Phase Vocoder

By J. L. FLANAGAN and R. M. GOLDEN

(Manuscript received July 18, 1966)

A vocoder technique is described in which speech signals are represented by their short-time phase and amplitude spectra. A complete transmission system utilizing this approach is simulated on a digital computer. The encoding method leads to an economy in transmission bandwidth and to a means for time compression and expansion of speech signals.

I. INTRODUCTION

Analysis-synthesis methods for speech transmission aim at efficient encoding of voice signals. A customary approach is to represent separately the important features of vocal excitation and tract transmission.[1] The well-known channel vocoder of Dudley[2] derives signals which fall into this dichotomy. The tract transmission is described by values of the short-time amplitude spectrum measured at discrete frequencies, and the excitation is described in terms of the fundamental frequency of the voice and the voiced-unvoiced character of the signal. Efforts to solve the long-standing problem of good-quality synthesis from such representations have centered on adequate analysis and specification of the excitation data.

One advance in surmounting the difficulties connected with pitch and voiced-unvoiced extraction is the voice-excited vocoder (VEV).[3] This device relys on transmission of an unprocessed subband of the original speech to carry the excitation information. The spectral envelope information is transmitted as in the channel vocoder by a number of slowly-varying signals. Through accurate preservation of excitation details, a transmission of improved quality and modest bandsaving is achieved.

The present paper proposes another technique for encoding speech to achieve comparable bandsaving and acceptable voice quality. In addition, the technique provides a convenient means for compression and expansion of the time dimension. The method specifies the speech signal in terms of its short-time amplitude and phase spectra. For this reason, it is called phase vocoder. Like the VEV, the phase vocoder does not

require the pitch tracking and voiced-unvoiced switching inherent in conventional channel vocoders. Elimination of these decision-making processes and the transmission of excitation information by phase-derivative signals contribute to improved quality in the synthesized signal.

II. PRINCIPLES

If a speech signal $f(t)$ is passed through a parallel bank of contiguous band-pass filters and then recombined, the signal is not substantially degraded. The operation is illustrated in Fig. 1, where $BP_1 - - - BP_N$ represent the contiguous filters. The filters are assumed to have relatively flat amplitude and linear phase characteristics in their pass bands. The output of the nth filter is $f_n(t)$, and the original signal is approximated as

$$f(t) \cong \sum_{n=1}^{N} f_n(t). \tag{1}$$

Let the impulse response of the nth filter be

$$g_n(t) = h(t) \cos \omega_n t, \tag{2}$$

where the envelope function $h(t)$ is normally the impulse response of a physically-realizable low-pass filter. Then the output of the nth filter is the convolution of $f(t)$ with $g_n(t)$,

$$\begin{aligned} f_n(t) &= \int_{-\infty}^{t} f(\lambda) h(t - \lambda) \cos [\omega_n (t - \lambda)] d\lambda \\ &= \text{Re} \left[\exp (j\omega_n t) \int_{-\infty}^{t} f(\lambda) h(t - \lambda) \exp (-j\omega_n \lambda) d\lambda \right]. \end{aligned} \tag{3}$$

The latter integral is a short-time Fourier transform of the input signal $f(t)$, evaluated at radian frequency ω_n . It is the Fourier transform of that part of $f(t)$ which is "viewed" through the sliding time aperture

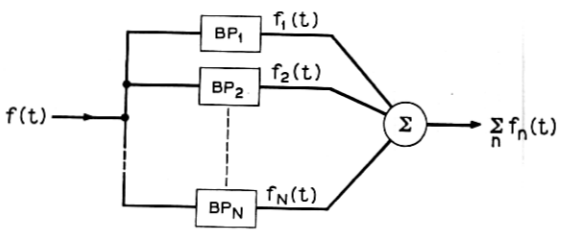

Fig. 1 — Filtering of speech by contiguous band-pass filters.

$h(t)$. If we denote the complex value of this transform as $F(\omega_n, t)$, its magnitude is the short-time amplitude spectrum $|F(\omega_n, t)|$, and its angle is the short-time phase spectrum $\varphi(\omega_n, t)$. Then

$$f_n(t) = \text{Re}[\exp(j\omega_n t)F(\omega_n, t)]$$

or

$$f_n(t) = |F(\omega_n, t)| \cos[\omega_n t + \varphi(\omega_n, t)]. \quad (4)$$

Each $f_n(t)$ may, therefore, be described as the simultaneous amplitude and phase modulation of a carrier ($\cos \omega_n t$) by the short-time amplitude and phase spectra of $f(t)$, both evaluated at frequency ω_n.

Experience with channel vocoders shows that the magnitude functions $|F(\omega_n, t)|$ may be band-limited to around 20 to 30 Hz without substantial loss of perceptually-significant detail. The phase functions $\varphi(\omega_n, t)$, however, are generally not bounded; hence they are unsuitable as transmission parameters. Their time derivatives $\dot{\varphi}(\omega_n, t)$, on the other hand, are more well-behaved, and we speculate that they may be band-limited and used to advantage in transmission. To within an additive constant, the phase functions can be recovered from the integrated (accumulated) values of the derivatives. One practical approximation to $f_n(t)$ is, therefore,

$$\tilde{f}_n(t) = |F(\omega_n, t)| \cos[\omega_n t + \tilde{\varphi}(\omega_n, t)], \quad (5)$$

where

$$\tilde{\varphi}(\omega_n, t) = \int_0^t \dot{\varphi}(\omega_n, t)dt.$$

The expectation is that loss of the additive phase constant will not be unduly deleterious.

