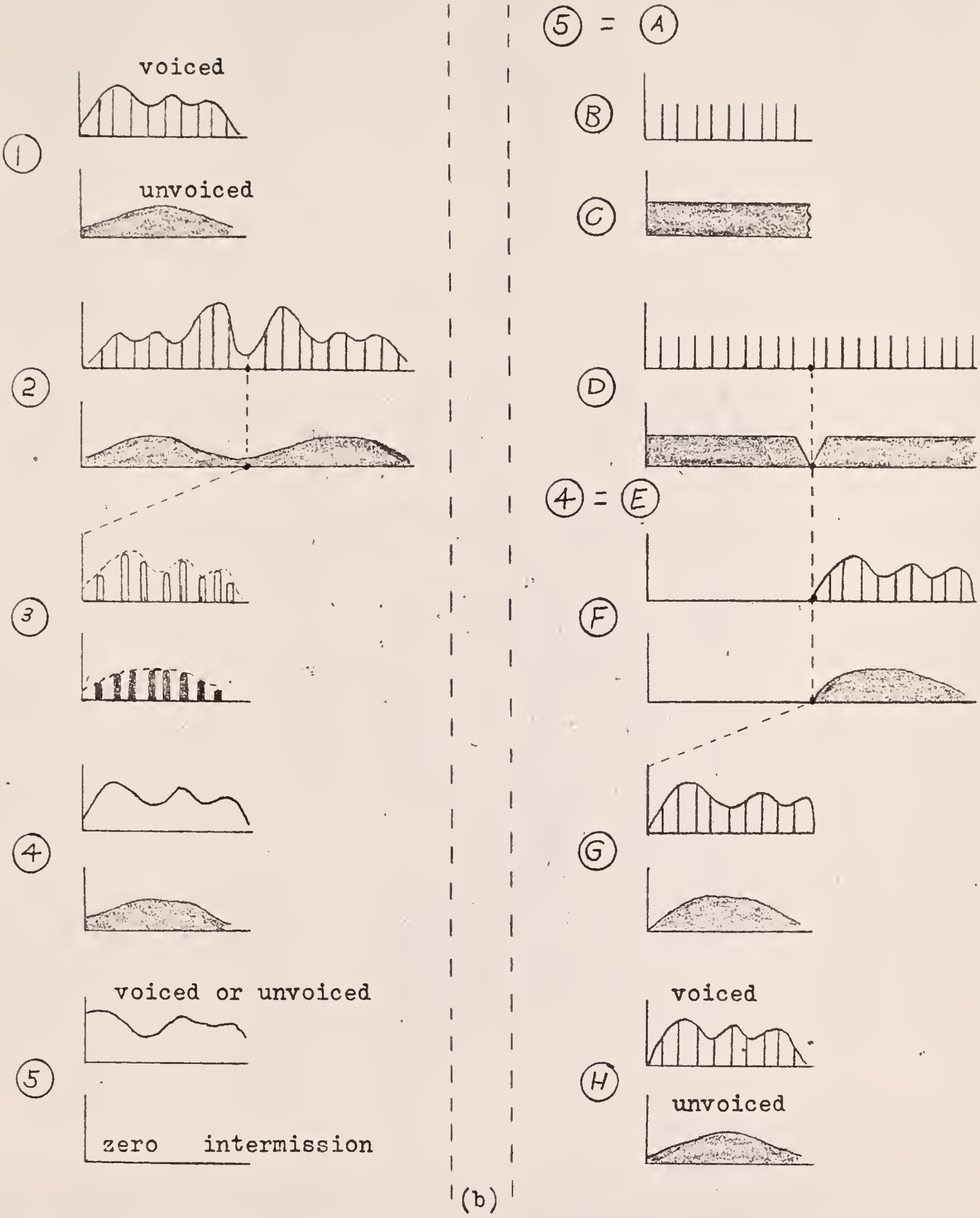

Fig. 3. The scan vocoder and frequency spectra at various points in the system (Vilbig and Haase, 1956).

the suppressed carrier output spectrum (see waveform D) is analyzed by another set of magnetostricicition filters. The filters are connected in pairs, so that there are only half as many outputs as there are in the analyzer. The filter outputs are connected to the high-frequency inputs of a set of modulators. The modulators receive a control voltage from the contacts of a rotating switch (see waveform (4) = (E).) Thus, the signal is sampled and stored. If the envelope changes between samples, the modulators receive the corresponding new voltage at the next sample. The modulator outputs are connected together in three groups. Group A contains the modulators 1, 4, 7, 10, ---; group B, the modulators 2, 5, 8, 11, ---; and group C, the modulators 3, 6, 9, 12, ---. A phase shift of 0°, 120°, and 240° is applied to the groups A, B, and C respectively, and then these three groups are added. This procedure restricts the modulation effect of any modulator to the frequency range of its corresponding filter. This is necessary since, otherwise, the envelope (see waveform F) of the sideband will be highly distorted. The complex output voltage of the three groups is demodulated by mixing with a carrier of 130 kc (see waveform G). The audio-frequency band thus detected is a close approximation of the original speech signal (Vilbig and Haase, 1956).

The Scan Vocoder is very complex and does not achieve a very large bandwidth reduction. For these reasons, this system has received little attention in the published literature on Speech Bandwidth Compression.

Vobanc

The Vobanc (Voice Band Compression) is a speech bandwidth compression system which utilizes frequency division and multiplication (Bogert, 1956). The general principle is to divide the speech based into three parts--0.2-1kc, 1-2kc, and 2-3.2kc--using filters after the speech signal has been pre-molulated (see Fig. 4). Each of these bands contains one of the vowel formants. The signal in each band is passed through a regenerative modulator (see Fig. 6) which halves the frequency of the strongest components of the formant, and translates the neighboring frequency components downward by a factor of F/2, where F is the frequency of the formant. The output of the regenerative modulator is filtered in order to obtain a bandwidth of one-half that of the original. At the receiving end, the frequency of each of the component bands is translated to double its value, and these are recombined in an attempt to generate the original spectrum.

A block diagram of the Vobanc is shown in Fig. 4. The input speech signal is modulated by a 108-kc oscillator. The difference frequency components are selected by "A" filters in three separate channels. The transmission characteristics of the "A" filters are shown in Fig. 5 (a). The A_1 filter transmits a band from 107.8 to 107 kc, which corresponds to the difference frequencies resulting from the modulation of the 108 kc carriers by a frequency in the range of 0.2 to 1 kc, corresponding to the first formant range. Second and third formant ranges are trans-

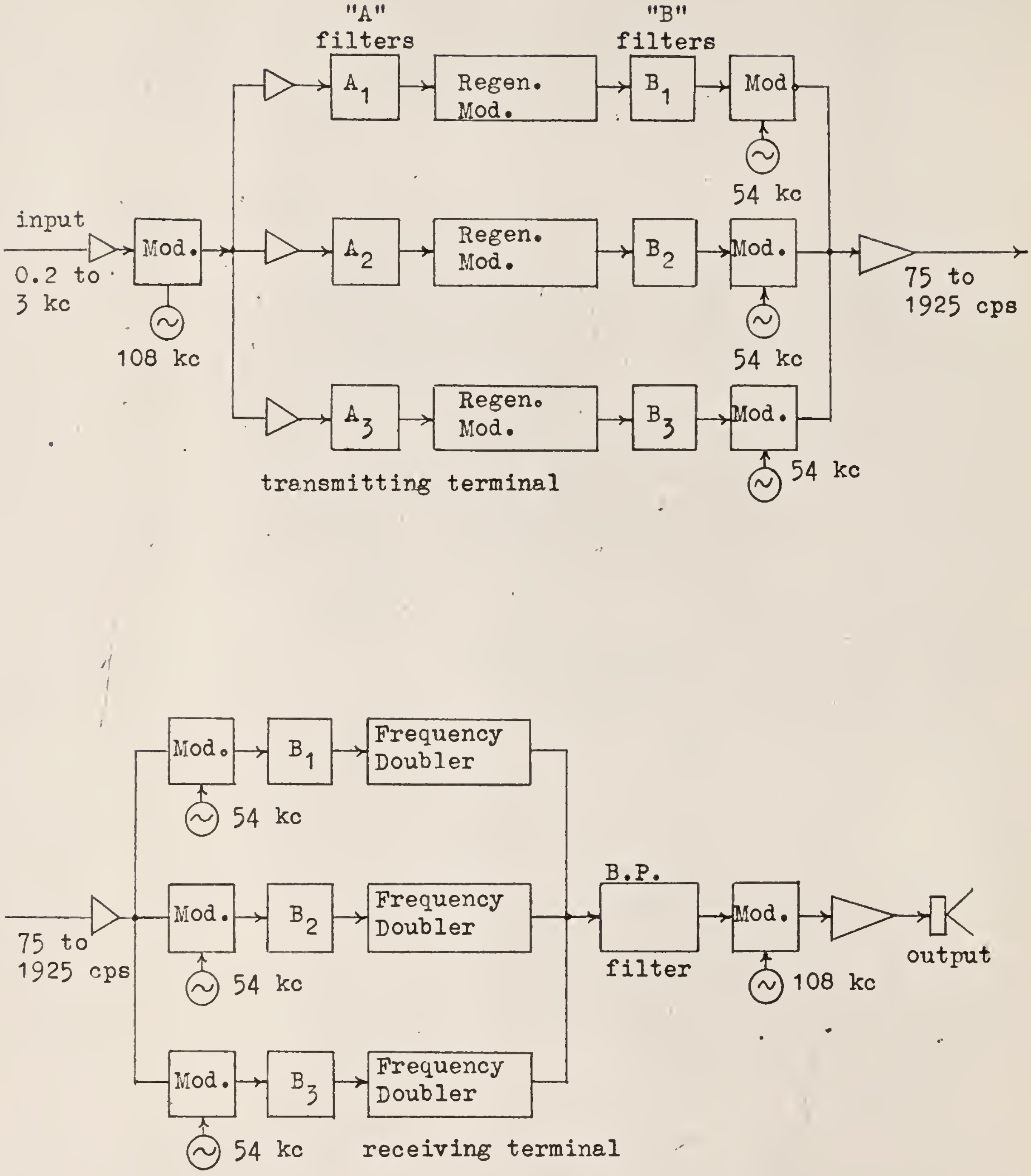

Fig. 4. Block diagram of the Vobanc (Bogert, 1956).

mitted by filters A_2 (107 to 106 kc) and A_3 (106 to 104.8 kc) respectively.

The output of each of the "A" filters is fed to a regenerative modulator (Fig. 6). The input of frequency f is modulated by a balanced modulator, whose output forms the input to a filter which selects only difference frequencies. This output is then amplified to form the carrier input signal to the balanced modulator. No feedback develops unless an input signal is applied. The circuit has a dynamic range of 35 db.

If two closely spaced frequencies are applied to the regenerative modulator, the average frequency of the input signal is halved, while the difference frequency between the two components remain the same. For speech signals, the formant frequencies are halved and the surrounding frequencies are reduced by half the formant frequency. The spacing between harmonic components of the speech signal remains the same in the process, but the range of formant variation is halved. Thus, at the output of the regenerative modulator the speech formant range can be included within a bandwidth one-half that of the corresponding A filter. The filters which select the half frequency components are labeled the "B" filters (Figs. 4 and 5 (b)). The frequencies passed by the "B" filters are then modulated down to frequency range 75 to 1925 cps by the output mixers of the transmitting terminal.

At the receiving terminal the compressed speech is again modulated up to the frequency range of the "B" filters. Each of

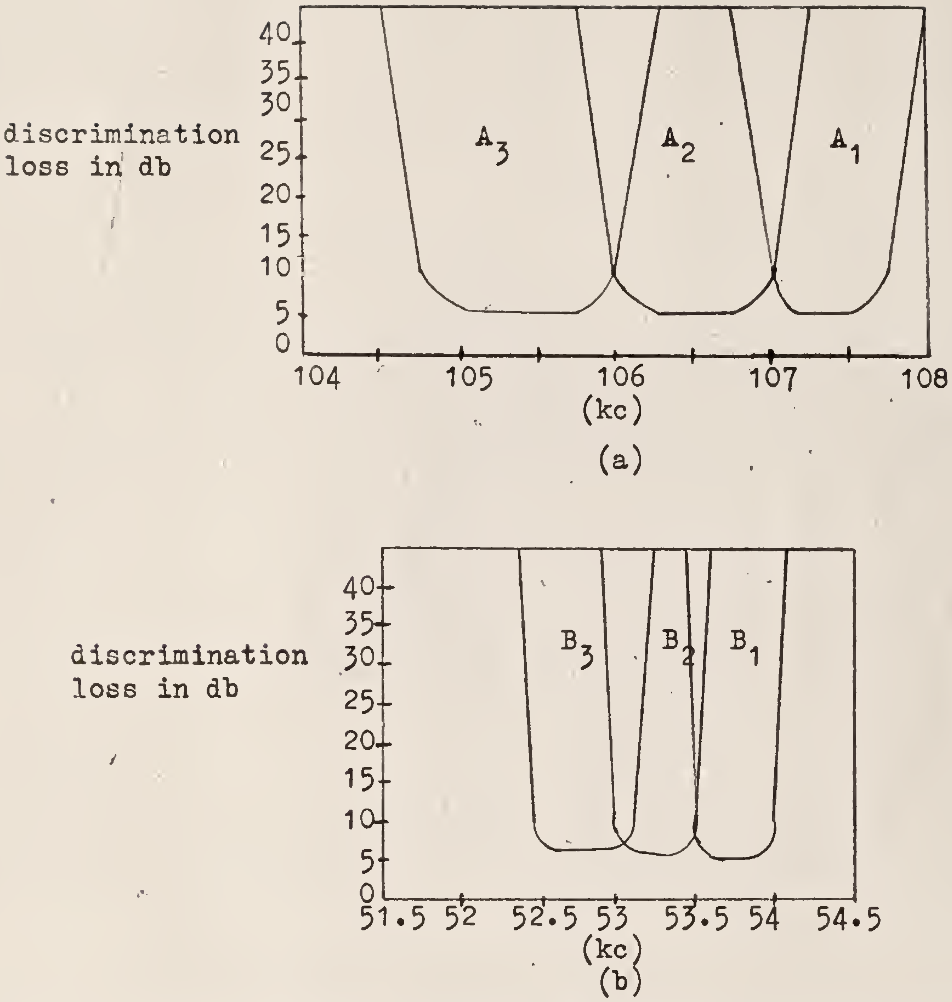

Fig. 5. Vobanc filter response characteristics (Bogert, 1956).

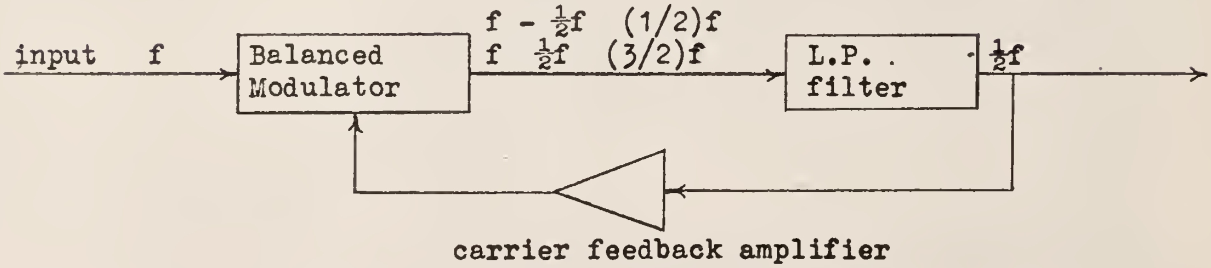

Fig. 6. Block diagram of regenerative modulator (Bogert, 1956).

the "B" filters output is doubled in frequency, summed, and filtered to restore the 108-kc carrier range. The resulting signal is then mixed with a 108 kc local oscillator to restore the signal to the audio-frequency range.

The Vobanc achieves a bandwidth reduction of slightly less than 2:1 due to the guard-band width allowed for the three channels. The quality of the speech is reasonable, as only slight distortion is introduced. The articulation effeciency ranges from 79 to 91 percent, where articulation efficiency is defined as the number of words understood divided by the total number of words transmitted.

