whose last form is essentially the invention made by Danielson and Lanczos [5]. (This is actually a particular case of an identity used earlier.) Thus, we see the transform of length T can be found from two length $T/2$ transforms on the two (real) series $X_0, X_2, \cdots, X_{T-2}$ and $X_1, X_3, \cdots, X_{T-1}$. Provided we can manage to split the series up this way, we can compute the two transforms using method 2) above and recombine them together with the complex exponential. The splitting (in place) has been programmed (by Bingham) taking advantage of the decomposition into disjoint cycles of the permutation to which it is equivalent.

References

[1] D. R. Brillinger, "An introduction to a polyspectra," *Annals of Mathematical Statistics*, vol. 36, pp. 1351–1374, 1965.
[2] M. D. Godfrey, "An exploratory study of the bispectrum of economic time series," *Applied Statistics*, Ser. C, vol. 14, no. 1, pp. 48–69.
[3] B. Bogert, M. J. Healey, and J. W. Tukey, "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Time Series Analysis*, M. Rosenblatt, Ed. New York: Wiley, 1963.
[4] J. W. Tukey, "Discussion, emphasizing the connection between analysis of variance and spectrum analysis," *Technometrics*, vol. 3, pp. 191–219, May 1961.
[5] G. C. Danielson and C. Lanczos, "Some improvements in practical Fourier analysis and their application to X-ray scattering from liquids," *J. Franklin Inst.*, vol. 233, pp. 365–380, 435–452, April–May 1942.
[6] P. Rudnick, "Note on the calculation of Fourier series," Marine Physical Lab., Scripps Inst. of Oceanography, La Jolla, Calif., Rept. MPL-U-68-65, 1965.
[7] I. J. Good, "The interaction algorithm and practical Fourier series," *J. Roy. Statist. Soc.*, ser. B, vol. 20, pp. 361–372, 1958; Addendum, vol. 22, pp. 372–375, 1960. (MR 21, 1674; MR 23, A 4231.)
[8] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of Fourier series," *Math. Comput.*, vol. 19, pp. 297–301, 1965.
[9] G. Sande, "On an alternative method of calculating autocovariance functions" (unpublished manuscript).
[10] T. G. Stockham, Jr., "High-speed convolution and correlation," *1966 Spring Joint Computer Conf., AFIPS Proc.*, vol. 28. Washington, D. C.: Spartan, 1966, pp. 299–233.
[11] H. Helms, personal correspondence.
[12] N. R. Goodman, "Measuring amplitude and phase," *J. Franklin Inst.*, vol. 270, pp. 437–450, December 1960.
[13] M. D. Godfrey, "Low-frequency variations in the German economy" (unpublished manuscript).
[14] G. Sande, an algorithm submitted to *Commun. ACM*.
[15] P. D. Welch, "A direct digital method of power spectrum estimation," *IBM J.*, pp. 141–156, April 1961.

Spectrum Analysis in Speech Coding

JAMES L. FLANAGAN, SENIOR MEMBER, IEEE

***Abstract*—In the process of hearing, the human ear develops a short-time spectrum of its acoustic input. Information-bearing features of the signal are retained in this spectral analysis. An understanding of the process by which the human auditory system preserves perceptually significant features is valuable in developing speech-transmission techniques. An example of effort in this direction is the phase vocoder.**

IN THE CONTEXT of this special issue, I would like to mention an area of research where spectrum analysis and related techniques play an important role. The area concerns the analysis and synthesis of speech signals, and particularly the transmission of speech by vocoder methods. Spectrum analysis is a cornerstone for this work, and with good reason.

At an early stage in its processing, the ear develops a short-time spectrum of its acoustic input. Features of a speech signal that are important to perception are obviously retained in such analysis. For efficient speech transmission we strive to represent these features with as little expenditure of channel capacity as is necessary. Although the full story of information processing in the auditory system remains to be discovered, we do know something of the peripheral processing. Let us consider briefly the nature of this analysis.

The diagram at the top of Fig. 1 represents the peripheral part of the ear. Shown in the diagram are the external ear or pinna, the external canal, the eardrum, the ossicular chain of the middle ear, and the inner ear or cochlea. The cochlea, which is shown here unrolled and stretched out, is a fluid-filled chamber in the temporal bone. It is divided over most of its length by a partition bounded by two membranes: Reissner's membrane and the basilar membrane. The first-order fibers of the auditory nerve terminate along the basilar membrane. Any motion of this membrane, beyond a certain threshold, causes electrical activity in the first-order fibers. This electrical activity is transmitted to the brain.