Reconstruction of the original signal is accomplished by summing the outputs of n oscillators modulated in phase and amplitude. The oscillators are set to the nominal frequencies ω_n, and they are simultaneously phase and amplitude modulated from band-limited versions of $\dot{\varphi}(\omega_n, t)$ and $|F(\omega_n, t)|$. The synthesis operations are diagrammed in Fig. 2.

These analysis-synthesis operations may be viewed in an intuitively appealing way. The conventional channel vocoder separates vocal excitation and spectral envelope functions. The spectral envelope functions of the conventional vocoder are the same as those described here by $|F(\omega_n, t)|$. The excitation information, however, is contained in a signal which specifies voice pitch and voiced-unvoiced (buzz-hiss) excitation. In the phase vocoder when the number of channels is reasonably

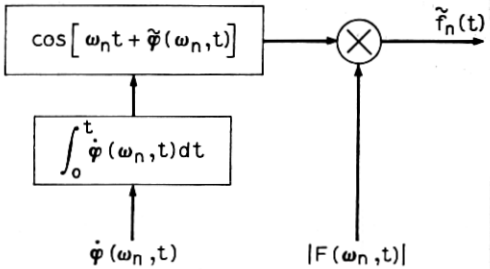

Fig. 2 — Speech synthesis based on the short-time amplitude and phase-derivative spectra.

large, the information about excitation is conveyed primarily by the $\dot{\varphi}(\omega_n, t)$ signals.* In the present technique, and if good quality and natural transmission are requisites, the indications are that the $\dot{\varphi}(\omega_n, t)$ signals may require about the same channel capacity as the spectrum-envelope information. This preliminary impression seems not unreasonable in view of our experience with voice quality in vocoders.

III. COMPUTER SIMULATION

We have simulated a complete phase vocoder analyzer and synthesizer on an IBM 7094 computer. The program, written in the BLODI-B language,[4,5] provides for the processing of any digitalized input speech signal. Flexibility built into the program permits examination of a number of design parameters such as number of channels, width of analyzing pass bands, band center frequencies, and band limitation of the phase and amplitude signals.

In the analyzer, the amplitude and phase spectra are computed by forming the real and imaginary parts of the complex spectrum

$$F(\omega_n, t) = a(\omega_n, t) - jb(\omega_n, t),$$

where

$$a(\omega_n, t) = \int_{-\infty}^{t} f(\lambda)h(t - \lambda) \cos \omega_n \lambda d\lambda$$

and

$$b(\omega_n, t) = \int_{-\infty}^{t} f(\lambda)h(t - \lambda) \sin {}_n\lambda d\lambda. \qquad (6)$$

* At the other extreme, with a small number of broad analyzing channels, the amplitude signals contain more information about the excitation, while the $\dot{\varphi}$ phase signals tend to contain more information about the spectral shape. Qualitatively, therefore, the number of channels determines the relative amounts of excitation and spectral information carried by the amplitude and phase signals.

Then,

$$| F(\omega_n , t) | = (a^2 + b^2)^{\frac{1}{2}}$$

and

$$\dot{\varphi}(\omega_n , t) = \left(\frac{\dot{a}b - \dot{b}a}{a^2 + b^2}\right). \tag{7}$$

The computer, of course, must deal with sampled-data equivalents of these quantities. Transforming the real and imaginary parts of (6) into discrete form for programming yields

$$\begin{aligned} a(\omega_n , mT) &= T \sum_{l=0}^{m} f(lT)[\cos \omega_n lT]h(mT - lT) \\ b(\omega_n , nT) &= T \sum_{l=0}^{m} f(lT)[\sin \omega_n lT]h(mT - lT), \end{aligned} \tag{8}$$

where T is the sampling interval. In the present simulation, $T = 10^{-4}$ sec. From these equations, the difference values are computed as

$$\Delta a = a[\omega_n , (m + 1)T] - a[\omega_n , mT]$$

and

$$\Delta b = b[\omega_n , (m + 1)T] - b[\omega_n , mT]. \tag{9}$$

The magnitude function and phase derivative in discrete form, are computed from (8) and (9) as,

$$\begin{aligned} | F[\omega_n , mT] | &= (a^2 + b^2)^{\frac{1}{2}} \\ \frac{\Delta\varphi}{T}[\omega_n , mT] &= \frac{1}{T}\frac{(b\Delta a - a\Delta b)}{a^2 + b^2}. \end{aligned} \tag{10}$$

Fig. 3 shows a block diagram of a single analyzer channel as realized in BLODI-B. Since this block of coding is required for each channel, it is defined as a new block type and thereafter used as though it were a single block. A parameter associated with the block determines the center frequency for each channel. The time-window analyzing filter, labeled $h(lT)$, is itself a special block and can be changed simply by the substitution of a different block of coding.[6]

In the present simulation, a sixth-order Bessel filter is used for the $h(lT)$ window. Its amplitude, phase, and delay responses are plotted in Figs. 4(a), (b), and (c), respectively. Its impulse and step responses are given in Figs. 4(d) and (e). The present simulation uses 30 channels ($N = 30$) and $\omega_n = 2\pi n(100)$ rad/sec. The equivalent pass bands of the