Codimex System

The Codimex (compression-division-multiplication-expansion) system falls into the category of "formant tracking" devices, which also includes the Vobanc. The Codimex system uses many of the principles used in the Vobanc (Daguet, 1963). The instantaneous frequency of single sideband, suppressed carrier signal is put through a dividing process to obtain a 4 to 1 bandwidth reduction.

The spectral analysis of the compressed formant reveal the following effects:

1. Reduction of the frequency scale excursion by a factor of 8.
2. Concentration of the spectrum about an average frequency.

3. Increase of the average amplitude level which shows only slight variation.

The Codimex system transmits signals representing the compressed formants. These signals are transmitted at similar and slightly varying levels. The energy of the signals is concentrated in a very narrow frequency band.

The voice signal is separated into three parts corresponding to the three formant frequency ranges, namely 300-700 cps, 700-2000 cps, and 2,000-3,400 cps. Each formant is reduced in bandwidth by a separate operation. The starting point is the separation of the signal amplitude, a(t), and phase, cos Ø(t), as functions of time, in such a way that the real signal, S(t), may be represented as

$$S(t) = a(t) \cos \phi(t) . \qquad (7)$$

Let S(t) be a signal occupying a limited frequency band with finite energy. The Hilbert transform pairs,

$$\sigma(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{S(\tau)}{(\tau - t)} \, d\tau \qquad (8)$$

and

$$S(t) = - \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sigma(\tau)}{(\tau - t)} \, d\tau , \qquad (9)$$

are then used to form orthogonal component functions in such a way that

$$\psi(t) = S(t) + j\sigma(t) = a(t) \cos \phi(t) + ja(t) \sin \phi(t) \qquad (10)$$

where ψ(t) has been named the "analytic signal" by Ville (1948) and it was first introduced by Gabor (1945).

Another form of ψ(t) used is

$$\psi(t) = a(t)e^{j\phi(t)} \tag{11}$$

where a(t) is always a positive function. The actual signal is given by

$$S(t) = \mathrm{Re}\,\psi(t) = a(t)\cos\phi(t) \tag{12}$$

as before. Thus, the function a(t) and Ø(t) are unique and entirely determined from S(t).

The process of frequency compression is accomplished by subjecting the analytic function to a square root extracting process so that

$$\sqrt{\psi(t)} = \sqrt{a(t)}\,e^{\frac{j\phi(t)}{2}}. \tag{13}$$

For signals corresponding to the speech formants, the spectral analysis shows that the bandwidth reduction is proportional to the frequency compression. In the Codimex system, the square root extracting process is repeated three times given the result

$$[\psi(t)]^{1/8} = [a(t)]^{1/8}e^{\frac{j\phi(t)}{8}}. \tag{14}$$

The received signal undergoes the reverse process at the receiving end.

The amplitude and phase of the analytic signal may be obtained by subjecting the real signal S(t) to a single sideband supressed carrier (S S B) modulation process. If w is the carrier frequency, the SSB signal will be a(t) cos $[wt + \phi(t)]$. The single sideband signal may be obtained by two modulation processes using carriers in quadrature.

$$S(t) \cos wt - \sigma(t) \sin wt = a(t) \cos \phi(t) \cos wt - a(t) \sin \phi(t) \sin wt$$
$$= a(t) \cos \left[wt + \phi(t)\right]. \qquad (15)$$

The square root of the signal $a(t) \cos \left[wt + \phi(t)\right]$ is obtained using steps:

1. Detection of the envelope $a(t)$ (this is possible due to the separation of the spectra of $a(t)$ and $\cos \left[wt + \phi(t)\right]$).
2. Addition of the envelope to the SSB signal giving $a(t) \left\{1 + \cos \left[wt + \phi(t)\right]\right\}$.
3. Feeding this signal into a network whose output is proportional to the square root of the input, results in $\left[a(t) \left\{1 + \cos \left[wt + \phi(t)\right]\right\}\right]^{\frac{1}{2}} = \sqrt{2a(t)} \left|\cos \tfrac{1}{2}\left[wt + \phi(t)\right]\right|$.
4. Division by two, effected by switching a scale of two each time the signal $\left|\cos \frac{wt + \phi(t)}{2}\right|$ passes through zero.
5. Reversing the sign of $\cos \frac{wt + \phi(t)}{2}$ each time the scale of two is switched, thus producing $\sqrt{2a(t)} \cos \frac{wt + \phi(t)}{2}$.

The square root operation is repeated three times and the three channels are transmitted by frequency multiplexing. At the receiver, the signal is demultiplexed and separated into the three formant bands. The compressed waveform is then squared by means of a fullwave rectifier operating along a parabolic characteristic

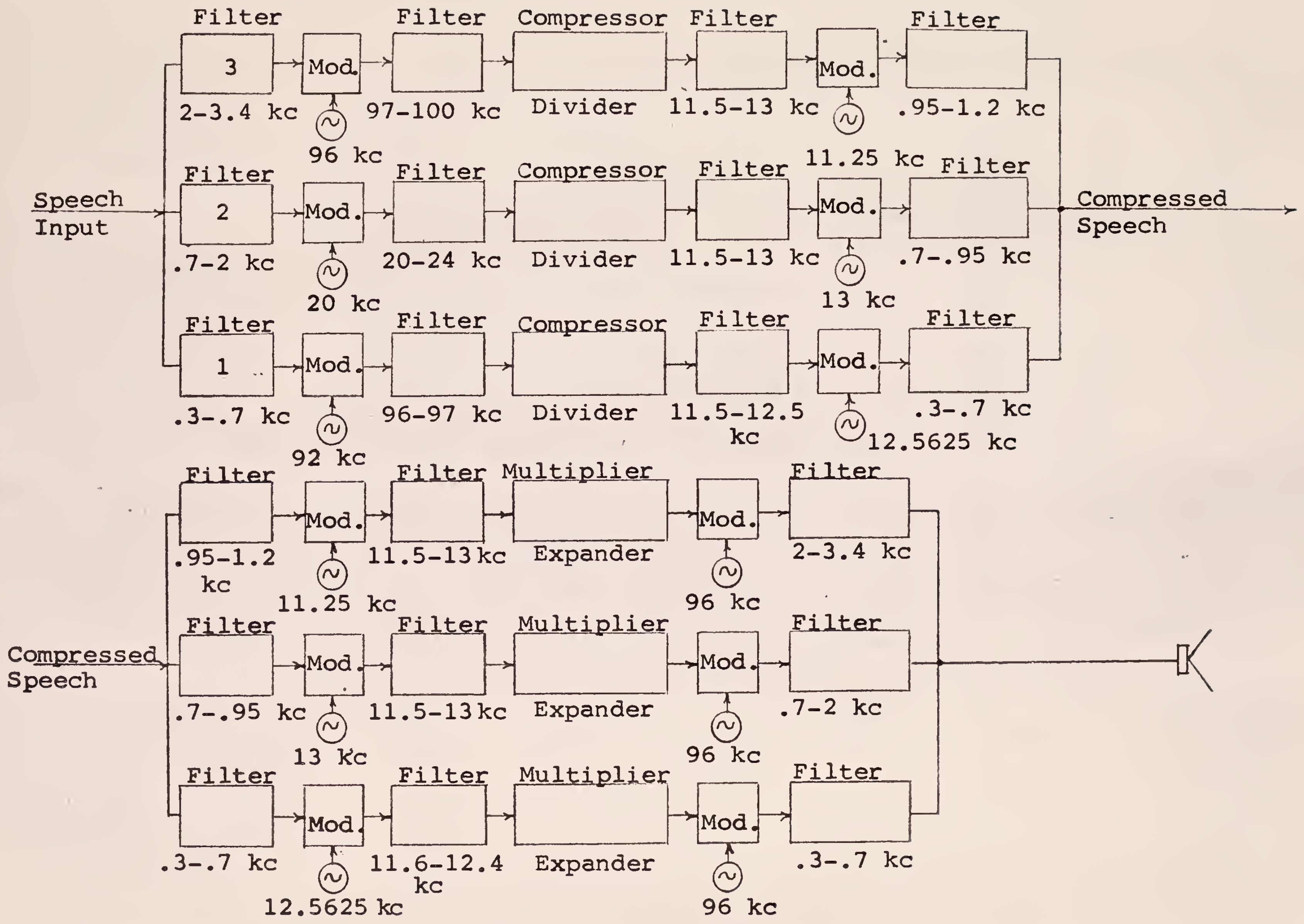

Fig. 7. Codimex speech compression system (Daguet, 1963).

$$\left[\sqrt{2a(t)}\ \cos \frac{wt + \phi(t)}{2}\right]^2 = 2a(t)\ \cos^2 \frac{wt + \phi(t)}{2} = 2a(t)\ \left\{\frac{1 + \cos\left[wt + \phi(t)\right]}{2}\right\} \qquad (16)$$

The low frequency signal a(t) is eliminated by a high pass filter leaving a(t) cos $[wt + \phi(t)]$, which is demodulated to restore the original signal $S(t) = a(t) \cos \phi(t)$. A block diagram of the Codimex system is shown in Fig. 7.

The Codimex system provides a rather modestbandwidth reduction, but reproduces good quality speech. Further bandwidth reduction is possible, but the system is said to become quite complex.

Correlation Vocoders

Correlation Vocoders (Schroeder, 1962) utilize speech analysis in the time domain by correlation techniques. The Wiener-Khinchin relationship suggests that correlation analysis can take the place of spectral analysis. The autocorrelation function and the power spectrum of a given signal form a Fourier transform pair,

$$\phi(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(w)\ \cos w\tau\ dw \qquad (17)$$

and

$$G(w) = \int_{-\infty}^{\infty} \phi(\tau)\ \cos w\tau\ d\tau. \qquad (18)$$

Fano (1950) has extended this relationship to include short-time analysis as

$$\phi t(\tau) = \frac{e^{a|\tau|}}{2\pi} \int_{-\infty}^{\infty} G_t(w)\ \cos w\tau\ dw \qquad (19)$$

and

$$G_t(w) = \int_{-\infty}^{\infty} e^{-\alpha|\tau|} \phi_t(\tau) \cos w\tau \, d\tau, \tag{20}$$

where $\phi(\tau)$ is the short-time autocorrelation function and $G_t(w)$ is the power spectrum and $1/\alpha$ is a time constant.

The autocorrelation function is taken over an interval of approximately 30 milliseconds for speech analysis, and is given by

$$\phi(\tau) = \overline{S(t) \cdot S(t-\tau)} \tag{21}$$

where the bar denotes the time average. $\phi(\tau)$ is symmetric in τ, and is bandlimited to the same frequency range as the signal. The spectrum $\Phi(f)$, of the autocorrelation function is the absolute square of the signal spectrum, S(f). Therefore,

$$\Phi(f) = |S(f)|^2. \tag{22}$$

Thus, $\phi(\tau)$ contains the same information as the amplitude spectrum of the signal $|S(f)|$.

An autocorrelation vocoder is shown in Fig. 8. A short-time autocorrelation function of the speech signal is derived for a number of discrete delays, τ_0, τ_1, $---\tau_N$, in the analyzer. The autocorrelation function is completely specified for discrete delays and has a spacing of $\Delta\tau = \frac{1}{2}f_c$, where f_c is the cut-off frequency of the speech. For $f_c \angle 3.3$ kc, a $\Delta\tau$ of 0.167 msec. suffices. The maximum delay for which the short-time autocorrelation function needs to be specified is of the order of 3 msec. In Schroeder's vocoder, there are 18 "delay channels", each with a bandwidth of 20 cps for a total bandwidth of 360 cps. This approaches a bandwidth compression ratio of 10:1, but when the pitch channel and guard-bands are taken into account the ratio

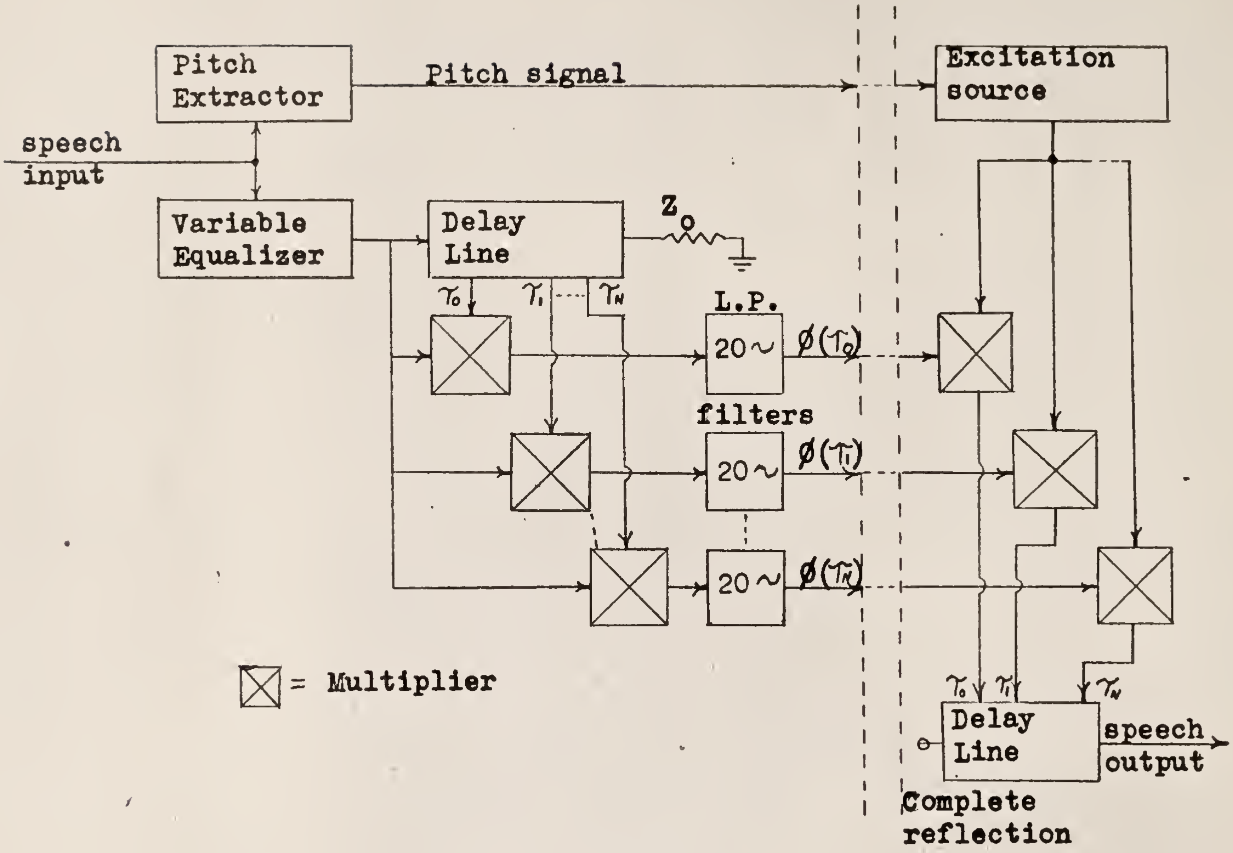

Fig. 8. Autocorrelation vocoder (Schroeder, 1962).

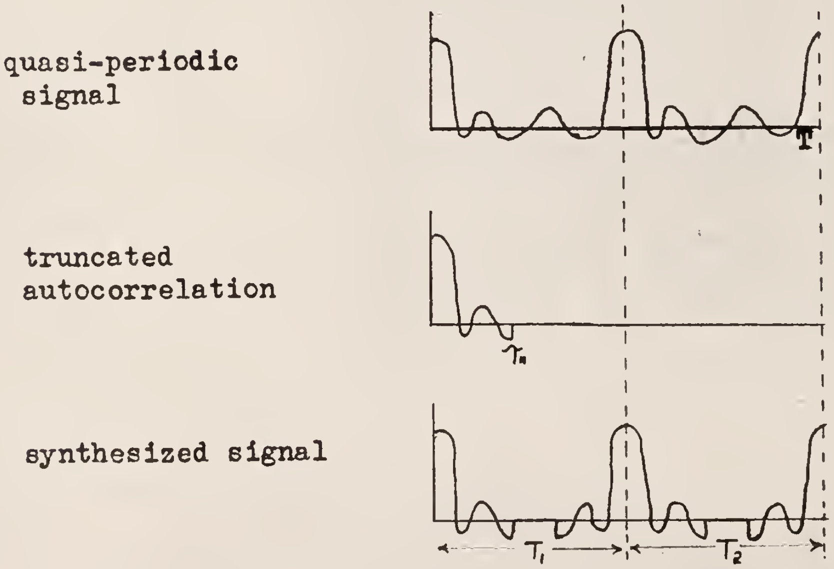

Fig. 9. Synthesis of speech by reciprocating scanning of auto-correlation function (Schroeder, 1962).

is reduced to the order of 9:1. The autocorrelation vocoder conserves bandwidth in a manner similar to the spectrum analyzing devices. The phase information is discarded, the autocorrelation is averaged over several fundamental periods, and some spectral resolution is sacrificed.