The basilar membrane is a mechanical frequency analyzer of a sort. It is light and stiff at the end toward the eardrum, the basal end, and is more massive and more compliant toward its far, apical end. If an input sound has high-frequency content, for example, a 10 000-Hz tone, the membrane vibrates maximally towards the basal end. If, on the other hand, the input

Manuscript received February 28, 1967.
The author is with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

has low-frequency content, say a 100-Hz tone, it vibrates maximally toward the apex. As already indicated, motion of this membrane is instrumental in producing the percept.

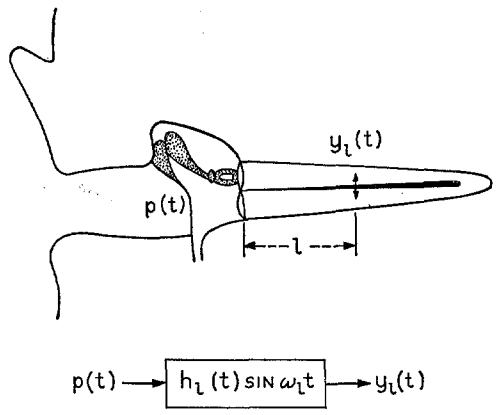
At a distance l along the membrane, we can characterize the displacement by the time function $y_l(t)$. Similarly, the sound pressure at the eardrum is $p(t)$. In a crude approximation, but one that is adequate for illustrative purposes, these quantities can be linked by the filter function shown as the impulse response in the box of Fig. 1.[1] Here ω_l is the characteristic frequency of the membrane place l-distance from the base, and the envelope function $h_l(t)$ is a non-negative, physically realizable low-pass function.

The motion at place l is the convolution of the input with this filter function, and can be represented as shown in Fig. 1. The quantity $|P(\omega_l, t)|$ is the magnitude of the short-time Fourier transform of $p(t)$, evaluated at frequency ω_l. Similarly, $\phi(\omega_l, t)$ is the short-time phase spectrum, also evaluated at ω_l. One sees that the portion of the input signal which is Fourier transformed at any given time is that part which is "viewed" through the sliding time aperture $h_l(-t)$. In the ear, at least over the apical half of the membrane, the envelope function has a form approximately proportional to $h_l(t) \sim [(\omega_l t)^2 e^{-\omega_l t/2}]$. That is, the filter function is nearly constant-percentage bandwidth in character.

A reasonable question is, "What kind of neural processing is applied to the membrane motion, and what features of the motion are important to perception?" Details of neural processing in the auditory system, and especially the conversion of mechanical response into neural activity, are very poorly understood. This area is, in fact, one frontier in auditory work. Nevertheless, there are certain facts from physiological and psychoacoustic work that provide grounds for speculation. Among the more obvious relations, for example, are indications that intensity perception is linked to temporally averaged values of the magnitude function $|P_l|$. Further, under certain conditions frequency perception appears related to temporally averaged values of the instantaneous frequency of motion $(\omega_l + \dot{\phi}_l)$.

Recognizing these gross relations, and other more specific ones which give additional insight into the processing, we hope to turn them to advantage for speech transmission. The aim is to duplicate, in some sense, the processing which the auditory system applies to acoustic signals, and to preserve those features which are perceptually significant and eliminate those which are not. By so doing, one hopes to conserve transmission bandwidth, while at the same time providing perceptually acceptable signals.

An example of effort in this direction is a device we call a phase vocoder. It is so named because it encodes speech information in terms of smoothed values of the



$$y_l(t) = p(t) * h_l(t) \sin \omega_l t$$

$$= I_m e^{j\omega_l t} \int_{-\infty}^{t} p(\lambda) h_l(t-\lambda) e^{-j\omega_l \lambda} d\lambda$$

$$= |P(\omega_l, t)| \sin [\omega_l t + \varphi(\omega_l, t)]$$

Fig. 1. Schematic diagram of peripheral ear. Analytic relations are shown for basilar membrane displacement at a place l distance along the membrane. The input sound pressure is $p(t)$ and the membrane displacement is $y_l(t)$.