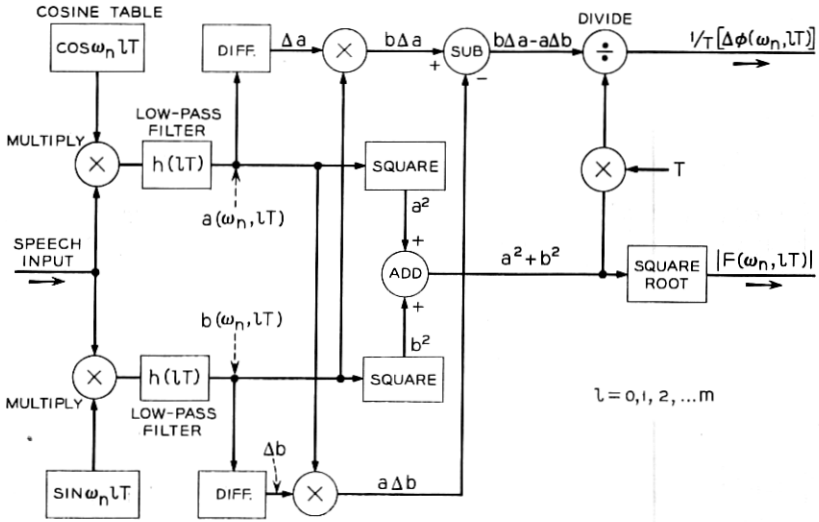

Fig. 3 — Programmed operations for extracting $| F(\omega_n , t) |$ and $\dot{\varphi}(\omega_n , t)$.

analyzing filters overlap at their 6 dB down points and a total spectrum range of 50 to 3050 Hz is analyzed.

Programmed low-pass filtering of any desired form may be applied to the amplitude and phase difference signals as defined by Fig. 3. Simulation of the whole system is completed by the synthesis operations for each channel performed according to

$$\tilde{f}_n(mT) = | F(\omega_n , mT) | \cos \left(\omega_n mT + T \sum_{l=0}^{m} \frac{\Delta\varphi(\omega_n , lT)}{T} \right). \quad (11)$$

Adding the outputs of the n individual channels, according to (1), produces the synthesized speech signal.

IV. TYPICAL RESULTS

As part of the present simulation, identical (programmed) low-pass filters were applied to the $| F(\omega_n , lT) |$ and $(1/T)\Delta\varphi(\omega_n , lT)$ signals delivered by the coding block shown in Fig. 3. These low-pass filters are similar to the $h(lT)$ filters except they are fourth-order Bessel designs. Their response characteristics are shown in Fig. 5. The cut-off frequency is 25 Hz, and the response is -7.6 dB down at this frequency. This filtering is applied to the amplitude and phase signals of all 30 channels in the present simulation. The total bandwidth occupancy of the system is therefore 1500 Hz, or a band reduction of 2:1.

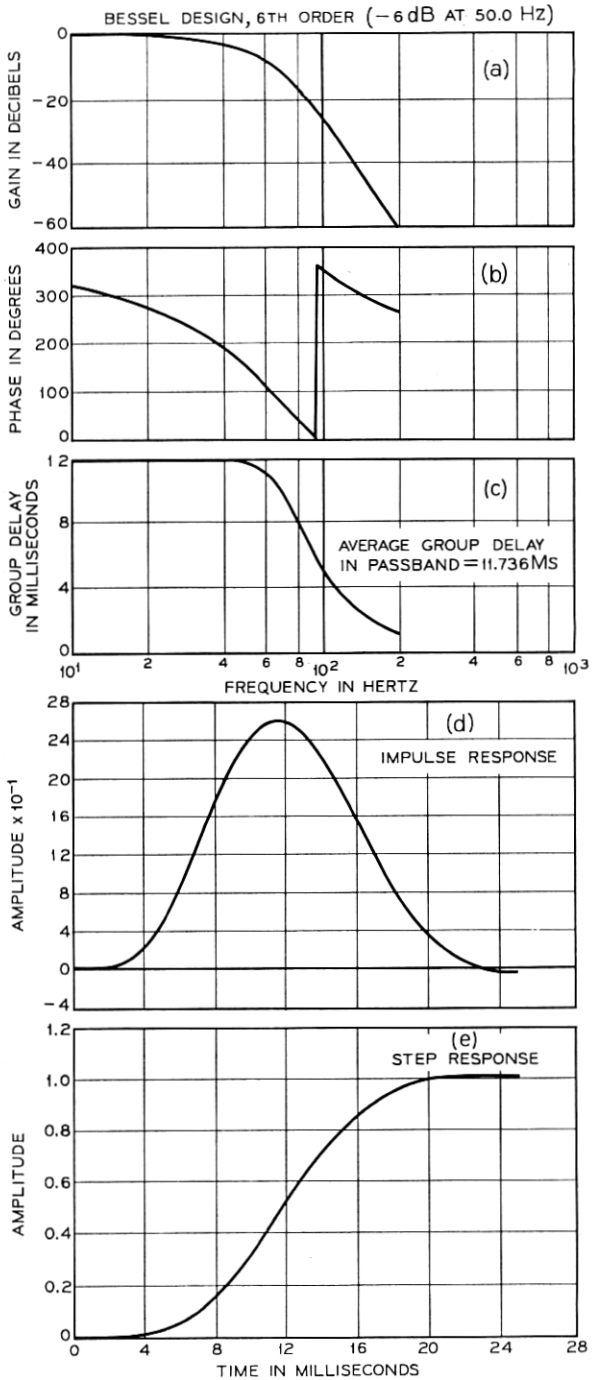

Fig. 4 — $h(t)$ analyzing function and its spectral transform used in one simulation of the phase vocoder. The function is a sixth-order Bessel filter having a −6 dB cut-off of 50 Hz.

