Speech spectrograms using the fast Fourier transform

Increased flexibility and the capability for on-line analysis are the two primary reasons for utilizing a digital computer for the generation and display of speech spectrograms

Alan V. Oppenheim *Massachusetts Institute of Technology*

An important aid in the analysis and display of speech is the sound spectrogram, which represents a time–frequency–intensity display of the short-time spectrum.[1–3] With many modern speech facilities centering around small or medium-size computers, it is often useful to generate spectrograms digitally, on-line. The fast Fourier transform algorithm provides a mechanism for implementing this efficiently.

A principal reason for the importance of the sound spectrogram in speech analysis is that many speech sounds can be considered to be produced by exciting a resonant cavity, the vocal tract, with either a quasi-periodic or a noiselike excitation. For many applications, then, the speech waveform is characterized by the frequencies of the vocal tract resonances and, for the quasi-periodic excitation, the fundamental frequency of the excitation, both of which are readily apparent on a spectrogram.

Many modern speech research facilities center around small or medium-size computers, which provide a mechanism for carrying out sophisticated studies in speech analysis and synthesis.[4,5] With such a facility, it is sometimes useful to generate speech spectrograms on-line rather than by making an analog recording, which is then analyzed off-line by a spectrograph machine. Furthermore, it is often advantageous to closely relate time displays with spectral displays and to be able to choose bandwidths flexibly—perhaps even time-dependently—during the analysis. Such flexibility is ideally suited to computer implementation. With the use of the fast Fourier transform algorithm for computing spectrums digitally, speech spectrograms can be implemented on small computers with analysis times approximately the same as required with modern analog spectrographic equipment.

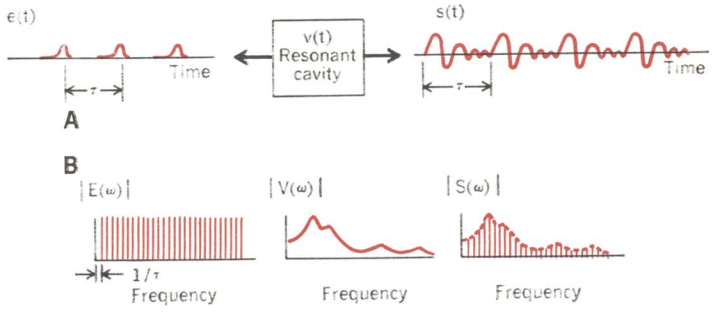
Spectral analysis of speech

A simplified but often useful model for speech production consists of a linear system with a quasi-periodic excitation function for voiced sounds (such as vowels) and a noiselike excitation during unvoiced (fricative) sounds.[6] The linear system represents the vocal cavity. The quasi-periodic excitation function during voiced sounds corresponds to the air flow through the vocal cords. During the production of a sustained vowel, the vocal tract configuration can be assumed to be fixed, and the sound produced corresponds to the response of a resonant cavity. Figure 1(A) represents the production of a steady-state vowel, where the excitation function $e(t)$ corresponds to air flow through the vocal cords, and the output $s(t)$ corresponds to air flow at the lips. If, for example, the vowel produced was the vowel /ah/ as in "father," a typical set of numbers for the first three resonances of the resonant cavity are 750, 1150, and 2500 Hz. Since the resulting vowel is periodic, its spectrum is a line spectrum with harmonics spaced in frequency by the reciprocal of the pitch period (typically about 125 Hz for a man). The envelope of the line spectrum will contain peaks corresponding to the resonant frequencies of the vocal cavity. Figure 1(B) illustrates pictorially the line spectrum corresponding to the excitation,* the spectral envelope corresponding to the frequency response of the vocal cavity, and the composite line spectrum corresponding to the spectrum of the steady-state vowel. During the production of fricative sounds such as /sh/, the excitation is a noiselike waveform produced by turbulence at the lips and teeth. Thus, for fricative sounds, the output is noiselike and has no line spectrum.

During the production of continuous speech, of course, the shape of the vocal cavity is not fixed, and the resonances vary to produce different sounds. Consequently, the linear system corresponding to the resonant cavity in Fig. 1 is time-varying, as is the excitation period. If the variation is not too rapid, it is reasonable to view the system as stationary on a short-time basis, so that a

* The line spectrum corresponding to the excitation will have an envelope representing the spectrum of an individual glottal pulse. For purposes of illustration, it was assumed that the spectrum of the glottal pulses was approximately constant.