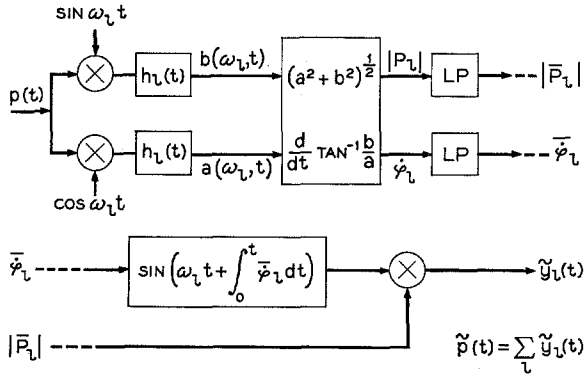


Fig. 2. Method for calculating values of the short-time amplitude spectrum and the phase-derivative spectrum, and for synthesizing a signal corresponding to low-passed values of these quantities.

phase-derivative spectrum and short-time amplitude spectrum, much along the lines of the preceding discussion. The device contains l channels, crudely analogous to l different places along the basilar membrane. Each channel accomplishes the processing shown in Fig. 2. The real and imaginary parts of the short-time spectrum, $a(\omega_l, t)$ and $b(\omega_l, t)$, respectively, are first formed, and $|P_l|$ and $\dot{\phi}_l$ are produced from them.[2] These quantities are averaged by low-pass filters, and are transmitted over a restricted bandwidth channel to the receiving terminal. At the receiver, $|\bar{P}_l|$ and $\bar{\dot{\phi}}_l$ are used to simultaneously modulate in amplitude and phase an oscillator of nominal frequency ω_l. An approximation $\tilde{y}_l(t)$ to the original bandpass signal $y_l(t)$ is thereby recovered. Summing the outputs of l such channels

[1] For simplicity, various multiplicative constants and ω_l-dependent delay are not shown.

[2] The phase derivative is calculated from $(a\dot{b} + b\dot{a})/(a^2+b^2)$. Note that, ideally, ϕ_l can be recovered to within an additive constant from $\int_0^t \dot{\phi}_l dt$.

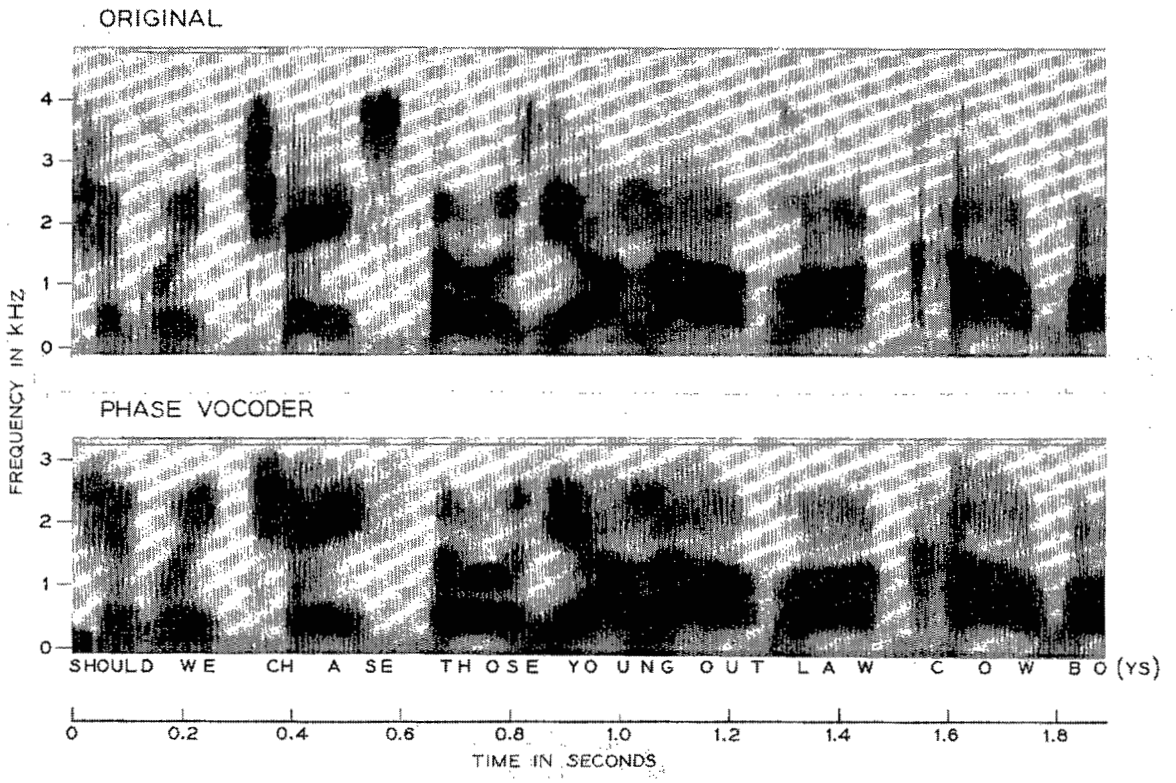


Fig. 3. Spectrograms illustrating speech transmitted by a 30-channel phase vocoder. The bandpass analysis is accomplished by sixth-order Bessel filters of 100-Hz bandwidth. Low-pass filtering of $|P_l|$ and $\dot{\phi}_l$ is by fourth-order Bessel filters of 25-Hz cutoff. Male speaker. The vocoder analysis and synthesis covers the frequency range 50 to 3050 Hz.