At the synthesizer a symmetrical replica of the autocorrelation function is generated for every pitch period by reciprocating scanning (Fig. 9). Neglecting the truncating distortion apparent at the ends of some scans, the synthesized signal has an amplitude spectrum that is the square of the original speech spectrum $\left[\phi(f) = |S(f)|\right]^2$.

The spectrum squaring, inherent in autocorrelation vocoders, needs to be compensated if natural sounding speech is to be obtained. While spectrum-squared speech is fairly intelligible, it has an unpleasant muffled and uneven quality. A time-varying equalizer (Fig. 10) compensated for the squared spectrum. It consists of three filters for formant extraction, rectifiers, low-pass filters, square root extractors, and dividers. In the equalized signal at the output, the formant amplitudes are reduced to the square root of their original amplitudes, thus compensating for the spectrum squaring of the autocorrelation vocoder.

Schroeder rated the autocorrelation vocoder as high in intelligibility and fair in quality. A certain distortion, attributed to the chopping of the individual pitch periods (see Fig. 9), was noticeable. To reduce this distortion, the autocorrelation coefficients have been tapered by a "Hamming" window function,

$$H(\tau) = 0.54 + 0.46 \cos \frac{\pi\tau}{\tau_{max}} \tag{23}$$

In order to minimize the number of autocorrelation signals to be transmitted, the maximum delay, τ_{max}, for speech frequencies above 2.5 kc has been reduced to 1.5 msec. τ_{max} has been maintained at 2.5 msec for the medium speech frequencies (1.5 to 2.5 kc). In order to improve the spectral resolution at low frequencies, τ_{max} has been extended to 5 msec for frequencies below 1.25 kc. The corresponding increase in the number of channels is small because the sampling interval for low frequencies is relatively large, for instance at 0.625 kc, the sampling interval is 0.4 msec. The total number of channels in the improved version of the autocorrelation vocoder is 27 for an input bandwidth of 5 kc. A bandwidth reduction of the order of 9:1 is maintained, however. Schroeder rated this version of the autocorrelation vocoder as superior or equal to the best known spectrum channel vocoder with comparable bandwidth compression.

The problem of spectrum squaring in a correlation vocoder can be avoided by cross-correlating the speech with a speech-derived signal having a flat spectral envelope. A cross-correlation analyzer is shown in Fig. 11. The spectrum flattener contains a non-linear network producing a flat distortion spectrum at the output for a variety of speech inputs. Several such spectrum flatteners have been invented at Bell Telephone laboratories and are described in the literature.

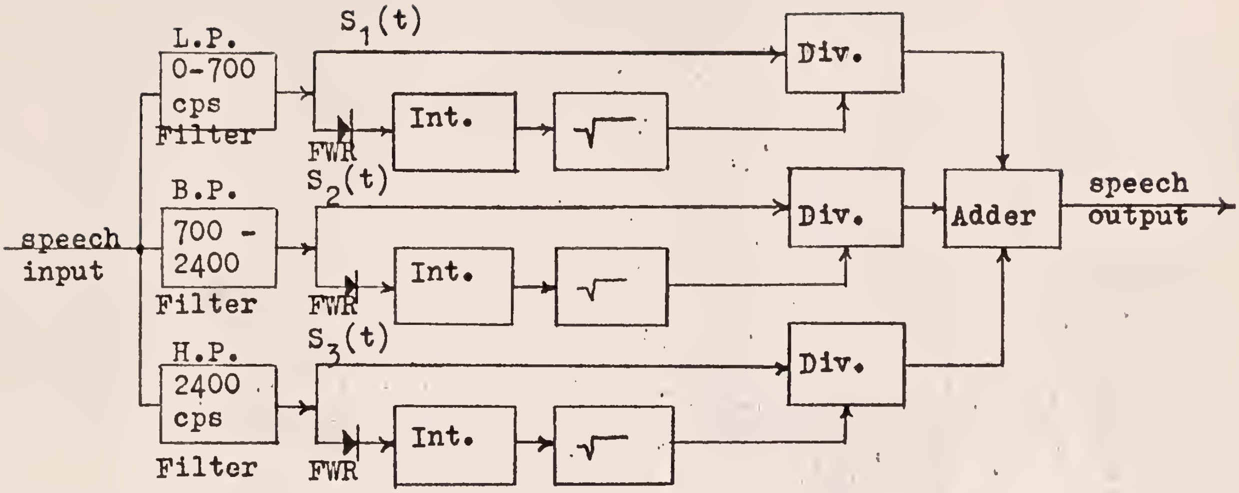

Fig. 10. Variable equalizer (Schroeder, 1962).

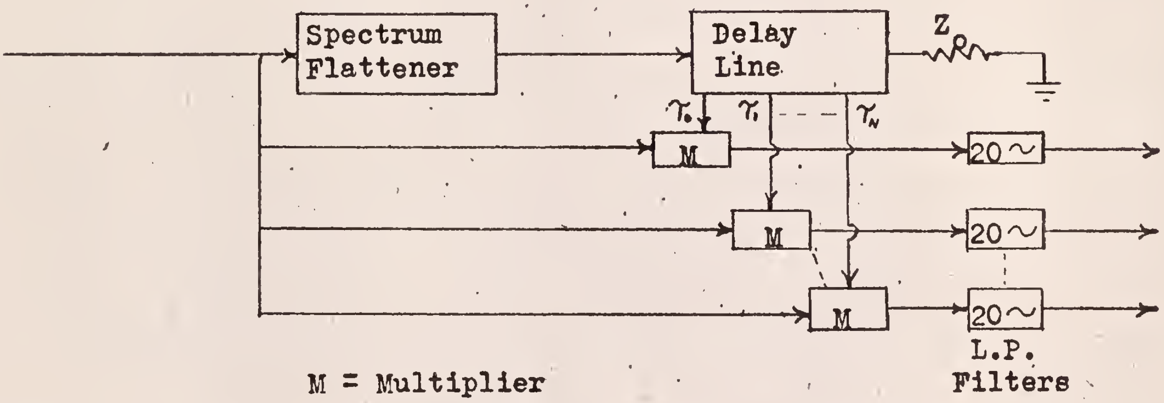

Fig. 11. Cross-correlation analyzer (Schroeder, 1962).

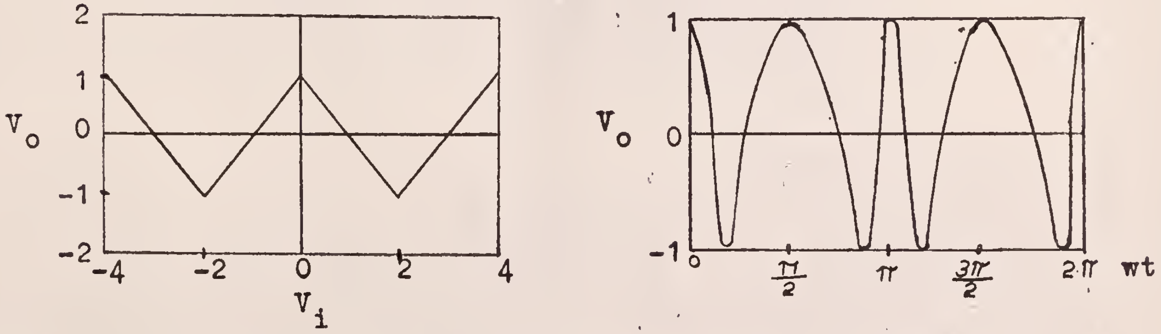

Fig. 12. Spectrum flattener characteristic and output waveform for input of 4 sin wt (Schroeder and David, 1960).

A spectrum flattener, described by Schroeder and David (1960), consists of a piece-wise linear network with an input-output characteristic of straight-line segments (see Fig. 12). The output is + 1 volt for inputs of 0 and ± 4 volts, and - 1 volt for an input of ± 2 volts. Thus, for an input voltage of greater than 3 volts, the output will contain 4 zero crossings for each one in the input. The spectrum flattener is quite independent of the form of the input (clipping or no clipping). This multiplication of zeros is accompanied by the desired spectral flattening.

The synthesizer of a cross-correlation vocoder is identical to that of an autocorrelation vocoder, provided that the cross-correlation function is reasonably symmetric so that reciprocal scanning can be made nearly symmetric by the proper choice of the reference signal with which the speech is cross-correlated. One such reference signal consists of pulses occurring at the relative maxima of the speech signal and having amplitudes proportional to the square of the speech maxima.

Schroeder termed the cross-correlation vocoders as inferior to the autocorrelation type for two reasons. First, the cross-correlation function is not truly symmetric, thus, a synthesizer requiring symmetry produces distortion in the center of each pitch period. Secondly, the reference signal does not have a truly flat spectrum. Thus, the synthesized signal differs from the original spectrum. The major advantage of cross-correlation analysis is that no analog multipliers are required if the reference signal is in binary form (Schroeder, 1962).

Sampling Techniques for Bandwidth Compression

Recently, various authors have suggested sampling speech signals both in the time-frequency domain and in the frequency domain alone. Peterson and Subrahmanyam (1959) attempted to compress the effective speech bandwidth by simultaneous sampling in the time and frequency domains, but the results were not very encouraging. On the other hand, Kryter (1960) has shown that satisfactory communication could be obtained by sampling the speech with three 500 cps bandpass filters, but the total bandwidth in this case comes out to about 2340 cps at 30 db. Also, due to the non-uniform sampling, the different bands have to be translated to form a compact spectrum so that frequency-division multiplex system may be used. A method of uniform sampling in the frequency domain along with a necessary correction to make direct multiplexing possible has been shown by Das (1961).

A finite sample of a time varying signal may be represented as a sum or integral of exponentials as

$$F(w,t) = \sum A_n e^{P_n t} \qquad (24)$$

where A and P are complex, if the waveform is known for all time. For any random function, such as speech, a possible representation in the time-frequency plane is of the form

$$F(w,t) = \sum \sum A_{mn} U_{mn}(t) \qquad (25)$$

where A_{mn}, given by $S(m\Theta, n/\Theta)$, are the coefficients of the sampling functions $U_{mn}(t)$ corresponding to the cross-points of any grid laid on the t-w plane, for which the separation of the crossline is Θ in time and $1/\Theta$ in frequency. Then, the signal may be represented by (m x n) numbers for a given Θ, and the sampling may be accomplished in either the frequency or time domain. Ideally, a signal of bandwidth W and duration T require 2TW numbers to specify it completely, but slowly-varying signals, such as speech, have much less essential information-contents than specified by 2TW numbers.

Dudley (1940) has shown that the voiced sound may be represented as

$$F_v(w,t) = S(t)\sum_{k=1}^{n} r(w,t)A_k \cos\left[kP\int_0^t P(t)\,dt + \Theta_k\right], \qquad (26)$$

where the carrier is composed of n audible harmonics of relatively high frequencies having amplitude A_k, frequency kP, and phase Θ_k. S(t) is the switching function, P(t) is the inflecting factor, and r(w,t) is the effect of selective transmission. The three message functions S(t), P(t), and r(w,t) produce the necessary modulation processes on the carrier at the low rate at which the syllables are formed. The total information is mainly dependent on these slowly-varying parameters. The unvoiced case is only a degenerate case of Equation 26.

The effect of sampling in the frequency domain is shown in Fig. 13, where the instantaneous amplitude of the component frequencies may be represented by A_1, A_2, ---, A_n. The sampled signal, consisting of A_2, A_5, A_8, etc., has holes in the amplitude-frequency curve and consequently the intelligibility and

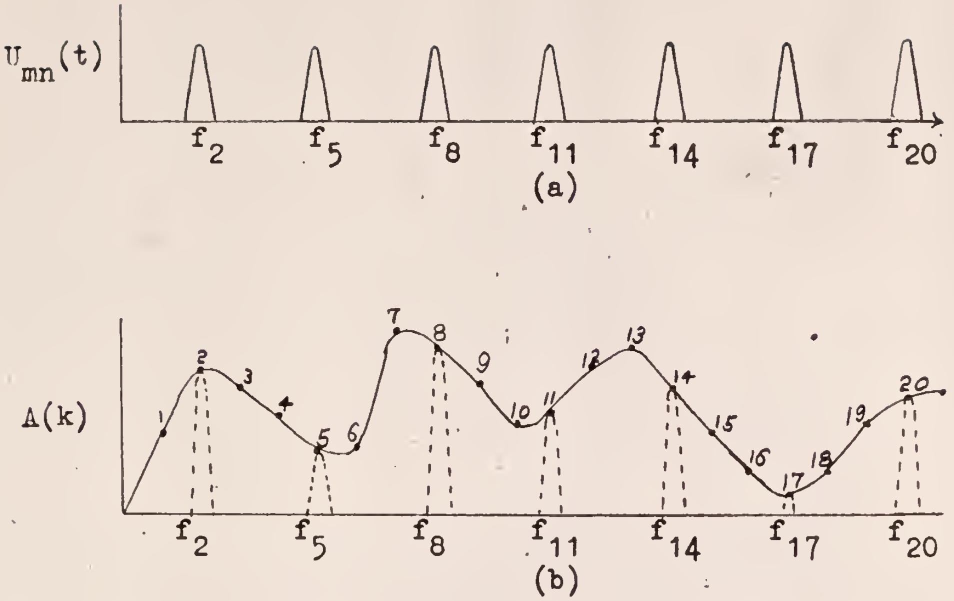

Fig. 13. Nature of sampling functions and a typical amplitude-frequency curve and its samples (dotted curves) (Das, 1961).