FIGURE 1. Simplified picture of speech production.



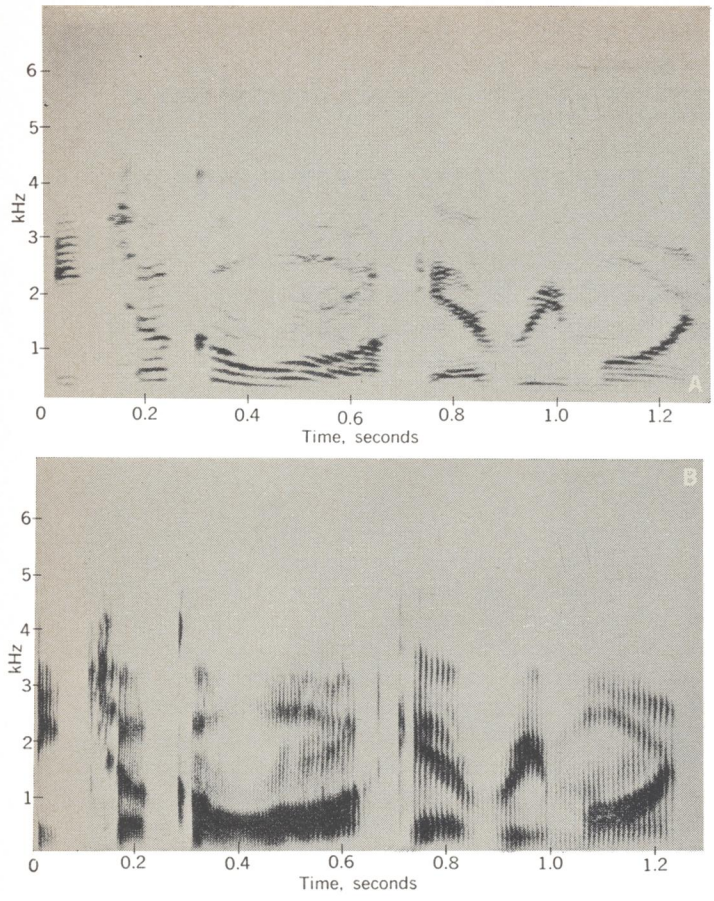


FIGURE 2. Spectrograms of the sentence, "He took a walk every morning," spoken by a male. A—Narrow-band spectrogram. B—Wide-band spectrogram.

FIGURE 3. Equivalent filter characteristic for rectangular time window.

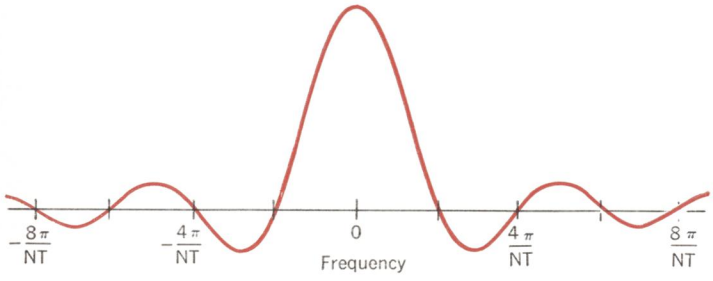
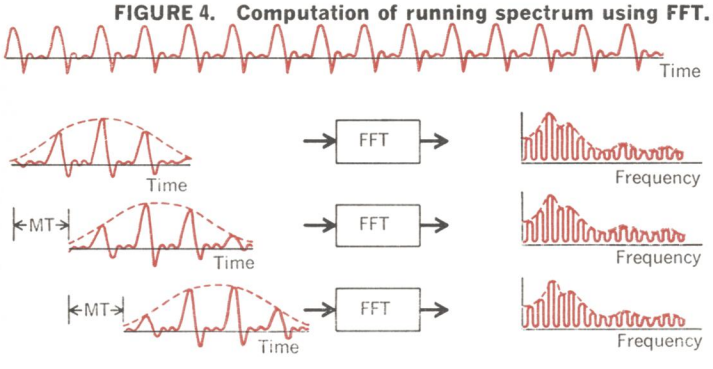


FIGURE 4. Computation of running spectrum using FFT.