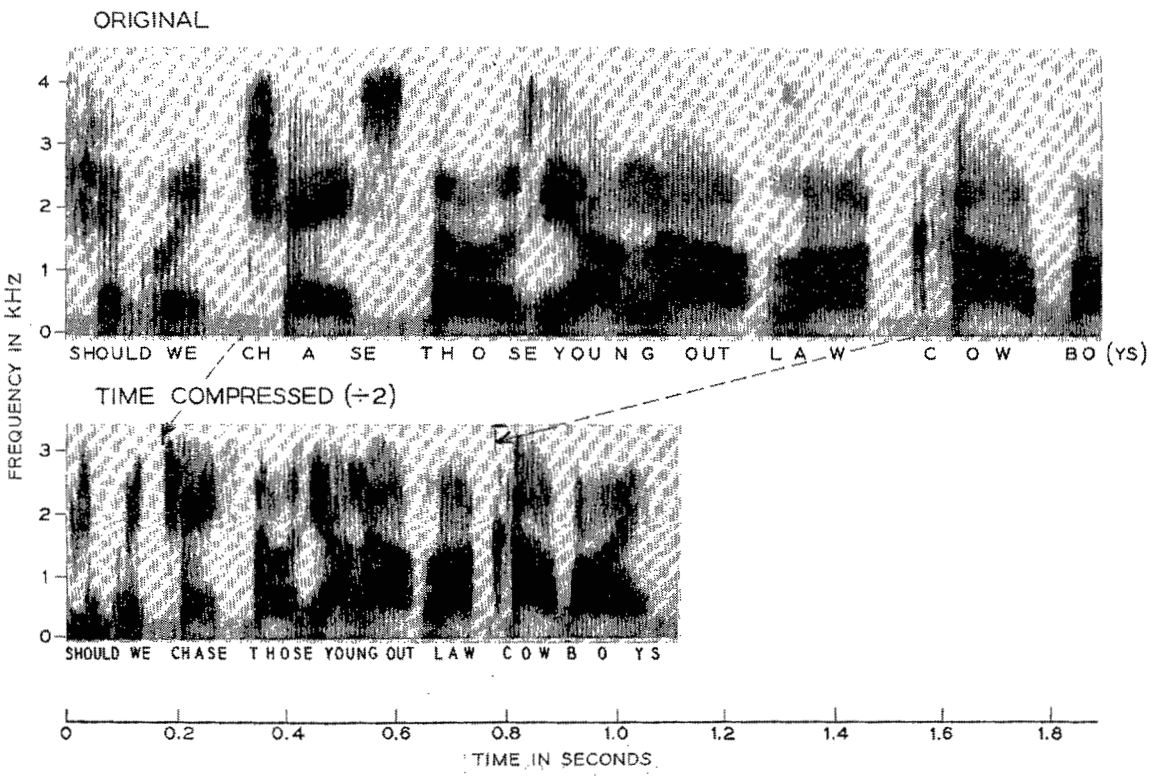


Fig. 4. Time compression of speech by a factor of 2. Male speaker. The top spectrogram is the original input speech. The lower spectrogram is the time-compressed output.