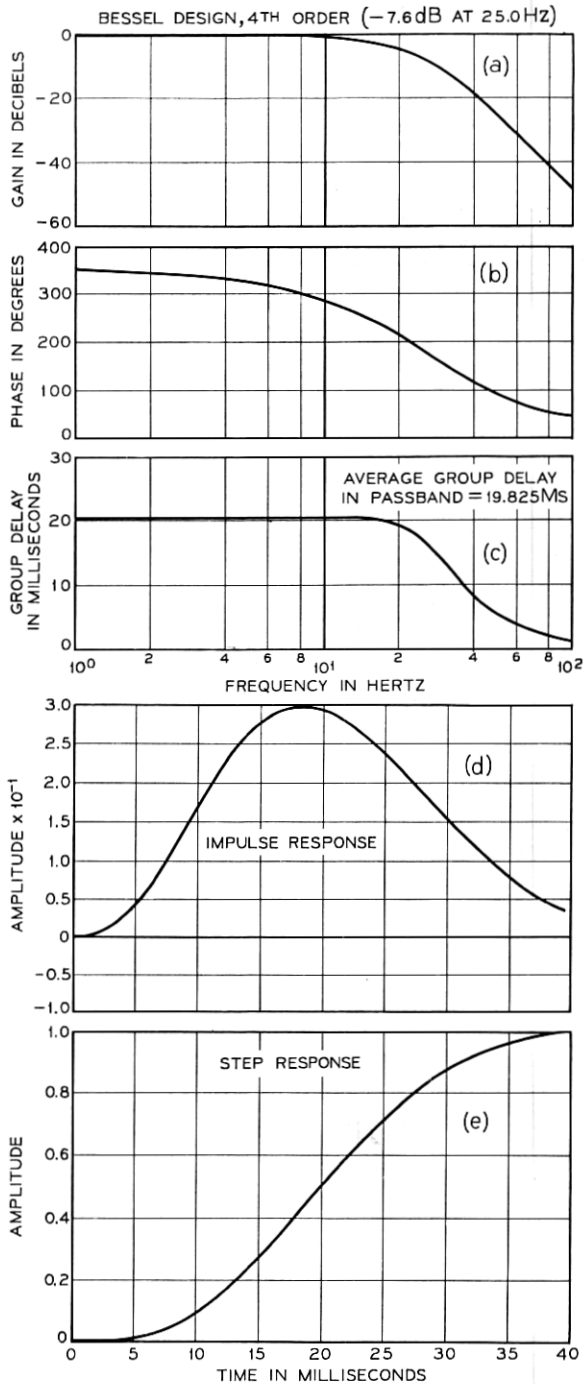

Fig. 5 — Fourth-order Bessel low-pass filter used to smooth the $|F_n|$ and $\dot{\varphi}_n$ signals.

After band-limitation, the phase and amplitude signals are used to synthesize an output according to (11). The result of processing a complete sentence through the programmed system is shown by the sound spectrograms in Fig. 6.* Since the signal band covered by the analysis and synthesis is 50 to 3050, the phase-vocoded result is seen to cut off at 3050 Hz. In this example, the system is connected in a "back-to-back" configuration, and the band-limited channel signals are not multiplexed.

Comparison of original and synthesized spectrograms reveals that formant details are well preserved and pitch and voiced-unvoiced features are retained to perceptually significant accuracy. The quality of the resulting signal considerably surpasses that usually associated with conventional channel vocoders.

V. MULTIPLEXING FOR TRANSMISSION

Besides conventional multiplexing methods for transmitting the band-limited phase and amplitude channel signals (that is, space-frequency or time-division multiplex), the coding technique suggests several other possibilities for transmission in a practicable communication system. As an example, suppose a limited-bandwidth analog channel is the available communication link. One advantageous procedure then is simply to divide (or scale down) all of the phase-derivative signals by some number, say 2 if the available channel has only one-half the conventional voice bandwidth. A synthetic signal of one-half the original bandwidth is then produced by modulating carriers of $\omega_n/2$ by the $\dot{\varphi}_n/2$ and $|F_n|$ signals. The synthetic analog signal now may be transmitted over the half-bandwidth channel.

At the receiver, restoration to the original bandwidth is accomplished by a second sequence of analysis and synthesis operations; namely, amplitude and phase analysis of the half-band signal, multiplication of the phase-derivative signals by a factor of 2, and modulation of ω_n carriers by the restored $\dot{\varphi}_n$ and reanalyzed $|F_n|$ signals. This "self-multiplexing" transmission is illustrated in Fig. 7. Spectrograms of the input signal, the half-band frequency divided signal, and the reanalyzed and resynthesized output are shown. It is clear that two trips through the process introduces measurable degradation, but the intelligibility and quality, particularly for high-pitched voices, remains reasonably good.

In effect, the greatest number q by which the ω_n and $\dot{\varphi}_n$'s may be

* The input speech signal is band limited to 4000 Hz. It is sampled at 10,000 Hz and quantized to 12 bits. It is called into the program from a digital recording prepared previously.