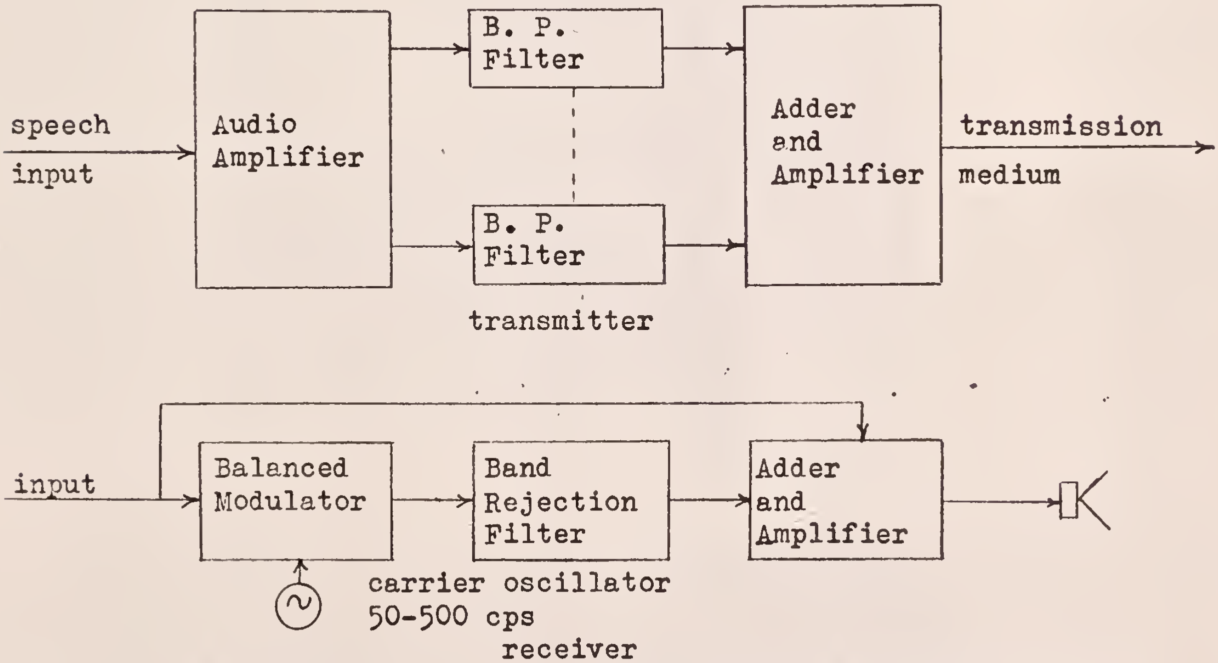

Fig. 14. Experimental apparatus for sampling and reconstruction of the speech signal (Das, 1961).

naturalness deteriorates. A process, similar to the pulse-lengthening generally used in time-division multiplex systems, is used to improve the quality of the signals. The holes created by sampling are filled by inserting sidebands corresponding to the accepted bands, but differing in frequency by $(\Delta f/3)$, where $\Delta f = (f_s - f_2) = (f_8 - f_5) = (f_n - f_{n-3})$. The reconstructed amplitude curve would then be of the staircase type and the new frequency-domain representation of the signal would be

$$F(w,t) = F_1(w,t) + F_2(w,t) \tag{27}$$

where

$$\begin{aligned} F_1(w,t) &= Am_2U_{m2}(t) + A_{m5}U_{m5}(t) + A_{m8}U_{m8}(t) + --- \\ &= \text{the sampled signal} \end{aligned} \tag{28}$$

and

$$\begin{aligned} F_2(w,t) &= A_{m2}\left[U_{m1}(t) + U_{m3}(t)\right] + A_{m5}\left[U_{m4}(t) + U_{m6}(t)\right] \\ &\quad + A_{m8}\left[U_{m7}(t) + U_{m9}(t)\right] + --- \\ &= \text{sidebands of the sampled signal} \end{aligned} \tag{29}$$

Some measure of the error in the synthesized signal may be obtained from the relations

$$\text{Variation} = \int \left|D'(f)\right| df \tag{30}$$

$$\text{Difference} = \int \left|D(f)\right| df \tag{31}$$

$$\text{Square of difference} = \int \left|D(f)\right|^2 df \tag{32}$$

where D(f) is the difference between the original spectra and the synthesized spectra, the difference over all f being normalized to zero.

In a practical sampling system, the coefficient A_{mn} will represent a small band of frequencies and it is necessary to determine the minimum number of bands that are to be transmitted as well as the width of each band. An experimental system is shown in Fig. 14. In the transmitter, the different bands are selected by bandpass filters, and, in the receiver, the transmitted signal as well as the sidebands generated in the balanced modulator are added together before final amplification. In order to avoid transient disturbances in the output, the sampling filters should have a smooth cut-off characteristic, even at the expense of the bandwidth of the system. Leakage of the carrier in the receiver causes masking of the signal and a sharp band-rejection filter is used to eliminate it from the output. Another alternative is to translate the speech signal to a higher band of frequencies, say 6-10 kc, then sample it, mix it with its sidebands, and then retranslate it back to its original frequency band.

Das (1961) found that with six filters, each with a bandwidth of 200 cps, and with Δf (gap width) equal to 600 cps, that without the addition of the sidebands the system had an articulation efficiency of about 85%. With the addition of the sidebands, the articulation efficiency approached 100%. The articulation efficiency was found to improve with increased bandwidth. The optimum shift of the samples was found to be about 150 to 250 cps, and any attempt to fill up completely the gaps having $\Delta f > 750$ cps tended to deteriorate the receiver output. Tests were initially performed with a time-varying carrier being fed to the modulator to cover the wider gaps in the spectrum, but the results with the fixed carrier were found to be better. The sampling bandwidth may be decreased to about 100 cps, but the filter transients become prominent in the output. The carrier suppression in the modulator has to be more than 50 db. Because of the smaller effective bandwidth of the system, the signal-to-noise ratio in the output is also improved.

Das concluded that since the sampling is uniform, different channels may be multiplexed without any frequency translation of the different filter outputs, and that the bandwidth compression possible by this method is superior to that of other similar methods.

CONTINUOUS ANALYSIS-SYNTHESIS METHODS

Channel Vocoders

In 1939, Dudley introduced the first channel vocoder shown in Fig. 15. It recognized that speech may be voiced or unvoiced, and that intelligibility is retained by preserving the short-time amplitude spectrum. A set of band-pass filters (BP_1-BP_n), with rectifiers and low-pass filters, produces the discrete short-time speech spectrum. A separate device, called the pitch extractor (Fig. 16), develops a voltage proportional to the fundamental frequency of the voiced sounds. The pitch control voltage is also used to control voiced-unvoiced selection. The pitch voltage takes on values above a certain threshold for voiced sounds, but remains at a steady state value below the threshold for silence and unvoiced sounds. The pitch signal modulates the frequency of a cord-tone generator (buzz) at the receiver and selects either the cord-tone or noise for excitation in the synthesis. The spectrum signals are applied to modulators (M in Fig. 15) at respective inputs to an identical set of band-pass filters. The filter outputs are summed and the short-time spectrum is reconstructed. Each spectrum channel requires about 20 cps bandwidth and a signal-to-noise ratio somewhat less than that of a conventional telephone circuit. Generally, the pitch channel requires about twice the bandwidth of the spectrum channel. The channel vocoder can transmit highly intelligible speech at an information rate of the order of 2000 bits per

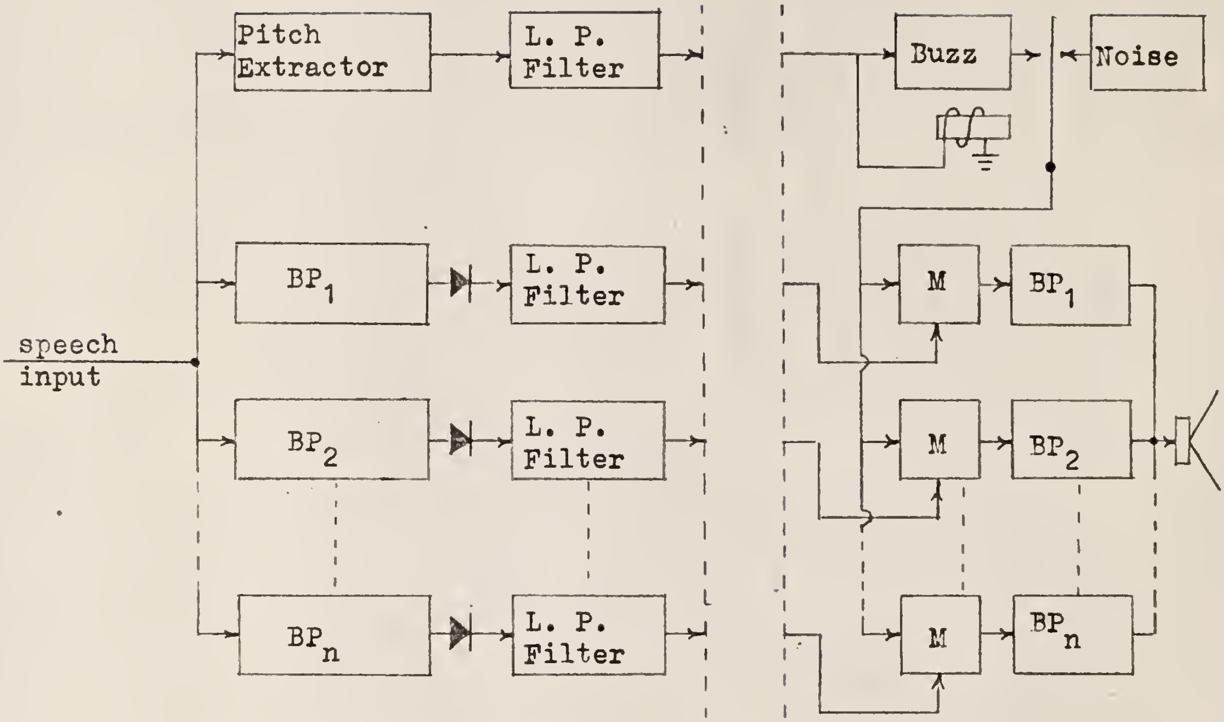

Fig. 15. Block diagram of channel vocoder (Dudley, 1939).

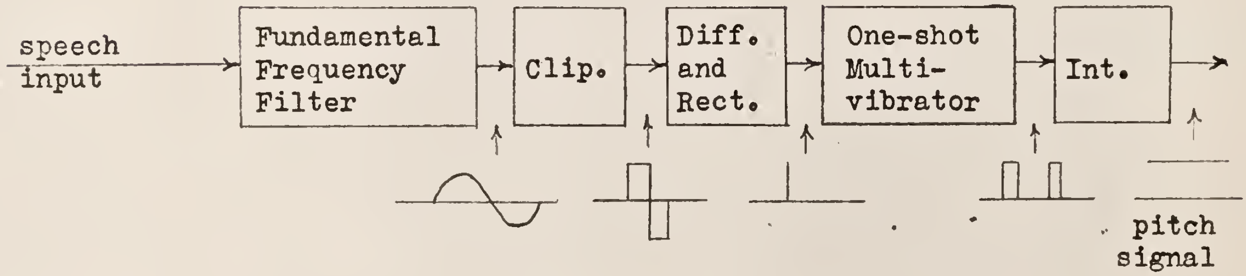

Fig. 16. Pitch extractor (Slaymaker, 1959).

second using 16 channels to span the telephone band (200-3200 cps). However, the speech quality of the vocoder is poor due to pitch errors and inadequate voiced-unvoiced detection.

Efforts to improve the speech quality led to the development of a split-band vocoder (Fig. 17). The split-band vocoder transmits a baseband (the lower one-third to one-half of the speech spectrum) over a conventional channel (no processing) and transmits the upper portion of the spectrum over several vocoder channels. The baseband is retarded at the receiver to equalize the delays in the vocoder channels and all the channels are recombined. Only noise excitation is used for the high-band synthesis, and voiced-unvoiced switching is performed by the baseband signal. The addition of the baseband corrected the pitch errors and substantially improved the speech quality. However, voiced-unvoiced detection still remained a problem (Flanagan, 1959) since no voiced excitation could be delivered to the high band.

Voice Excited Vocoders

The speech spectrum reflects the nature of the vocal excitation, and its description in vocoder transmission requires a voiced-unvoiced decision, based upon the amount of energy contained in a low frequency band which includes the pitch frequency. The pitch frequency is determined by an average zero-crossing count of the lowest frequency component of the speech spectrum. The reliability of this decision and the accuracy of the measurement of the pitch frequency depends critically upon the input speech quality, and in particular on signal-to-noise ratio and low-frequency equilization.

A voice-excited vocoder (Schroeder et al., 1962) avoids such difficulties by generating the excitation from an uncoded baseband of the original speech. The baseband may be added directly to the output, but its main function is to provide excitation for the synthesizer. A wide band excitation is generated from a narrow baseband by nonlinear distortion (see cross-correlation vocoder spectrum flattener, Fig. 12), which produces either a flat spectrum of noise or harmonic frequency components, depending on its input. Thus, the excitation is reproduced from the original speech and is not a result of any coding procedure.

This method is quite insensitive to input conditions and thus avoids the pitch problem. The voiced-unvoiced decision is also bypassed since this information is carried explicitly in the baseband. The voice-excitation also removes much of the electrical accent inherent in channel vocoders. However, bandwidth is sacrificed for the baseband transmission.

A major technical problem in voice-excitation is the required spectrum flattening. All schemes use nonlinear distortion to multiply the number of zero-crossings as described for the cross-correlation vocoder. The characteristics of such a device is similar to that shown in Fig. 12.

In order to achieve a reasonable bandwidth reduction factor, the baseband must be as narrow as possible. The nominal minimum for a wide range of speakers without incurring some degradation is about 700 cps (Schroeder et al., 1962). A filter bank is also needed in order to flatten such a band. The baseband is first spread by rectification into a wider band, and the spectral shape fluctuation are similar to that of the baseband. This fluctuation is removed by narrow-band filtering and clipping.

The intelligibility of the voice-excited vocoder does not depend on whether the baseband is added to the output or used only for excitation, because the voice-excitation mechanism preserves the rapid and inherent speech pitch fluctuations. The channel vocoder pitch circuit removes such desired fluctuations by averaging. Also, the conventional excitation is either a quasi-periodic waveform or noise, while the voice-excited vocoder (VEV) has a mixture (quasi-periodic for some frequencies, random noise for others) which can be appropriately reproduced.

Schroeder et al., (1962) found the VEV to be superior to the channel vocoder in the quality of reproduced speech, but inferior in bandwidth reduction due to the extra bandwidth required for the baseband. Recent work by these authors has made it possible to reduce the baseband to between 500 and 600 cps without appreciable degradation of the speech. Ten to twelve vocoder channels were found to be needed for satisfactory operation. The bandwidth of this system is between 800 and 1000 cps as compared to approximately 400 cps for the channel vocoder.

The transmission bandwidth for a bandwidth compression system depends on the type of modulation. Ordinarily single-sideband transmission should be used for the baseband but the

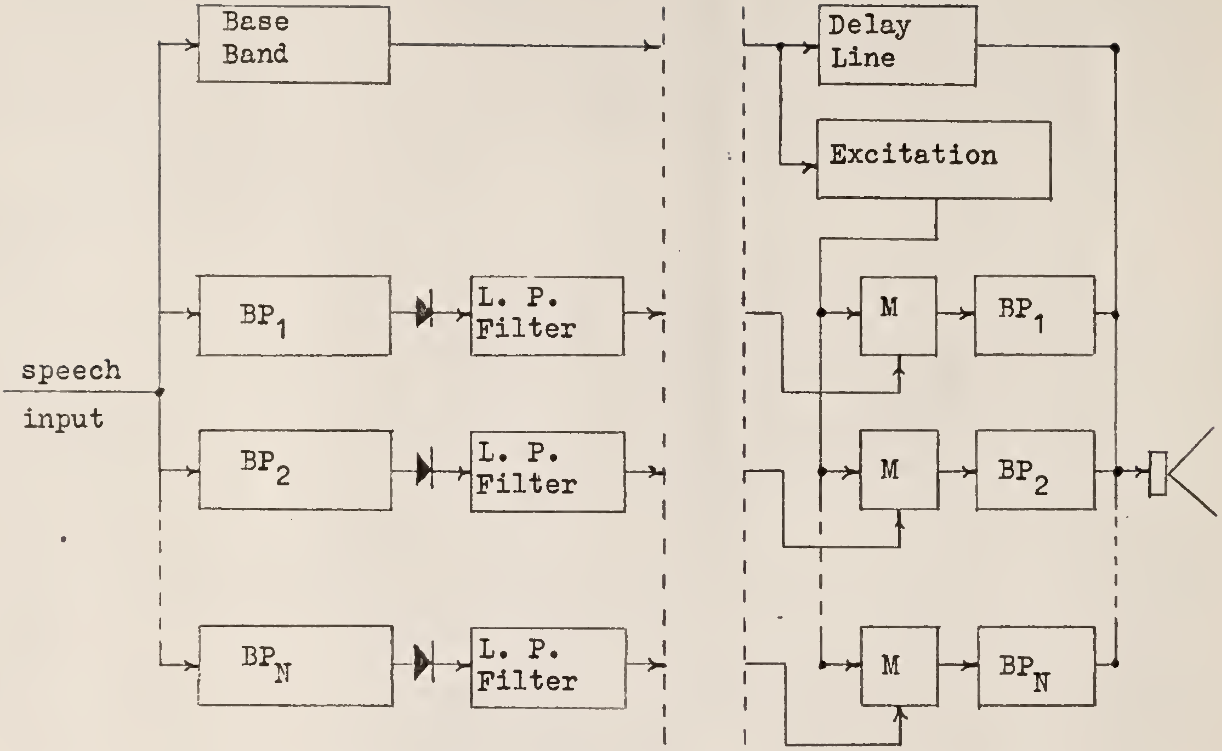

Fig. 17. Block diagram of split-band vocoder (Flanagan, 1959).