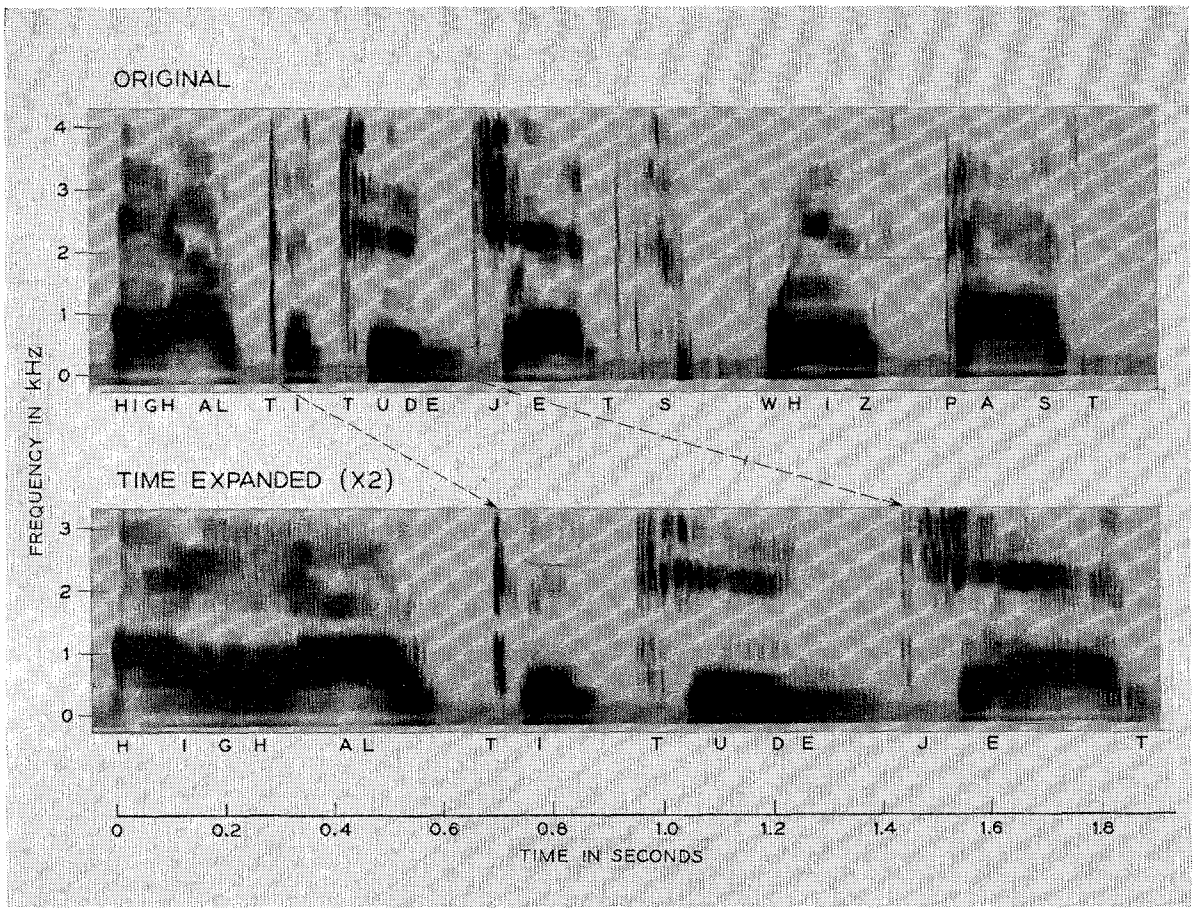


Fig. 5. Time expansion of speech by a factor of 2. Female speaker. The top spectrogram is the original input speech. The lower spectrogram is the time-expanded output.

produces an approximation to the input signal in which, ideally, perceptually significant information has been retained and certain irrelevant features have been discarded.

Using computer simulation techniques, we have investigated several configurations of this transmission scheme. One example is for $l=30$ and analysis passbands 100 Hz in width. Bessel low-pass filters of 25-Hz cutoff were used to limit the $|P_l|$ and $\dot{\phi}_l$ signals. A band of speech covering the frequency range 50 to 3050 Hz and transmitted over this system is illustrated in Fig. 3. For these conditions, the total transmission bandwidth is one half that of the input signal.

Because phase derivative signals or, in effect, instantaneous frequency measures, are produced in the analysis, the system provides a convenient means for frequency and time scaling. For example, suppose one divides all the $(\omega_l+\dot{\phi}_l)$ quantities by some factor k, and synthesizes a signal occupying $1/k$th the original band. Recording this result and then replaying it at a speed k times faster restores the spectrum to the original frequency range and compresses the time scale by the factor k. An example of a 2-to-1 time compression is shown by the spectrograms in Fig. 4. By the same token, multiplication of $(\omega_l+\dot{\phi}_l)$ by k leads to an expansion of the time scale. A 1-to-2 expansion is illustrated by the spectrograms of Fig. 5. Although integer values have been selected for examples, it is clear that the scale factor k can be a non-integer.

In summary, we can restate three points. The human ear makes a type of spectrum analysis in processing acoustic signals. Guided by physiological and psychoacoustic findings, we can gain insight into the perceptual importance of the parameters of this analysis. By trying to duplicate the elements of the analysis, we can be led to useful band conservation in speech transmission.