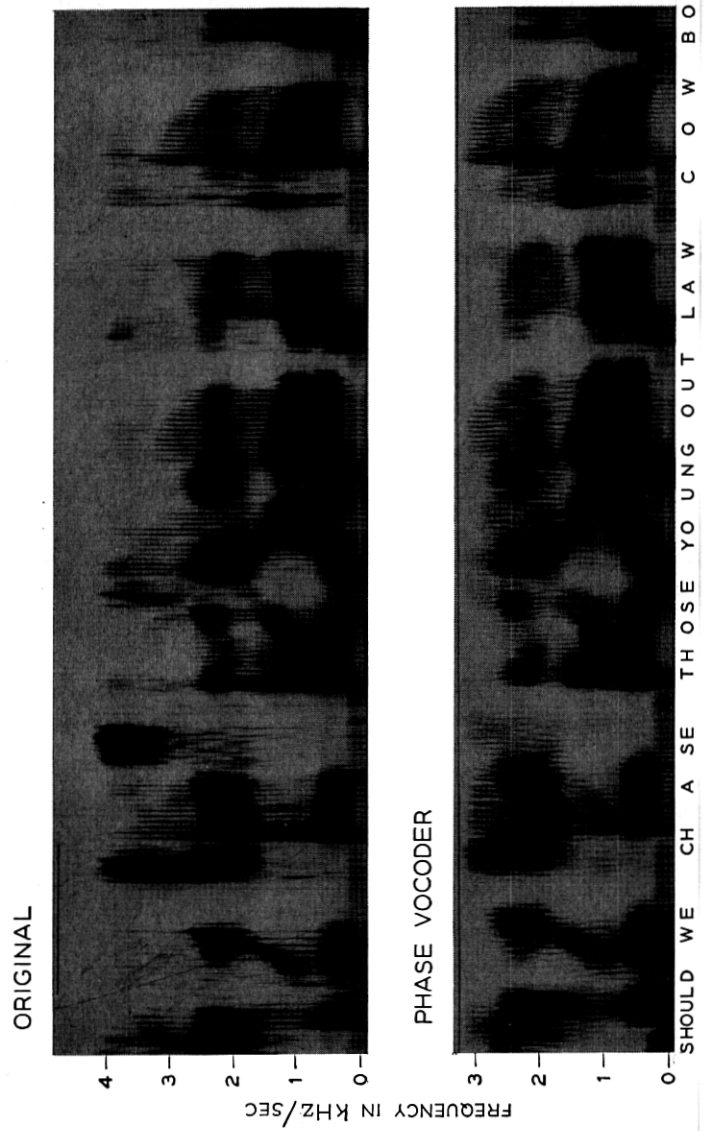

Fig. 6 — Spectrograms illustrating speech transmitted by the phase vocoder ($N = 30$). The band-pass analysis is by sixth-order Bessel filters of 100-Hz band-width. Low-pass filtering of $|F_n|$ and $\dot{\varphi}_n$ is by fourth-order Bessel filters with 25 Hz cut-off. Male speaker A. "Should we chase those young outlaw cowboys."

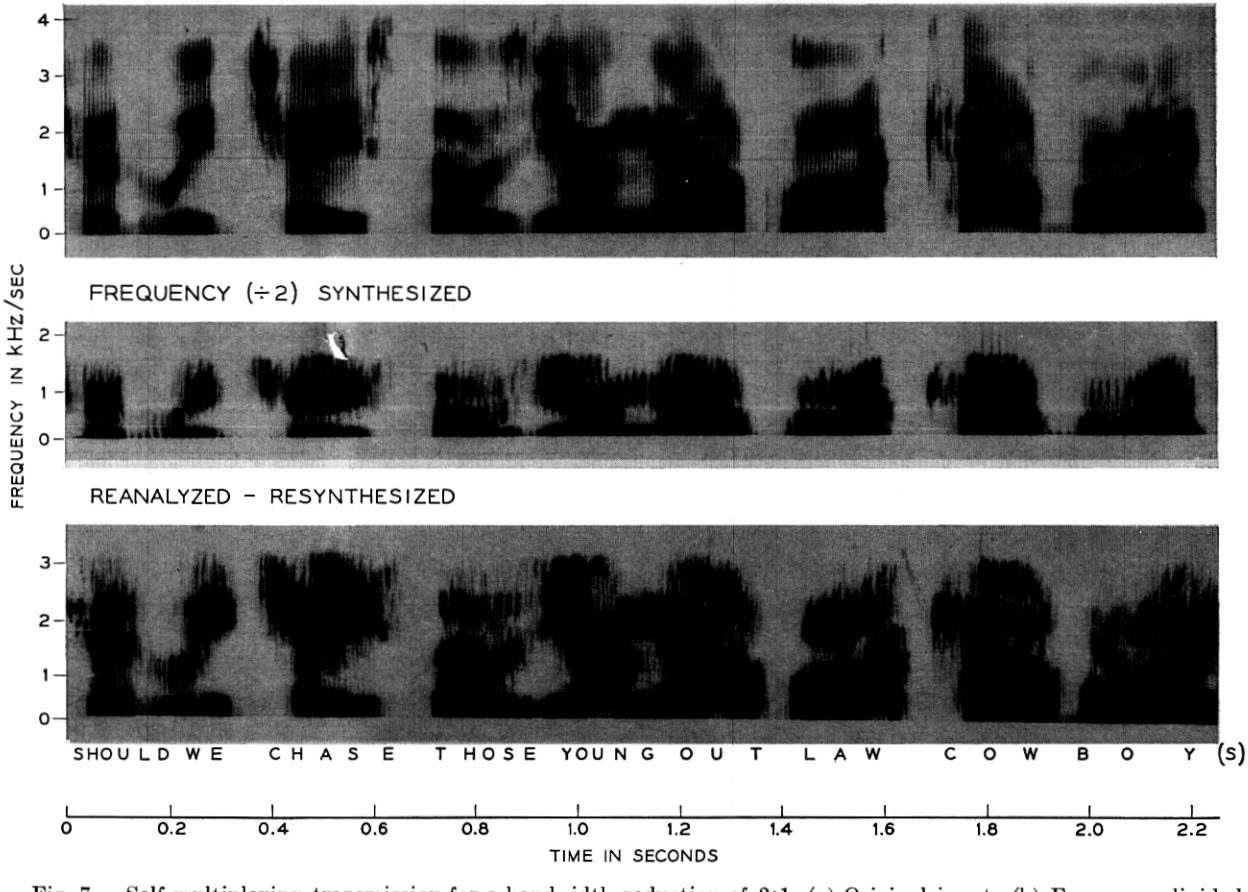

Fig. 7 — Self-multiplexing transmission for a bandwidth reduction of 2:1. (a) Original input; (b) Frequency-divided synthetic signal for analog transmission over one-half bandwidth channel; (c) Synthesized output from the reanalyzed, frequency-multiplied, half-band signal. Male speaker B.