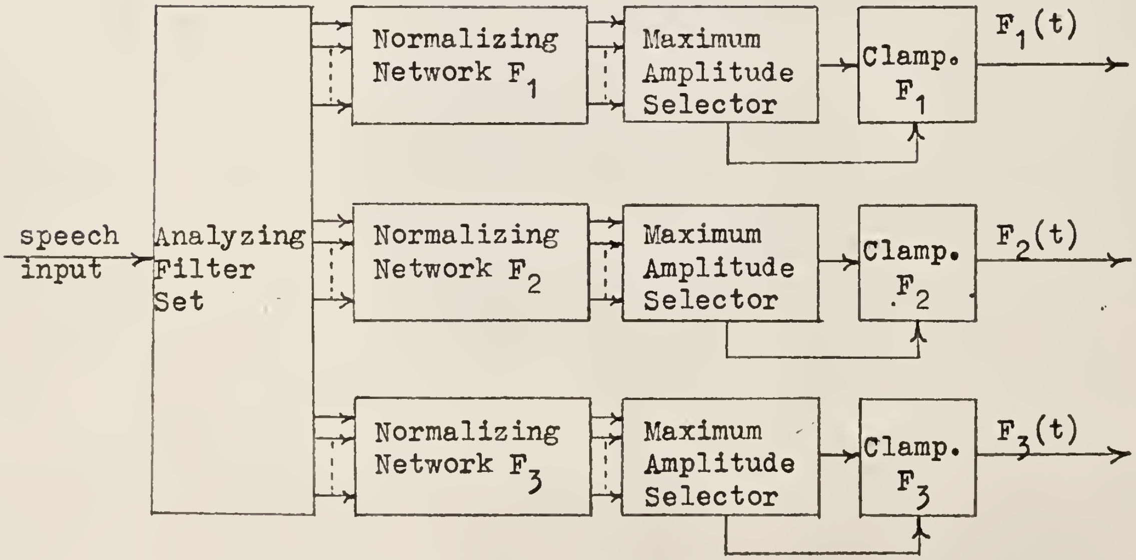
speech input
Analyzing Filter Set
Normalizing Network F$_1$
Maximum Amplitude Selector
Clamp. F$_1$
F$_1$(t)
Normalizing Network F$_2$
Maximum Amplitude Selector
Clamp. F$_2$
F$_2$(t)
Normalizing Network F$_3$
Maximum Amplitude Selector
Clamp. F$_3$
F$_3$(t)

Fig. 18. Formant extractor (Flanagan, 1956).

requirement to preserve the dc component of the channel signals prevents its use, therefore Schroeder et al. (1962) proposed quadrature modulation. In such a process, two signals are amplitude-modulated (DSB) onto a single carrier frequency, w_o, such that one of the signals is modualted onto cos w_ot and the other onto sin w_ot. Coherent carriers and a product detector are required at the receiver. In the usual multiplex situation, for instance, two normal voice-channels, carrier coherence cannot be assured to within the tolerance necessary for holding "cross-talk" between channels within acceptable bounds. Crosstalk is defined as mutual overlapping of information of adjacent channels. However, vocoder channel signals are highly correlated and because the ear is not too sensitive to spectral overlap within the same speech subbands, the crosstalk between adjacent channels needs to be held to within about 20 db, which is easily realizable by quadrature modulation systems.

Resonance Vocoders

Experiments in the analysis and preception of speech show that vowel sounds may be identified and synthesized from a knowledge of the formant frequencies (Flanagan, 1956). A formant extracting device for use in speech compression systems must accept continuous speech at its input and produce output voltages with time varying amplitudes representing the formant frequencies. A formant extracting device (Fig. 18) developed by Flanagan (1956) divides the speech spectrum into formant frequency ranges.

The frequency with maximum spectrum amplitude within each frequency range is then detected. This spectrum-segmentation method is based upon the fact that the first three formant fall in frequency ranges which, on the average do not appreciably overlap. An appropriate short-time spectrum of the input speech signal is obtained using a set of analyzing filters. This set is composed of 36 contiguous band-pass filters having a common input, but separate outputs. Each channel of the set includes a tuned circuit, an amplifier, a full-wave rectifier, and a smoothing network with a time constant of 10 milliseconds. The center frequencies of the filter channels are set on a Koenig frequency scale (logarithmic) extending from 150 cps to 7 kcs. The bandwidth of the filter channels are 100 cps for frequencies below 1 kc and increase logarithmically from 100 cps at 1 kc to 450 cps at 7 kc. The adjacent channels overlap at the half-power frequencies. The gain and bandwidth of each channel may be adjusted independently. The useful dynamic range of each channel is greater than 30 db (Flanagan, 1956).

The speech spectrum slopes downward at about 10 db/octave, on the average. Thus, it is desirable to perform a frequency equalization that permits all of the filter channels to operate at about the same signal level. It is also desirable to obtain a spectral output in which all the maxima are approximately of the same amplitude in order to alleviate the problem of dynamic range in the formant-analyzing equipment. A driver amplifier employing an equalizing network is used to supply the input to the filter set. The frequency response of the equalizing network rises at approximately 10 db/octave between 750 and 3000 cps, and is essentially flat outside this range. The network and the driver amplifier are an integral part of the filter set. Any reference to the filter set in the following discussion assumes that the input speech is equalized according to the frequency characteristic mentioned above.

The outputs of the analyzing filter set are separated into groups to cover the formant frequency ranges, $0 \leq F_1 \leq 800$, $800 \leq F_2 \leq 2280$, and $2280 \leq F_3$ cps respectively. The outputs of each group of filter channels are monitored and the channel having the maximum output within each group is selected and sampled at a rate of 60 times per second to indicate the formant frequency.

A normalizing circuit computes the mean value of its set of input voltages and subtracts this mean value from each of the inputs. It provides one-half of this difference at each corresponding output. For example, if e_k is the voltage input to the normalizing circuit from the kth filter channel of a group of N channels, then the normalized kth channel voltage is

$$e'_k = \tfrac{1}{2}\left[e_k - (1/N)\sum_{n=1}^{N} e_n\right]. \qquad (33)$$

This constraint on the mean value of the set of voltages to zero without altering the relative amplitudes permits reliable selection of the maximum voltage over a range of mean amplitude greater than 30 db.

The normalized set of voltages of any one group is sent to

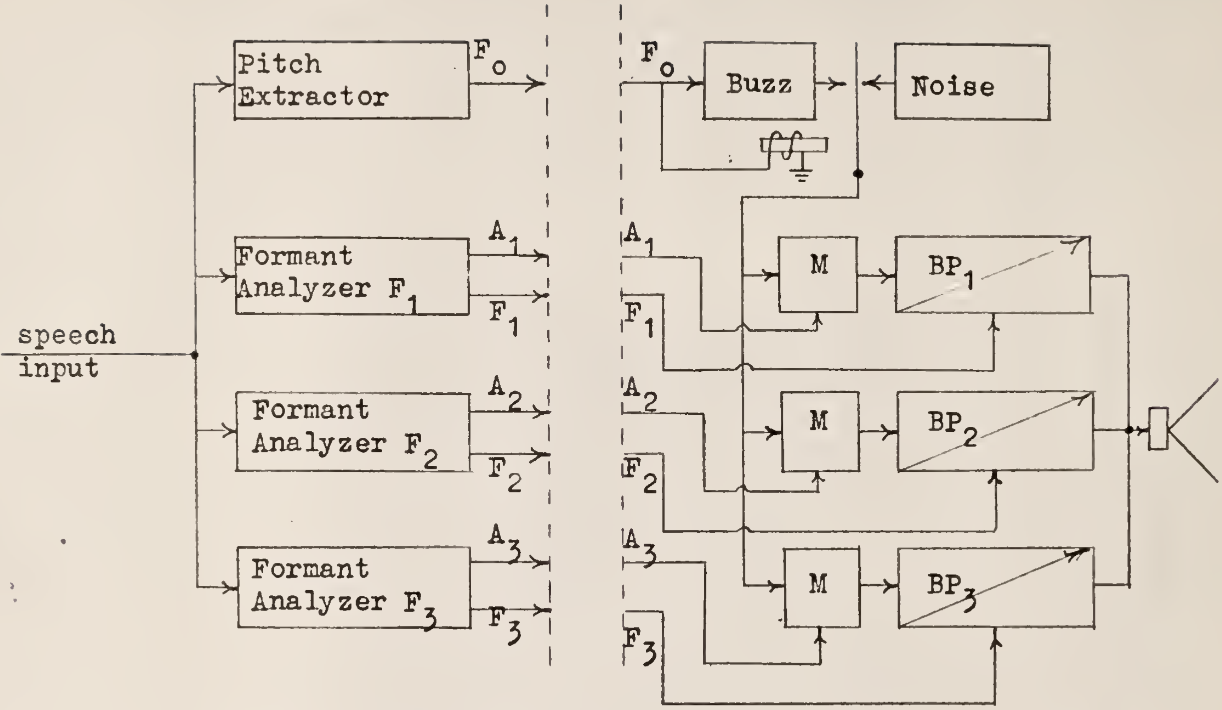

Fig. 19. Resonance vocoder (Flanagan, 1959).

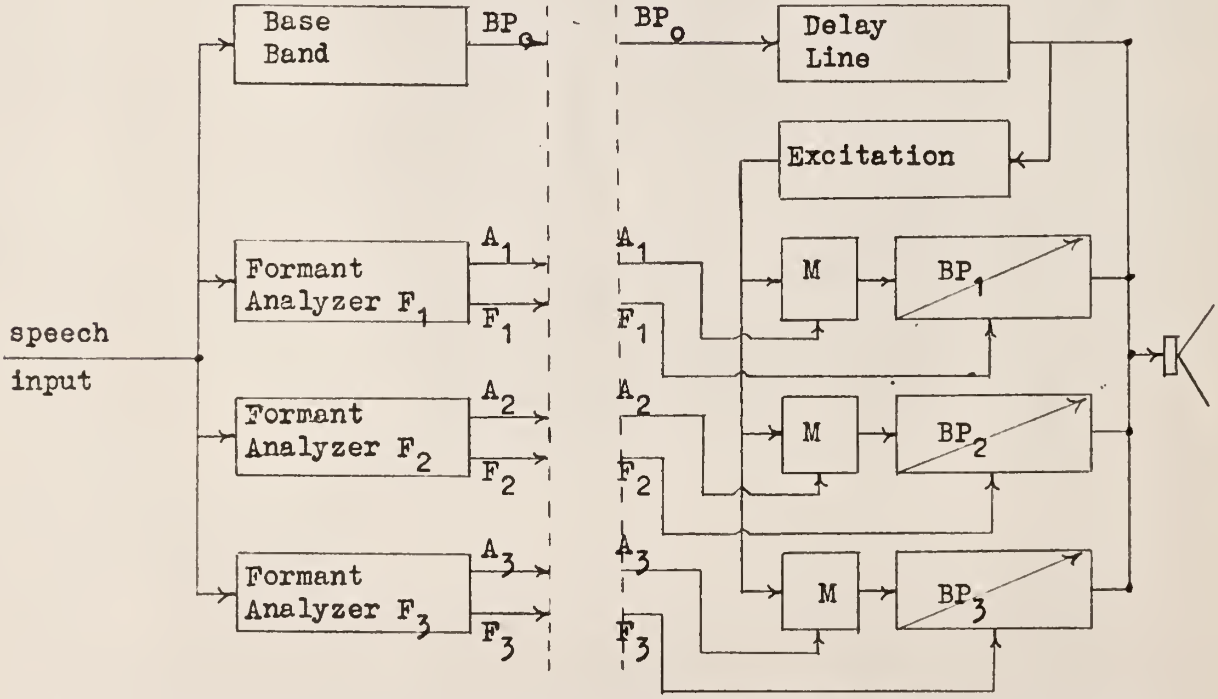

Fig. 20. Resonance vocoder and baseband complement (Flanagan, 1959).

the appropriate grids of a thyratron maximum-amplitude selector. The set of thyratron tubes has a common load resistor and the plate supply voltage is effectively turned on and off at a rate of 60 times per second. The thyratron having the highest positive grid voltage will fire first and preclude the firing of any other tube in the presence of the plate voltage. A potentiometer is connected as the cathode resistor of each thyratron and the output is taken from the arm of the potentiometer. The potentiometer is set so that the output voltage (when the tube fires) is proportional to the frequency of the channel that the tube is monitoring. All of the potentiometer arms are connected to a resistance adder to provide a single output from the selector. The output voltage from the selector is a series of rectangular pulses whose heights correspond to the frequency of the channel selected as having the maximum output.

The clamper is a filter which has an impulse response with Laplace transform

$$(1/s)(1 - e^{-sT}) \tag{34}$$

where T is the selecting or firing period of the selector. The selector output pulses are fed into the clamper for smoothing. A gate pulse is generated in the clamper by a one-shot multivibrator that is triggered each time a thyratron fires. The gate pulse samples the heights of the successive output pulses from the selector. This is the output of the formant extracting system (Flanagan, 1956). This system is very stable, and its calibration can be matched to essentially any single-valued function relating formant frequency and output voltage.

The resonance or formant vocoder (Fig. 19) uses the above described formant extractor as the vocoder analyzer. The excitation data are handled essentially in the same manner as in the channel vocoder. Two voltages are taken from the formant analyzers. One is proportional to the amplitude (A_i) of the spectral maximum, and the other to the frequency (F_i) of the maximum. The frequency voltages tune formant resonators, and the amplitude voltages modulate (see blocks M, Fig. 19) the inputs to the resonators.

The problems that plague the channel vocoder are also present in the resonance vocoder. Flanagan (1959) found the intelligibility of the resonance vocoder to be inferior to that of the channel vocoder. The split-band or hybrid idea again offered a practical compromise. Flanagan (1959) developed a resonance vocoder with a baseband complement (Fig. 20). The baseband covers the frequency range from 300 to 800 cps and is transmitted without further processing. Another band covers the range of 800 to 3200 cps and is transmitted by a resonance vocoder. Each amplitude and frequency signal is passed through a low-pass filter with a bandwidth of 15 cps and an 18 db per octave cut-off rate.

The baseband is delayed by 15 milliseconds at the synthesizer to equalize its delay to that of the processed band. The excitation is derived by means of nonlinear distortion of the baseband as discussed previously. The frequency response of the resonance vocoder and baseband complement is shown in Fig. 21.

An improvement in the system can be made by widening the baseband, or by increasing the number of vocoder channels for the higher band, but this results in a reduction of bandwidth compression.

The Harmoniphone

The harmoniphone, as shown in Fig. 22 (Pirogov, 1959), employs harmonic functions for the coding and synthesis of speech information. The analyzer at the transmitting end performs a Fourier analysis of the speech signal. Any analyzing filter set such as that described for the formant extractor (Flanagan, 1956) may be used. The spectrum is then sampled at a rate of 25 to 50 times per second, depending on the quality of speech required. Each sample K'(w) of the spectral speech function can be described by six to ten discrete levels, on the average, although a good reproduction of the vowels can be made with spectra with envelopes defined by three to five discrete levels only. This means that the form of a spectral function can be transmitted in the frequency band limited to third to fifth harmonics of the sampling frequency;

$$BW = n\ F_s = (3 \text{ to } 5)(25 \text{ to } 50) = 75 \text{ to } 250 \text{ cps} \qquad (35)$$

The same channel must also transmit pitch information which requires another band of 50 cps. The total frequency band of a harmoniphone telephone channel can be transmitted in a bandwidth of 100 to 300 cps, depending on the required quality of

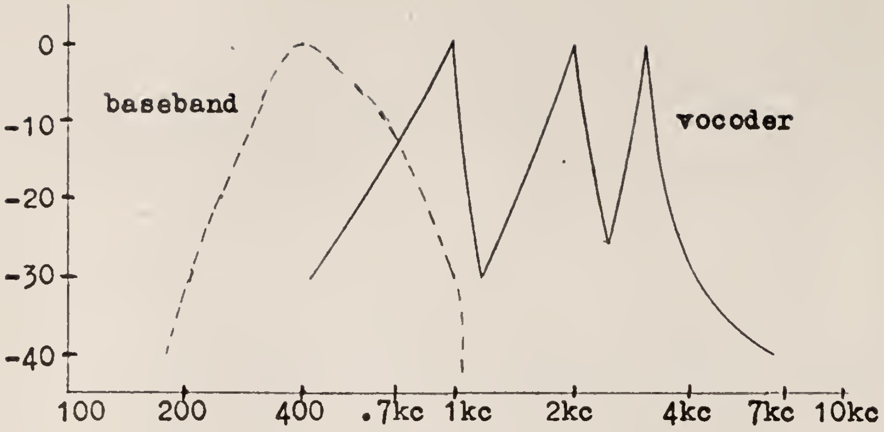

Fig. 21. Frequency response of resonance vocoder and baseband system (Flanagan, 1959).

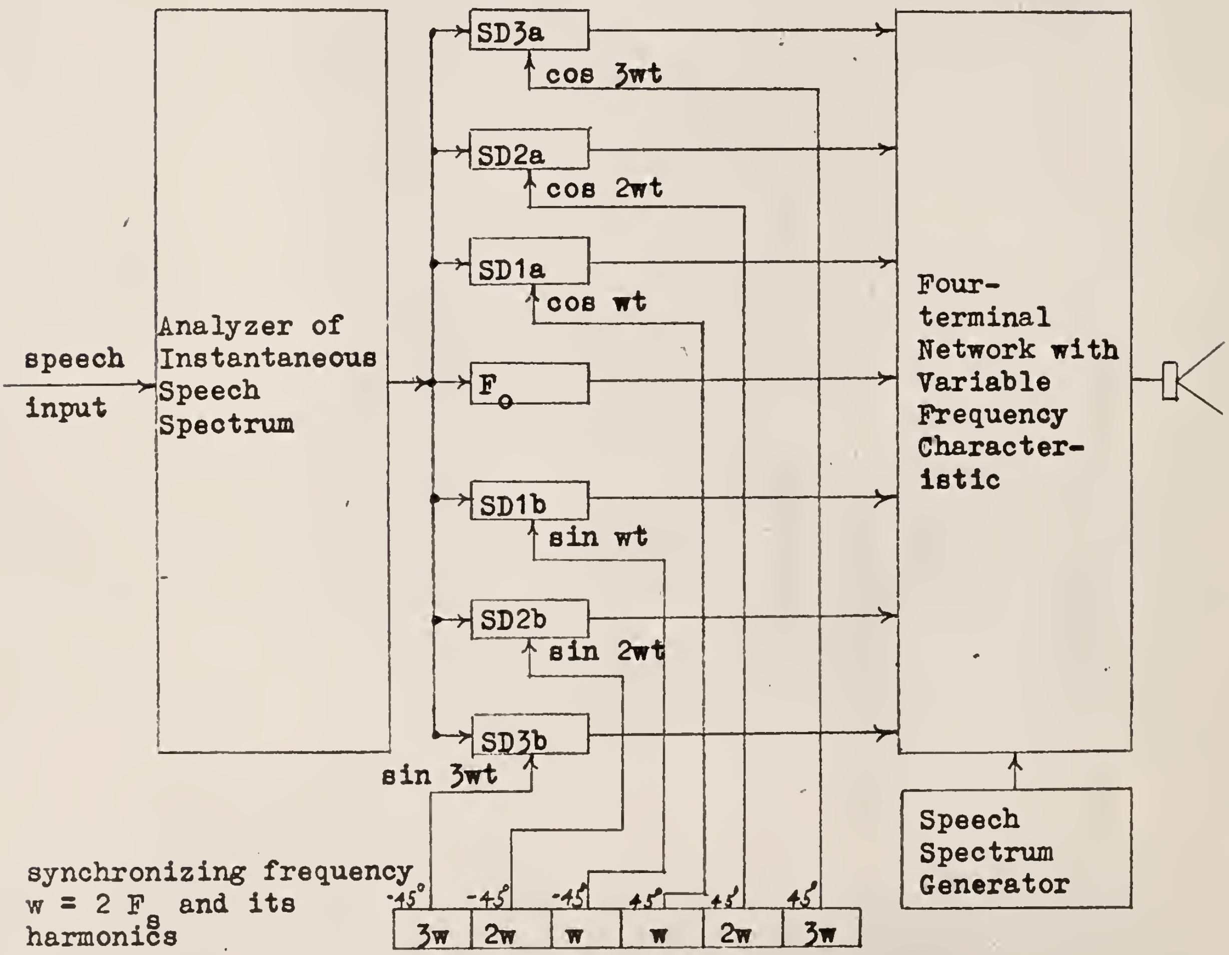

Fig. 22. The harmoniphone (Pirogov, 1959).

speech.

The analyzed signal, S(t) is applied to the synchronous detectors SD1a, SD2a ... SD1b, SD2b... and the filter of the constant component F_o(See Fig. 22). The synchronous detectors are controlled by the sampling frequency F_s and its harmonics, which are obtained from the synchronizer equipment in orthogonal phase relations (sin wt, cos wt, sin 2wt, and cos 2wt, etc. where $w = 2\pi F_s$). A synchronizer is generally used for many vocoders.

The synchronous detectors work into integrating circuits with outputs of

$$a_k = (2/T)\int_o^T S(t) \text{ cos kwt dt} \qquad (36)$$

and

$$b_k = (2/T)\int_o^T S(t) \text{ sin kwt dt} \qquad (37)$$

which are proportional to the coefficients of the Fourier series of S(t). When coefficients a_k and b_k are known it is possible to synthesize a four-terminal network whose frequency characteristic will, with an accuracy up to the highest harmonic of the Fourier analysis of S(t), correspond to the envelope of the shorttime spectrum K'(w), because the coefficients fully determine the shape of the frequency characteristic K(w).

A Chebyshev-Type Vocoder

A speech bandwidth compression system based on the transmission of a limited number of signal parameters proportional to the coefficients of expansion of the instantaneous pulse response

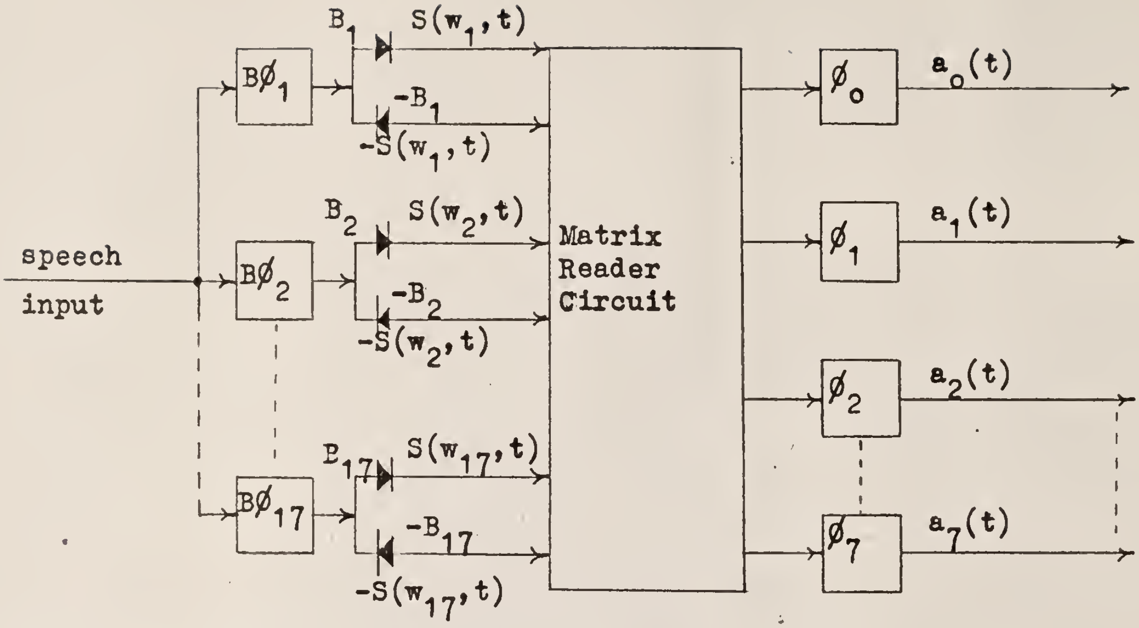

Fig. 23. Chebyshev vocoder analyzer (Kulya, 1963).

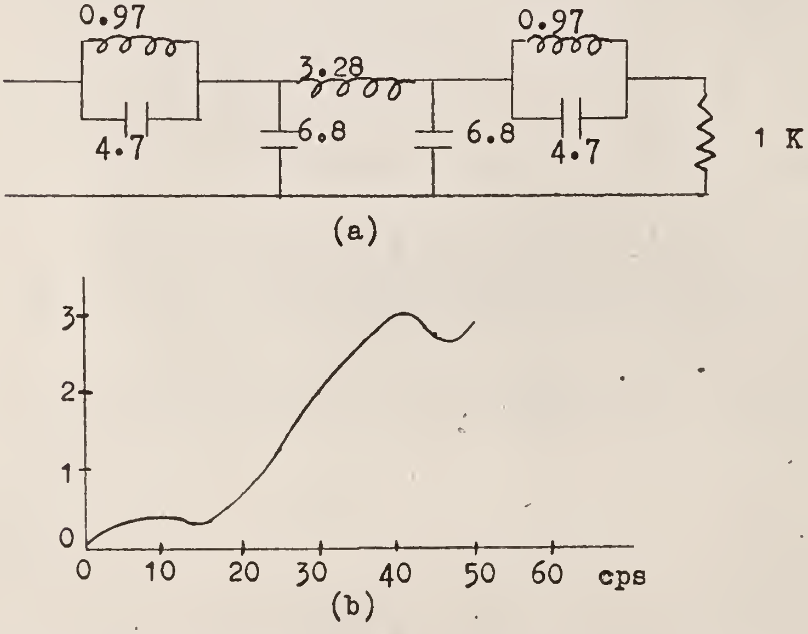

Fig. 24. Low-frequency filter ($Ø_n$) and characteristic (Kulya, 1963).

of the synthesizer in a series of weighted Laguerre polynomials was developed by Kulya (1963). The system, similar to the harmoniphone system, belongs to the class of orthogonal systems in which the transfer constants and the pulse response of the synthesizer are expressed as orthogonal functions. As a result, increasing the number of terms of the approximating series makes it possible to achieve a transmission of the approximated spectral function as accurately as desired. Furthermore, it is possible to retain any small number of the signal parameters without modifying any of the apparatus.

The application of Laguerre functions in an orthogonal vocoder leads to two desirable qualities. First, there is an improvement in the synthesized signal quality for a limited number of signal parameters due to the practical acceptance of nonuniform approximation to the frequency scale in the relationship between the instantaneous spectral intensity of the speech signal and frequency. The second merit of the system is the simplification of circuit solutions, and the reduction in size, weight, and cost of the apparatus due to the exclusion of low-frequency inductance coils.

If the envelope of the modulus of the speech signal spectrum is of the form S(w,t), then the pulse response of the synthesizer can be represented as

$$g(\tau,t) = \frac{1}{\pi} \int_{0}^{\infty} S(w,t) \cos w\tau \, dw \qquad (38)$$

and the signal parameters, which are proportional to the coefficients of their expansion into a series of Laguerre functions on the right semiaxis of time Υ are

$$A_n(t) = \int_0^\infty g(\Upsilon,t)L_n(\Upsilon)\, d\Upsilon = \int_0^\infty S(w,t)\psi_n(w)\,dw, \text{ for } \Upsilon \geq 0 \tag{39}$$

where

$$L_n(\Upsilon) = e^{\frac{1}{2}\lambda\Upsilon}\frac{d^n}{d\Upsilon^n}\left(\frac{\Upsilon^n}{n!}\quad e^{-\lambda\Upsilon}\right) \tag{40}$$

which are the orthogonal Laguerre functions of nth order for integral values of n.

The cosine transform of $L_n(\Upsilon)$ is

$$\psi_n(w) = \int_0^\infty L_n(\Upsilon)\ \cos w\Upsilon\ d\Upsilon = \frac{T_{2n+1}(\sqrt{1+4w^2/\lambda^2})}{\sqrt{1+4w^2/\lambda^2}} \tag{41}$$

where $T_{2n+1}(x)$ are the Chebyshev polynomials of the first kind and (2n + 1)th order. Equation 39 can be approximated by the finite sum

$$a_n(t) = \sum_{k=1}^{m} S(w_k,t)\psi_n(w_k)\Delta w_k \tag{42}$$

where $S(w_k,t)$ are the readings of the envelope of the instantaneous spectrum taken along the frequency axis for the values $w = w_k$, m is number of readings, and $\Delta w_k = w_{k+1} - w_k$.

The signal parameters $a_n(t)$, Equation 42, are obtained from the speech signal with the aid of the analyzer shown in Fig. 23. The readings $S(w_k,t)$ of the approximate amplitude of the envelope of the instantaneous speech spectrum are obtained at the outputs of the 17 spectral channel filters $BØ_1$ to $BØ_{17}$ and amplifier-rectifiers B_1 to B_{17}. Since the functions $\psi_n(w)$ are of dual polarity, the amplifier-rectifiers should provide both positive and negative output voltages. The output voltages of each of the spectral channels are summed at each of the eight outputs $a_o(t)$, $a_1(t)$, ... $a_7(t)$ with coefficients proportional to the readings of the corresponding functions $\psi_n(w)$ at the points $w = w_k$. This conversion is achieved by means of a resistive matrix reader (Kulya, 1963).

The signal parameters obtained are smoothed out by low-frequency filters $\emptyset_0$ to $\emptyset_7$ with a pass-band of 25 cps. The circuit and typical frequency characteristic of a filter are shown in Fig. 24.

The instantaneous pulse response of the synthesizer as a Fourier transformation of S(w,t) should be of the form

$$g(t, |\tau|) = \sum_{n=0}^{7} a_n(t) \; L_n(|\tau|). \tag{43}$$

A response of such a form is not physically realizable since it extends throughout the entire time axis, $-\infty \leq \tau \leq \infty$. The synthesizer pulse response may be expressed in the following form instead of as in Equation 43:

$$g(t, \tau) \quad \sum_{n=0}^{7} a_n(t) \left[L_{m-n}(\tau) + L_{m+n+1}(\tau)\right] \tag{44}$$

for $\tau \geq 0$ which is physically realizable. The corresponding circuit is shown in Fig. 25. The response of the cascaded "a" and the n "b" sections is in the form of a Laguerre function of nth order.