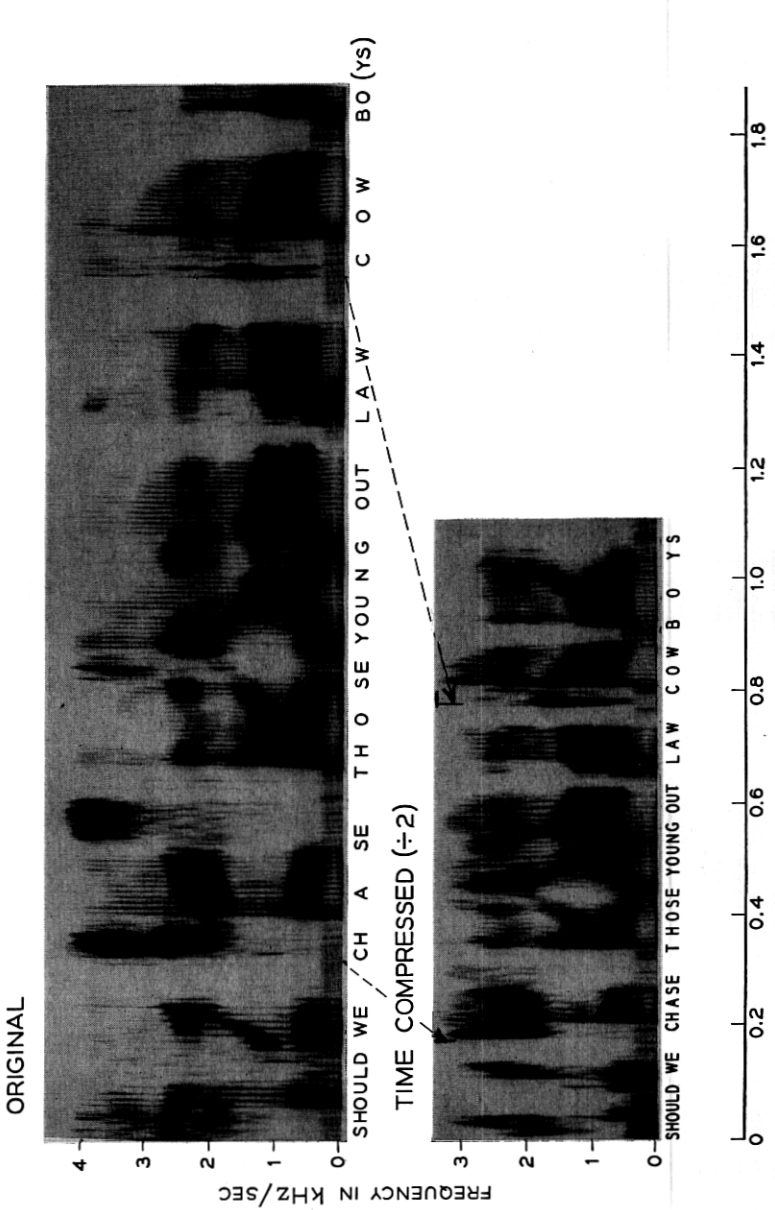

Fig. 8 — Time compression of speech by a factor of 2. Male speaker A. (a) Original input; (b) Time-compressed output.

divided is determined by how distinct the side-bands about each ω_n/q remain, and by how well each $\dot{\varphi}_n/q$ and $|F_n|$ may be retrieved from them.* Practically, the greatest number appears to be about 2 or 3 if transmission of acceptable quality is to be realized.

VI. COMPRESSION AND EXPANSION OF THE TIME SCALE

As mentioned above, a synthetic frequency-divided signal may be produced through division of $[\omega_n t + \int \dot{\varphi}_n dt]$ by some number q. This signal may be essentially restored to its original spectral position by a time speed-up of q. Such a speed-up can be accomplished by recording at one speed and replaying q-times faster. The result is that the time scale is compressed and the message, although spectrally correct, lasts $1/q$th as long as the original. An example of a 2:1 frequency division and time speed-up is shown by the sound spectrograms in Fig. 8. This feature of the phase vocoder is completely parallel to the time-compression feature of the "harmonic compressor" reported earlier.[7] However, the techniques for analysis and synthesis in the two cases are basically different, and the phase vocoder allows compression by non-integer factors.

Time-scale expansion is likewise possible by the frequency multiplication $q[\omega_n t + \int \dot{\varphi}_n dt]$; that is, by recording the frequency-multipied synthetic signal and then replaying it at a speed q-times slower. An example of time-expanded speech is shown by the spectrograms in Fig. 9. The expansion feature provides an interesting "auditory microscope" for directing attention to the spectral properties of specific elements of speech sounds — such as rapidly articulated consonants. In both compression and expansion of the time scale, a perceptual limit exists, of course, to how greatly the time scale may be altered and still have the signal sound like human speech.