To determine the optimum number of terms to be retained in Equation 43, the following relationship can be developed:

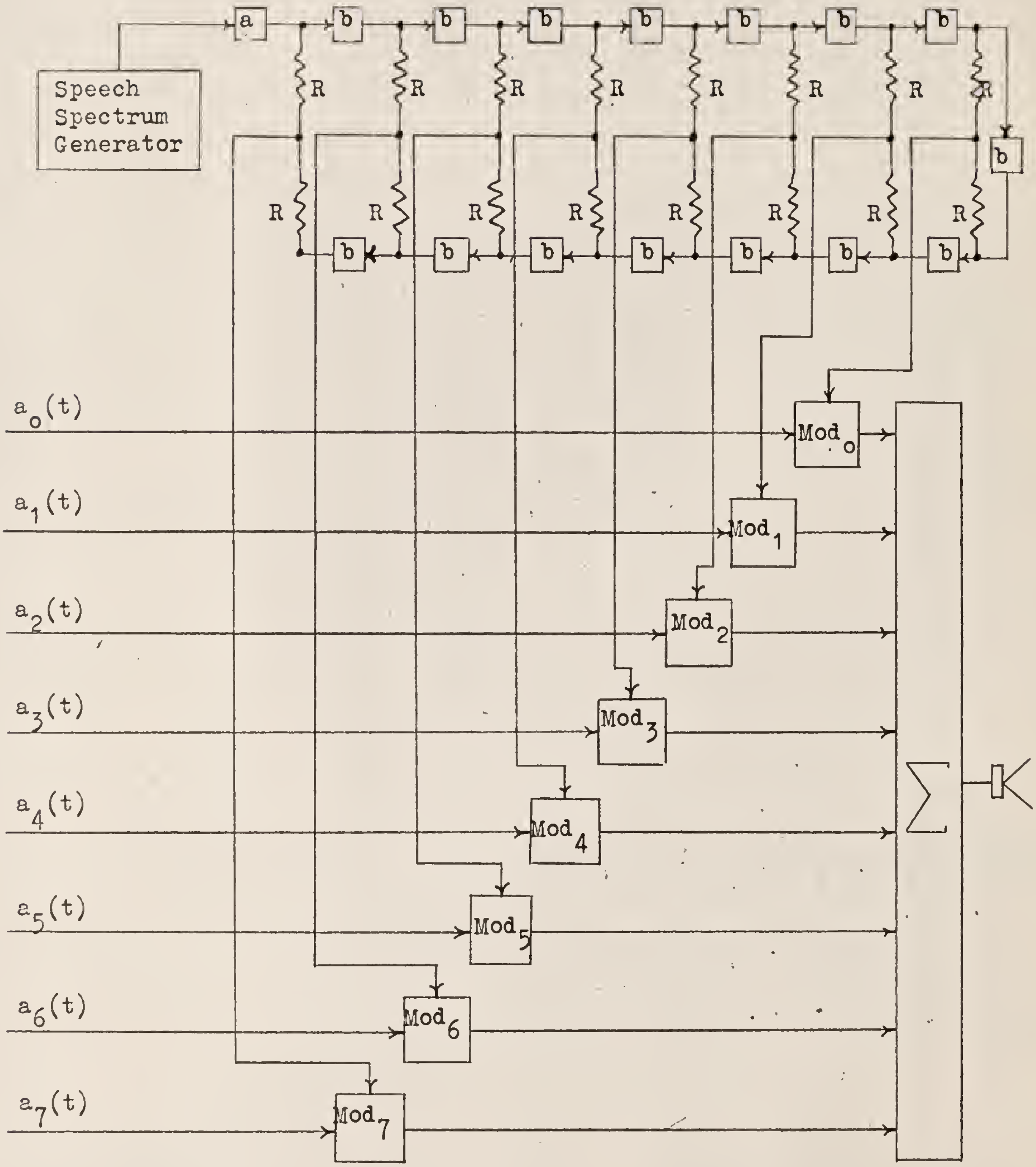

Fig. 25. Chebyshev vocoder synthesizer (Kulya, 1963).

$$S(w,t) = \frac{e^{-2j \arctan 2w/\lambda}}{\sqrt{1 + 4w^2/\lambda^2}} \sum_{n=0}^{7} a_n(t) T_{2n+1}(\sqrt{1 + 4w^2/\lambda^2})$$

$$= \frac{e^{-2j \arctan 2w/\lambda}}{\sqrt{1 + 4w^2/\lambda^2}} \sum_{n=0}^{7} a_n(t) \emptyset_n(w/\lambda) \tag{45}$$

Each term under the summation sign in Equation 45 is an oscillating function with a number of zeros, on the w-axis, equal to the order n of the function. The zeros are not equally spaced and, for the finite expression, the accuracy of approximation of the function S(w,t) decreases with an increase in frequency w. A measure of this change of accuracy may be obtained from the relationship

$$\theta = \arctan 2w/\lambda \tag{46}$$

where θ is in degrees.

Kulya found the optimum value of λ to be $\lambda = 5.34\pi \times 10^3$ and the required bandwidth proportional to the frequency. The bandwidth required for transmission is 175 cps plus the guard space. Assuming a guard space of 10 cps the total bandwidth would be 255 cps, or a compression ratio of 12:1 with an articulation efficiency of approximately 85%.

DISCRETE SOUND ANALYSIS-SYNTHESIS METHODS

Phonetic Pattern Recognition Vocoder

The phonetic pattern recognition vocoder (Dudley, 1958), as shown in Fig. 26, compares observed phonetic patterns with stored standard patterns and transmit sufficient information to synthesize the recognized pattern. The speech is filtered by a set of ten band-pass filters each 300 cps wide except the first which is 250 cps. Each band-pass circuit contains an amplifier, a rectifier, and a low-pass filter to smooth the speech power to a syllabic rate. An amplifier gain adjustment is used as a convenient means of frequency equalization. The output of the band-pass circuit is fed to the memory circuit, which consists of a 10 x 10 matrix of potentiometers. These are set to yield the spectral frequency pattern for the sound i, as in seat, in the first-row, I as in sit, in the second row, and so on to the tenth row for s. The potentiometer settings are determined from measurements of the band-pass circuit outputs as sounds are sustained. A relay, not shown, set to operate slightly above the noise level in each band, feeds a capacitor through high resistance. As the reference sounds are spoken (sustained) by the speaker, the charges on the capacitors build up according to the spectral pattern, at the conclusion of the sound, the relays open, holding the capacitor voltages for measurement. The voltage data thus obtained is normalized so that the smallest voltage is reduced to zero. The voltage settings for the matrix

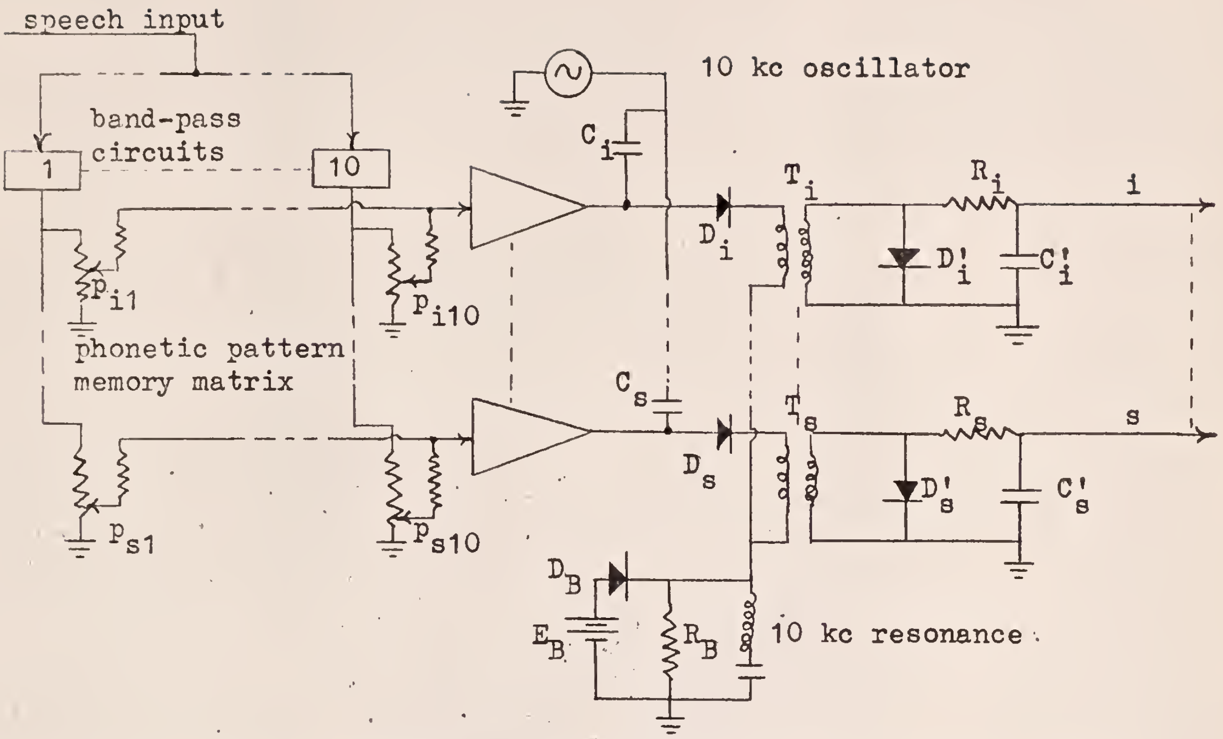

Fig. 26. Phonetic pattern recognizer (Dudley and Balashek, 1958).

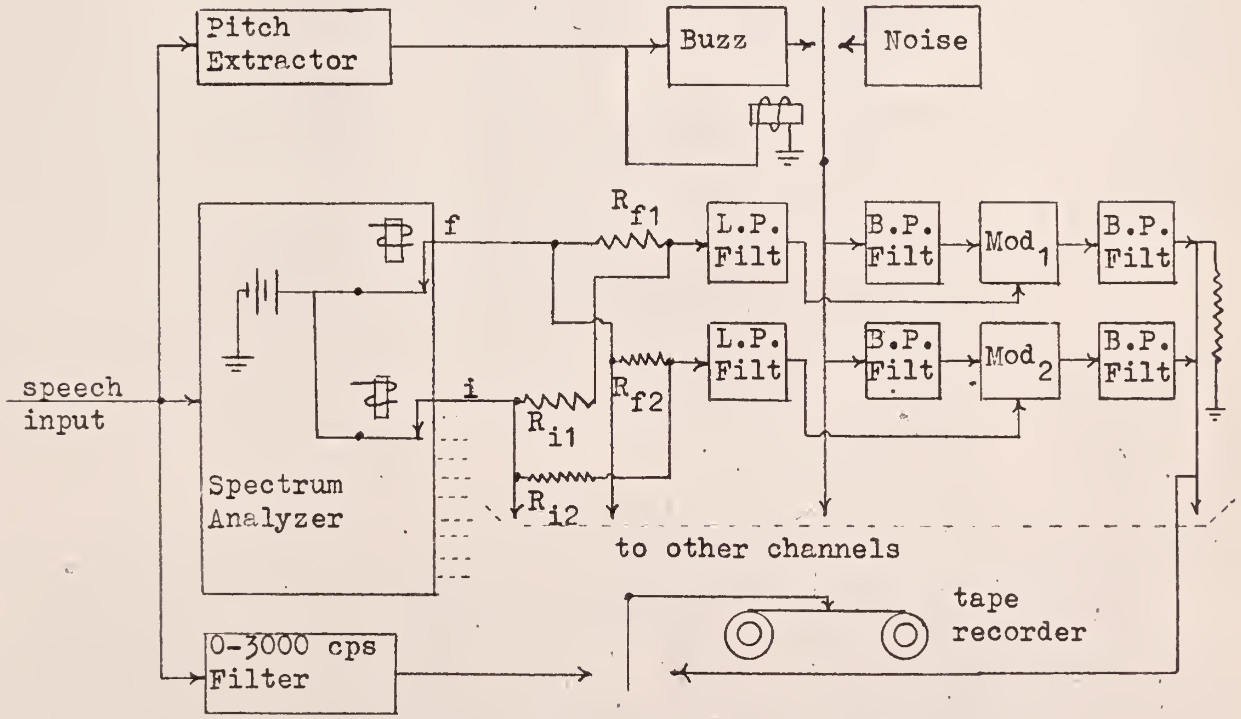

Fig. 27. Phonetic pattern recognition vocoder (Dudley, 1958).

are listed in Table IV. The listing for row p and column b are normalized voltages from the matrix potentiometer output, in the pth row of the bth column (band-pass circuit) as the sound prototype for the pth phonetic pattern sustained long enough to give a set of reasonably large voltages. The ratios of these values to 150v are the matrix transfer ratios.

The output voltage v_{ps} from the matrix for the pth phonetic pattern branch at any instant as the sth speech sound is spoken is given by the summation

$$v_{ps} = \sum_{b=1}^{10} r_{pb}\, v_{sb} \qquad (47)$$

where v_{sb} is the instantaneous smoothed rectified voltage output of the bth band-pass circuit as the sth sound is spoken, and r_{pb} is the voltage transfer ratio for the potentiometer setting of the bth band-pass circuit for the pth phonetic pattern as given in Table IV (after normalization). At any instant, for one value of p, v_{ps} will have a maximum value,

$$v_{Ps} = \text{largest } v_{ps}, \text{ for } p = P. \qquad (48)$$

The Pth phonetic pattern would then be the pattern assigned by the apparatus to that portion of the sth sound. In general, memory pattern P would have the best spectral match at that instant with the portion of sth sound being spoken.

The actual selection of phonetic pattern for best achievable instantaneous match takes place to the right of the resistance matrix in Fig. 26. The ten voltages v_{ps} from the phonetic pattern matrix are amplified by the individual buffer

TABLE IV. Voltage Settings for Phonetic Pattern Matching (Dudley, 1958).

	Band-pass Filter Number									
	1	2	3	4	5	6	7	8	9	10
	Mid-Band Frequencies - (cps)									
	125	400	700	1000	1300	1600	1900	2200	2500	2800
Phonetic Pattern	Voltages - Volts									
i	11.4	2.1	0.3	0.0	0.3	0.3	0.4	3.7	5.0	9.6
I	5.9	10.2	1.4	0.7	0.7	1.4	4.7	4.7	7.1	5.2
E	4.3	5.3	5.1	1.4	1.0	3.1	9.9	4.7	6.6	4.1
a	2.7	2.5	5.2	8.9	10.0	3.3	2.7	4.8	1.0	1.7
o	4.6	7.2	9.7	9.4	1.5	0.8	1.3	2.2	0.5	1.1
u	13.0	9.3	2.9	0.6	0.2	0.0	0.0	0.3	0.3	0.6
n	14.0	4.3	0.8	1.2	4.0	3.0	1.3	3.0	3.3	1.7
r	3.5	6.0	6.9	1.6	5.8	9.5	6.0	2.2	0.9	1.5
f	0.5	0.6	0.6	1.7	2.4	3.5	3.0	4.1	3.9	14.3
s	0.4	0.8	0.4	0.9	1.1	2.3	2.3	2.6	4.5	15.0

amplifiers, whose outputs are connected through biasing diodes D_p and transformer winding T_p to ground via the 10kc series resonant circuit. A 10kc oscillator supplies the sensing voltage through capacitors C_i and C_s for all of the matrix outputs. A voltage-biasing circuit attached to the 10kc resonant circuit provides a minimum threshold to prevent operation by noise. When any branch voltage v_{ps} become large enough to overcome this bias, it will pass dc through its own branch diodes D, and lower its resistance so that 10kc current flows through the transformer winding T_p. The dc current adds bias in the same direction as the original bias from noise-bias battery E_B, increasing the bias against current from any of the other branches. Thus, the selection is completed with the 10kc current transmitted only in the branch having the strongest signal. The resulting 10kc pattern recognition current is rectified by diode D_p' and then after smoothing is fed to the synthesizer control circuit (Dudley 1958).

A phonetic pattern recognition vocoder is shown in Fig. 27. The direct speech is recorded through a 3 kc low-pass filter for later comparison with the speech produced by the vocoder. The analyzer has devices for both pitch and spectrum determination. The pitch circuit is the same as that used in the channel vocoder. The spectrum analyzer is the phonetic pattern recognizer described above. The analyzer recognizes only ten patterns corresponding to four consonants and vowels.

A set of ten resistors for each pattern recognized control a 10 channel vocoder synthesizer at the receiver. Channel resistors are so chosen as to provide proper amounts of current for each pattern recognized. Each set of resistances is followed by a 25 cps low-pass filter so that the current to the synthesizing modulators passes from one value to another smoothly. The fixed voltage for each pattern recognized allocates energy source via the ten resistors in the set for that pattern in such a way that the appropriate spectrum is produced in the vocoder output.

If the recognition process could be made to approach the human facility, the number of phonetic patterns would be roughly comparable to the number of alphabetic characters used in telegraphy. In other words, if only 32 characters need be transmitted to give the 26 letters of the alphabet, then for the spoken word as well only 32 characters are required. That is, five bits of information or at the most 64 characters or six bits, need to be transmitted provided the information is limited to the sort produced in the written word.

The phonetic pattern recognition vocoder has several shortcomings. There are, chiefly, limited number of patterns available and no adequate provision for recognition of patterns where the change of power with time is an essential characteristic, as in "plosives" (t,b,p ---).

If the pattern recognition were perfect, the required bandwidth for 32 characters would be

$$B = \tfrac{1}{2}\ ICW = \tfrac{1}{2} \times 5 \times 5 \times 4 = 50 \text{ cps} \tag{49}$$

where I is the number of "Nyquist" time intervals per character, C is the average number of characters per word, and W is the average number of words per second; with I = 5, C = 5 characters per word, and W = 4 words per second.

However, the recognition is not perfect, and the required bandwidth for a 16 character alphabet was found to be approximately 100 cps (Dudley, 1958) and a resulting bandwidth reduction factor of 30:1.

SOUND GROUP ANALYSIS-SYNTHESIS METHODS

A system which automatically recognizes entire words, transmits a unique code for each word, and at the receiver converts the word code to synthetic speech may offer a means of transmission at extremely low information rates for a limited vocabulary. For example, a vocabulary of 32 words can be transmitted at information rates of 10 bits per second. A literature search shows that no attempt has been made to produce a complete system, but both analyzers and synthesizers have been investigated separately.

Work was done on an analyzer (Kock, 1956 and Dudley, 1958) which would automatically recognize spoken digits. The device analyzes the speech input to determine which sound in its memory is most similar to the observed sound. First it breaks the spoken digit into a series of sound identifications and then by comparison determines which of the ten digits in its memory has the same sequence. It can recognize the digits as spoken by the voice for which the system is calibrated, but fails to perform satisfactorily for any other voice, or even a change in the manner of speaking for the calibrating voice.

Magnetic tape playbacks operated from a source of digital information provide a satisfactory device for the synthesis of complete words. The Automatic Voice Readout System (AVRS) (Poppe and Suhr, 1957) shown in Fig. 28 is an example of such a device. The AVRS is intended for use as a digital code to voice converter to read out commands from a digital computer vocally.

Five digit control signals required to control a relay pyramid are supplied from a coding unit and synchronizer. The output of the relay pyramid drives an audio amplifier. The synchronizer serves as the basic timing within the system. Synchronization pulses are received every one-half second, and are used to advance the synchronizer to the next word interval. The function of the coding unit is to convert the input code to a five digit parallel code required to drive the relay pyramid.

A system consisting of a limited vocabulary coding analyzer and a limited vocabulary decoding synthesizer such as the AVRS would constitute an operable voice communication system of very low information rate. Although such a system has a very limited vocabulary, there are certain situations such as aircraft traffic control, where such a limited vocabulary would be entirely satisfactory. A diagram of such a system is shown in Fig. 29.

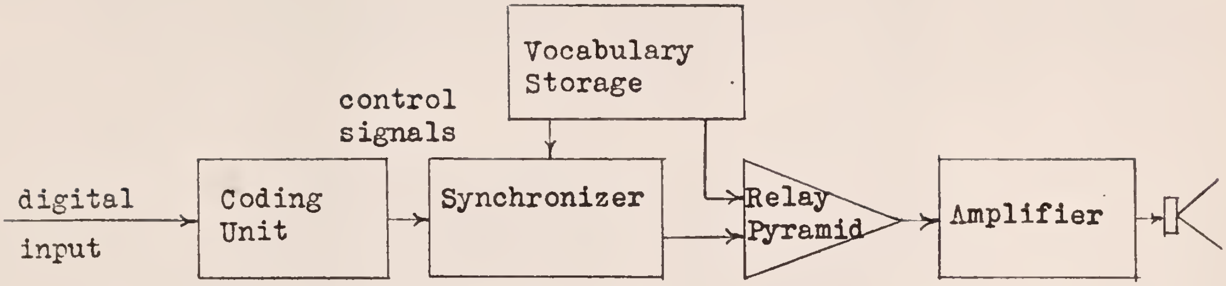

Fig. 28. Automatic voice readout system (AVRS) (Poppe and Suhr, 1957).

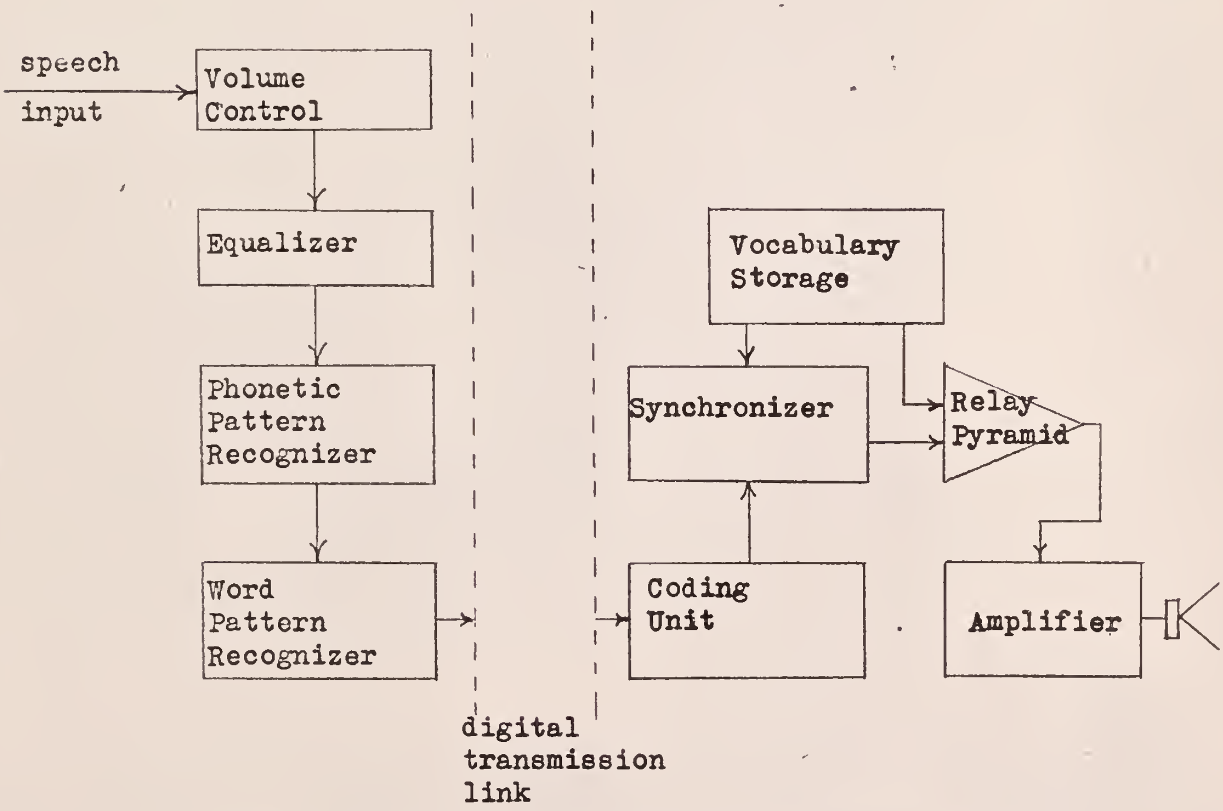

Fig. 29. Complete sound group analysis-synthesis system (Dudley and Balashek, 1958, and Poppe and Suhr, 1957).

CONCLUSIONS

The problem of speech bandwidth compression centers around the question as to what is the essential information contained in the speech spectrum that must be transmitted in order that intelligible speech may be reproduced; and as to what is the most reliable means of accomplishing it. Bandwidth compression techniques may be grouped into four general categories, namely, time and frequency compression methods, continuous analysis-synthesis methods, discrete sound analysis-synthesis methods, and sound group analysis-synthesis methods. All of these techniques attempt to eliminate the non-essential information for the purpose of the transmission.

The envelope of the power spectrum carries the essential information, and therefore, it is desired to reconstruct the power spectrum envelope as nearly as possible from the limited transmitted information produced by the analyzer of such a system. Some techniques produce good quality speech but exhibit very modest bandwidth reduction, such as the Vobanc (2:1) and Codimex (4:1) systems. Other techniques, such as phonetic pattern recognition vocoders, channel vocoders, and resonance vocoders, achieve high bandwidth reduction factors of approximately 30:1, 20:1, and 10:1 respectively, but lack in quality of the reproduced speech. This loss of quality is caused by errors in the pitch extraction and voiced-unvoiced decisions. Also, since the synthesizers can only produce either a continuous spectrum or a discrete spectrum, no semi-vowels can be produced by systems employing such excitation. This difficulty can be overcome, to a great extent, by employing a baseband of the original speech and deriving the excitation from the baseband. Such a device is called a voice-excited vocoder. Such vocoders reproduce higher quality speech than their counterparte, but the bandwidth reduction (4 to 5:1) suffers due to the bandwidth required for the transmission of the baseband. In general, the quality of the speech has little effect on the intelligibility of the speech, but any attempt to improve the quality results in less bandwidth reduction. The autocorrelation vocoder seems to be the only system that produces both reasonable quality and reasonable bandwidth reduction (10:1). This vocoder also experiences the pitch and excitation difficulties, but performs creditably in spite of them.

Discrete sound analysis-synthesis methods and sound group analysis-synthesis methods are capable of very low information rates of transmission of 60 bits per second and 5 bits per second, respectively, but have several serious drawbacks. First of all, in order to use such systems in general speech communication links, a very large vocabulary would be required. Thus, situations where limited vocabularies are used are the more likely possibilities for these systems.

These constitute the major bandwidth compression techniques.

ACKNOWLEDGMENT

The author wishes to express his thanks to Dr. H. S. Hayre for his guidance and helpful suggestions in the preparation of this report.

BIBLIOGRAPHY

Beddoes, M. P.
A Slope Feedback Method for Speech Compression. IRE Transactions on Communications Systems. Vol. C.S. 8, No. 4. Dec. 1960. 254.

Billings, A. R. - Lloyd, P. J.
Correlator Employing Hall Multipliers Applied to Analysis of Vocoder Control Signals. Institution of Electrical Engineers Proceedings. Vol. 107. Part B, No. 35. Sept. 1960. 435-438.

Billings, A. R.
Simple Multiplex Vocoder. Electronic and Radio Engineer. Vol. 36, No. 5. May 1959. 184-188.

Bogert, B. P.
Journal of Acoustical Society of America. Vol. 28. May 1956. 399-404.

Campanella, S. J.
A Survey of Speech Bandwidth Compression Techniques. IEEE Transaction on Audio. Au 6, No. 5. Sept. Oct. 1958. 104-116.

Daguet, J. L.
Speech Compression Codimex System. IEEE Transactions on Audio. Vol. Au 11, No. 2. 1963. 63-71.

Das, J.
Bandwidth Compression of Speech. Electronic Technology. Vol. 38, No. 8. Aug. 1961. 298-300.

Dudley, H.
Phonetic Patterns Recognition Vocoder For Narrow-Band Transmission. Journal of Acoustical Society of America. Vol. 30. Aug. 1958. 733.

Dudley, H. - Balashek, S.
Automatic Recognition of Phonetic Patterns. Journal of Acoustical Society of America. Vol. 30, No. 8. Aug. 1958. 721-732.

Fano, R. M.
Short Time Autocorrelation Function and Power Spectra. Journal of Acoustical Society of America. Vol. 22. 1950. 546.

Flanagan, J. L.
Automatic Extraction of Formant Frequencies From Continuous Speech. Journal of Acoustical Society of America. Vol. 28. Jan. 1956. 110-118.

Flanagan, J. L.
Bandwidth and Channel Capacity Necessary to Transmit The Formant Information of Speech. Journal of Acoustical Society of America. Vol. 28. July 1956. 592-596.

Flanagan, J. L.
Resonance Vocoder and Baseband Complement: Hybrid Speech Transmission. IRE WESCON Convention Record. Part 7. 1959. 5-16.

Gabor, D.
Theory of Communication. Journal of IEE. Part 3 G.B. Vol. 93. Nov. 1945. 429-457.

Gill, J. S.
A Study of Requirements For Excitation Control In Synthetic Speech. Third ICA Congress. 1959. 221-224.

Kharkevich, A. A.
Possibilities of Spectrum Compression. Telecommunications. Nov. 1958. 1121-1128.

Kock, W. E.
Speech Bandwidth Compression. Bell Labs Reports. March 1956. 81-85.

Kryter, K. D.
Speech Bandwidth Compression Through Spectrum Selection. Journal of Acoustical Society of America. Vol. 32. May 1960. 547-556.

Kulya, V. I.
Application of Laguerre Functions To Parametric Coding Of Speech Signals. Telecommunications. No. 7. 1962. 34-47.

Kulya, V. I.
Analysis of A Chebyshev-Type Vocoder. Telecommunications. No. 3. March 1963. 23-36.

Pirogov, A. A.
Harmonic System For Compressing Speech Spectra. Telecommunications. No. 3. 1959. 229-242.

Poppe, C. W. - Suhr, P. J.
An Automatic Voice Readout System. Eastern Joint Computer Conference Proceedings. Dec. 1957. 219-221.

Schroeder, M. R.
Recent Progress in Speech Coding At Bell Labs. Proceedings of The Third International Congress On Acoustics. Vol. 1. 1959. 201-210.

Schroeder, M. R. - David, E. E. Jr.
A Vocoder For Transmitting 10Kc/s Speech Over A 3.5Kc/s Channel. Acoustica. Vol. 10, No. 1. 1960. 35-43.

Schroeder, M. R.
Correlation Techniques for Speech Bandwidth Compression. Journal of Audio Engineering Society. Vol. 10. No. 2. April 1962. 163-166.

Schroeder, M. R. - David, E. E. Jr. - Logan, B. F. - Prestigiacomo, A. J.
Voice-Excited Vocoders For Practical Speech Bandwidth Reduction. IRE Transactions On Information Theory. Vol. IT8, No. 5. Sept. 1962. S-101- S105.

Slaymaker, F. H.
Bandwidth Compression By Means of Vocoders. IRE Transactions On Audio. Vol. Au8. Jan-Feb. 1960. 20-26.

Subrahmanyam, D. L. - Peterson, G. E.
Time-Frequency Scanning In Narrow-Band Speech Transmission. IRE Transactions On Audio. Vol. Au7. No. 6. Nov.-Dec. 1959. 148-160.

Ville, J.
Theorie De La Notion Signal Analytique. Cables Et Transmission. Vol. 1. Jan. 1948. 61-74.

SPEECH BANDWIDTH COMPRESSION

by

LEON ALVIN HOLLOWAY

B. S., Kansas State University, 1963

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1964

ABSTRACT

Application of speech bandwidth compression techniques to voice communications provides the promise of more efficient use of the radio spectrum and improved performance of noisy, long distance communications links. Proof of the need for bandwidth compression can be found in the fact that conventional speech transmission requires a transmission rate of approximately 24,000 bits per second, while the same information transmitted by teletype requires a rate of only 75 bits per second. This need for conservation of communication space brought about the development of various techniques of bandwidth compression which is the subject of this report.

There is a considerable difference between the information rate necessary to communicate the speech signal in the conventional manner and the actual rate which information appears to be generated by the vocal mechanism. The additional information rate is manifested in the identity of the speaker and his emotional status, redundancy, and inefficient use of the spectrum. In general, all speech bandwidth compression systems attempt to exploit one or more of these factors to obtain a reduction in the bandwidth and thus the required channel capacity.

The bandwidth compression techniques may be grouped into four general categories:

1) Time and frequency compression methods.

Such methods exploit the redundancy or regularities existing in the speech signal by sampling and frequency division techniques.

These systems generally exhibit bandwidth compression in the order 2:1 to 10:1 and can be transmitted in binary code form over channels of 2,000 to 10,000 bits per second capacity.

2) Continuous analysis-synthesis methods.

In place of the speech signal spectrum such methods transmit a description of the spectrum in terms of a number of analog parametric control signals. As such, they exploit both the redundancy and inefficiency existing in the speech signal. In general, these systems exhibit bandwidth compression in the order of 10:1 to 20:1 with a required channel capacity of 1,000 to 2,000 bits per second.

3) Discrete sound analysis-synthesis methods.

Such methods transmit in place of the speech signal code groups which identify the fundamental sounds that constitute the speech. As such, they exploit the redundancy and inefficiency of the speech signal, and remove the identity and emotional status cues. It is expected that such systems should be capable of transmitting speech at information rates as low as 60 bits per second.

4) Sound group analysis-synthesis methods.

Such methods transmit only certain groups of sounds (particular words and phrases), each identified by a code group. Information rates in this case are a function of the size of the vocabulary. Such a system appears to be most useful at information rates in the order of 5 to 10 bits per second.

Although some of the methods achieve very high compression ratios, their articulation efficiency, in some cases, may be too low for practical use. In general, bandwidth compression must be sacrificed to improve the articulation efficiency.