An attractive feature of the phase vocoder is that the operations for expansion and compression of the time and frequency scales can be realized by simple scaling of the phase-derivative spectrum. Since the frequency division and multiplication factors can be non-integers, and can be varied with time, the phase vocoder provides an attractive tool for studying non-uniform alterations of the time scale.[8]

* More precisely, the maximum divisor is determined by how closely

$$1/q \int_0^{qt} \dot{\varphi}_n dt$$

represents

$$\int_0^t \dot{\varphi}_n dt .$$

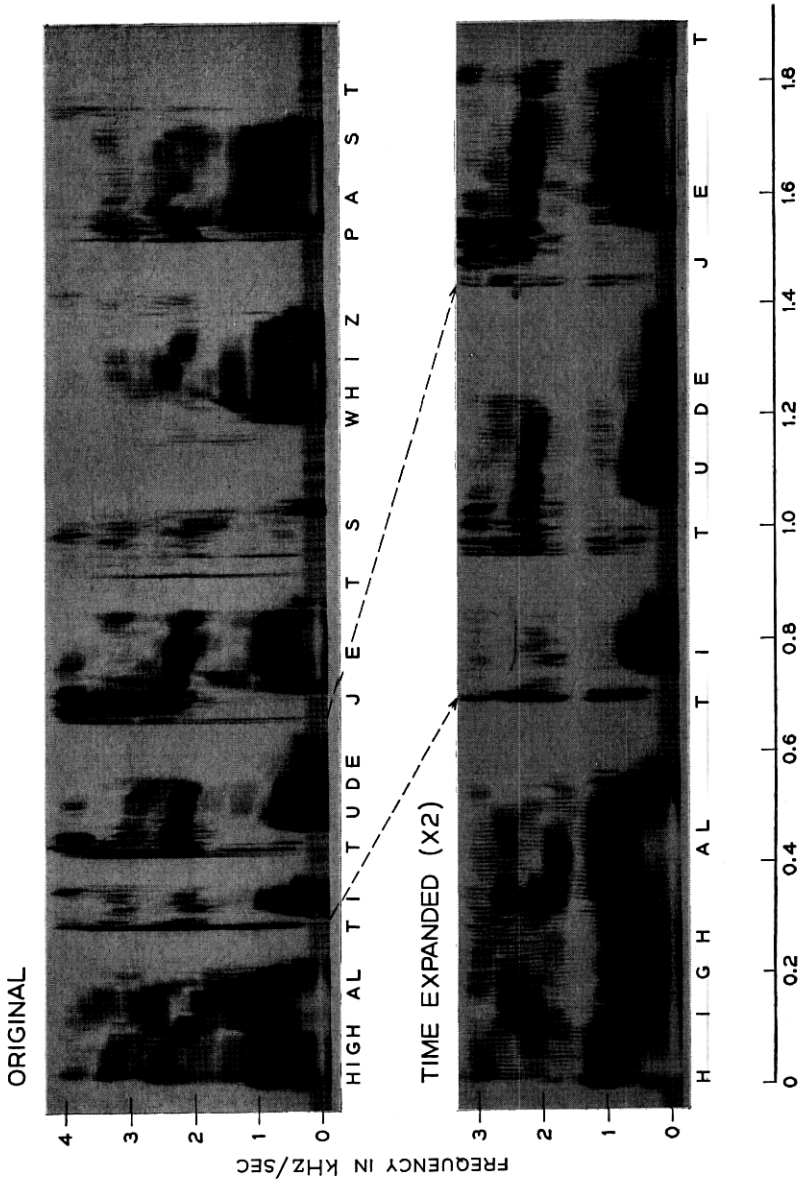

Fig. 9 — Time expansion of speech by a factor of 2. Female speaker. "High altitude jets whiz past screaming." (a) Original input; (b) Time-expanded output.

VII. FURTHER REMARKS ABOUT BAND OCCUPANCY

The possibilities of frequency division imply that the $|F_n|$ and $\dot{\varphi}_n$ signals are, in practical effect, band-limited. As described previously, modest bandwidth reduction of the order of 2:1 can be accomplished by a simple scaling of all the $\dot{\varphi}_n$ signals by $\frac{1}{2}$. (Overt low-pass filtering of the $\dot{\varphi}_n$ signals is not required.) Also, low-pass filtering the analyzed signals to a total band occupancy of one-half the original bandwidth results in relatively good speech quality upon synthesis (Fig. 6). If, however, some further trade between band saving and speech quality is desired, the control signals may be low-passed more severely, with concomitant loss in quality. The impairment resulting from low-passing the $\dot{\varphi}_n$ signals is a comb-filtering, reverberant effect in the reconstituted signals. Qualitatively, low-pass filtering of the $\dot{\varphi}_n$ signals apparently restricts the rate at which pitch changes can be duplicated, and "narrows" the sidebands produced about each ω_n-carrier at the synthesizer.

The discussion connected with (4) has pointed out that each band-pass signal in the phase vocoder may be considered as the simultaneous amplitude and phase modulation:

$$f_n(t) = |F_n| \cos(\omega_n t + \varphi_n),$$

where $|F_n|$ and φ_n are non-band limited, real-valued functions of ω_n and time. Practically, the bandwidth of $f_n(t)$ is confined to $2W$, where W is the cut-off frequency of the low-pass time aperture $h(t)$. This fact does not, however, suggest in an explicit way the band occupancy of the signals $|F_n|$ and φ_n. The experimental results of the present study indicate that each of the latter, at least for practical purposes, can be limited to around $W/2$ or less, but analytical treatment leading to explanation is difficult. Even the inverse problem, that is, calculation of the band occupancy of a simultaneously amplitude and phase modulated carrier, can only be bounded loosely.[9] To apply these bounds requires a precise description of the $|F_n|$ and $\dot{\varphi}_n$ signals. Although these parameters can be measured for a given speech signal, a general mathematical specification is not presently available. It is easy to indicate the difficulties involved. Consider the usual model of voiced speech sounds; that is, a periodic pulse source, whose frequency (pitch) may change with time, supplying excitation to a linear, passive, time-variable network. Variation of the network transmission represents the spectral changes both in the vocal sound source and the vocal tract transmission. For an analysis in terms of narrow pass-bands (large N), the $\dot{\varphi}_n$ signals depend primarily upon voice pitch. The $|F_n|$ signals, on the other hand,

depend both upon source spectrum and vocal transmission at any given instant.

VIII. CONSIDERATIONS FOR DIGITAL TRANSMISSION

Applications of the phase vocoder technique to digital transmission are of course obvious. Given an acceptable band-limitation of the $|F_n|$ and $\dot{\varphi}_n$ signals, each may be sampled at its Nyquist rate, or higher, and quantized to an accuracy that is perceptually sufficient. At this writing, optimum parameters for sampling and quantizing the control signals have not been studied in detail. Based upon past experience, however, a nonuniform distribution of the pass bandwidths of the analyzing filters would appear advantageous. For example, center frequencies and bandwidths chosen according to the Koenig scale, the mel (pitch) scale, or the auditory critical-band function should yield dividends.*

All of these bandwidth tapers are characterized by widths which monotonically increase with frequency. In such cases, the low-pass filtering applied to the amplitude signals would have cut-off frequencies also increasing monotonically with frequency. On the other hand, the low-pass filters applied to the phase signals might have cut offs which decrease with frequency. As a result, sampling rates would increase with ω_n for amplitude signals and diminish for phase signals. In addition, quantization levels for all signals might be made more coarse (less numerous) with increasing channel frequency. This is indicated because the ability of the ear to perceive frequency and amplitude changes in the higher end of a complex spectrum is, in general, less acute than for the lower part.

Although detailed study is yet to be made of optimum digital formats, experience in this area with related vocoder devices suggests that transmission at bit rates somewhat less than ten kilobits/sec should be possible without impairment due to digitalization. This rate is several times less than that normally associated with comparable quality PCM encodings of the speech waveform. Besides the questions of design optimization and data format for digital transmission, the trade which may be effected between signal quality and total bit rate is also a subject for further investigation.

IX. CONCLUDING COMMENTS

Because the phase vocoder produces phase derivative signals, it pro-

* Preliminary tests along these lines indicate that a phase vocoder with as few as eight non-uniform channels is capable of relatively good transmission (J. J. Kalsalik, unpublished work).

vides a particularly convenient means for multiplying or dividing the frequency spectrum of a broadband signal. By the same token, it is a convenient method for compressing or expanding the time scale of a signal. Frequency division of speech appears to hold potential as a communication aid for persons with hearing deficient in the high frequencies. Time compression shows promise for auditory "speed-reading" by persons with impaired sight.

Psychoacoustic and physiological studies show that the human ear makes a type of short-time spectral analysis of acoustic signals. This analysis occurs at an early level in the auditory processing; in fact, at a preneural level. It is also clear that the auditory system utilizes information corresponding to smoothed values of the short-time amplitude and phase spectra. The phase vocoder aims to turn these facts to advantage by describing speech signals in terms of band-limited values of the short-time amplitude and phase-derivative spectra. Indications are that band-limited spectral samples, occupying a bandwidth on the order of one half that of the original signal, preserve perceptually-significant features of the signal. Further band conservation can be realized, but at the expense of signal quality. As in many other transmission systems, a continuum of band conservation (or bit rate) versus signal quality exists, and one may choose the point of operation to suit requirements.

REFERENCES

1. Flanagan, J. L., *Speech Analysis, Synthesis and Perception*, Springer Verlag and Academic Press, New York, 1965.
2. Dudley, H., The Vocoder, Bell Labs. Record, *17*, 1939, pp. 122–126.
3. David, E. E., Schroeder, M. R., Logan, B. F., and Prestigiacomo, A. J., New Applications of Voice-Excitation to Vocoders, Proc. Stockholm Speech Comm. Seminar, R.I.T., Stockholm, Sweden, September, 1962.
4. Karafin, B. J., The New Block Diagram Compiler for Simulation of Sampled-Data Systems, AFIPS Conf. Proc., *27*, Pt. 1, 1965, pp. 55–61, Fall Joint Computer Conference, Spartan Books, Washington, D. C.
5. Golden, R. M., Digital Computer Simulation of Sampled Data Communication Systems Using the Block Diagram Compiler, BLODI-B, BSTJ, *45*, March, 1966, pp. 345–358.
6. Golden, R. M. and Kaiser, J. F., Design of Wideband Sampled-Data Filters, BSTJ, *43*, Pt. 2, July, 1964, pp. 1533–1546.
7. Schroeder, M. R., Logan, B. F., and Prestigiacomo, A. J., Methods for Speech Analysis-Synthesis and Bandwidth Compression, Fourth International Congress on Acoustics, Copenhagen, August 21–28, 1962.
8. Hanover, S. L. and Schroeder, M. R., Nonlinear Time Compression and Time Normalization of Speech, 72nd Meeting Acoustical Society of America, November, 1966.
9. Kahn, R. E. and Thomas, J. B., Some Bandwidth Properties of Simultaneous Amplitude and Angle Modulation, IEEE Trans. Inform. Theor., *IT-11*, October, 1965, pp. 516–520.