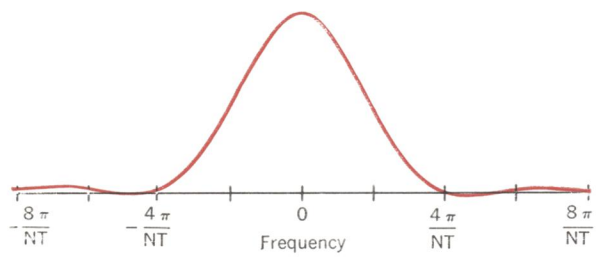
short-time spectral analysis of the speech waveform would exhibit peaks in the envelope corresponding to the resonances of the vocal tract as well as a harmonic structure corresponding to the excitation. As the time window is increased in length, frequency resolution improves and the harmonic structure becomes more evident, but the spectral analysis loses its ability to follow rapid changes.

A spectral analysis with a shorter time window, and consequently a wider frequency window, would provide better time resolution at the expense of spectral resolution; that is, it would tend not to resolve individual pitch harmonics in the spectrum but would be better able to track rapid changes. Because of this tradeoff between time and frequency resolution, it is common in spectral analysis of speech to utilize both narrow-band analysis, corresponding to good frequency resolution and poor time resolution, and wide-band spectral analysis, corresponding to good time resolution and poor frequency resolution. A typical means for obtaining and displaying speech spectrums is the spectrograph machine, for which the analysis corresponds to playing the speech through a bank of equal-bandwidth filters (usually implemented by heterodyning the signal past a single fixed filter). In a narrow-band analysis, the filter bandwidths are typically 45 Hz; for a wide-band analysis, they are 300 Hz. The recording is made on Teledeltos paper. Figure 2 shows narrow-band and wide-band spectrograms for the sentence, "He took a walk every morning," as spoken by a male. In the former, it is clear that the individual pitch harmonics have been resolved in frequency, whereas in the latter they are no longer evident. However, in the wide-band spectrogram, vertical striations can be seen that correspond to individual pitch periods. This is a consequence of the good time resolution of the wide-band spectrogram; they are not evident in the narrow-band case. Also evident on both wide- and narrow-band spectrograms are the frequency regions corresponding to high spectral amplitude designating the vocal tract resonances, referred to as formants.

Computation of spectrograms using the FFT

The fast Fourier transform (FFT) algorithm has assumed tremendous importance as a means for computing spectrums and implementing spectral displays on a digital computer.[7-10] In particular, for implementing on a digital computer spectral analysis of speech similar to that implemented by a spectrograph machine, it is considerably more efficient to carry out the analysis using the FFT algorithm than to implement a filter bank. The FFT is an algorithm for computing the discrete Fourier transform (DFT), defined as

FIGURE 5. Equivalent filter characteristic for Hanning window.



$$F(k) = \sum_{n=0}^{N-1} f(nT)e^{-j\frac{2\pi}{N}nk} \tag{1}$$

where $f(nT)$ corresponds to equally spaced samples of an analog time function $f(t)$. Assuming that the sampling has been done at a rate equal to or higher than the Nyquist rate ($2f_m$, where f_m is the highest frequency in the analog time function), it can be shown that the magnitude of the kth spectral point $|F(k)|$ in Eq. (1) corresponds to the magnitude that would be obtained at a time $t = (N - 1)T$ when *samples* of the analog function $f(t)$ are played through an analog filter with a frequency response $H(\omega)$ given by

$$H(\omega) = \frac{\sin\frac{NT}{2}\left(\omega - \frac{2\pi k}{NT}\right)}{\left(\omega - \frac{2k}{NT}\right)} \tag{2}$$

This filter characteristic is sketched in Fig. 3 for $k = 0$. The set of numbers $F(k)$, for k equal to 0 through $N - 1$, then corresponds to the set of outputs from a filter bank, each filter of which has a spectral shape similar to Fig. 3, with a center frequency at $\omega = 2\pi k/NT$.

The computation in Eq. (1) provides only one spectral section, that is, the output of the filter bank at a time $t = (N - 1)T$. To obtain a short-time spectral analysis, we would like to perform this computation at successive instants of time, and, in addition, to be able to modify the filter shape. For example, we may wish to reduce the sidelobes in the filter characteristic in Fig. 3. Furthermore, as we change from a wide-band to a narrow-band analysis, we would be required to change the width of the central lobe in the filter. To determine a running spectrum and provide flexibility in terms of the filter characteristic, the expression in (1) can be modified as

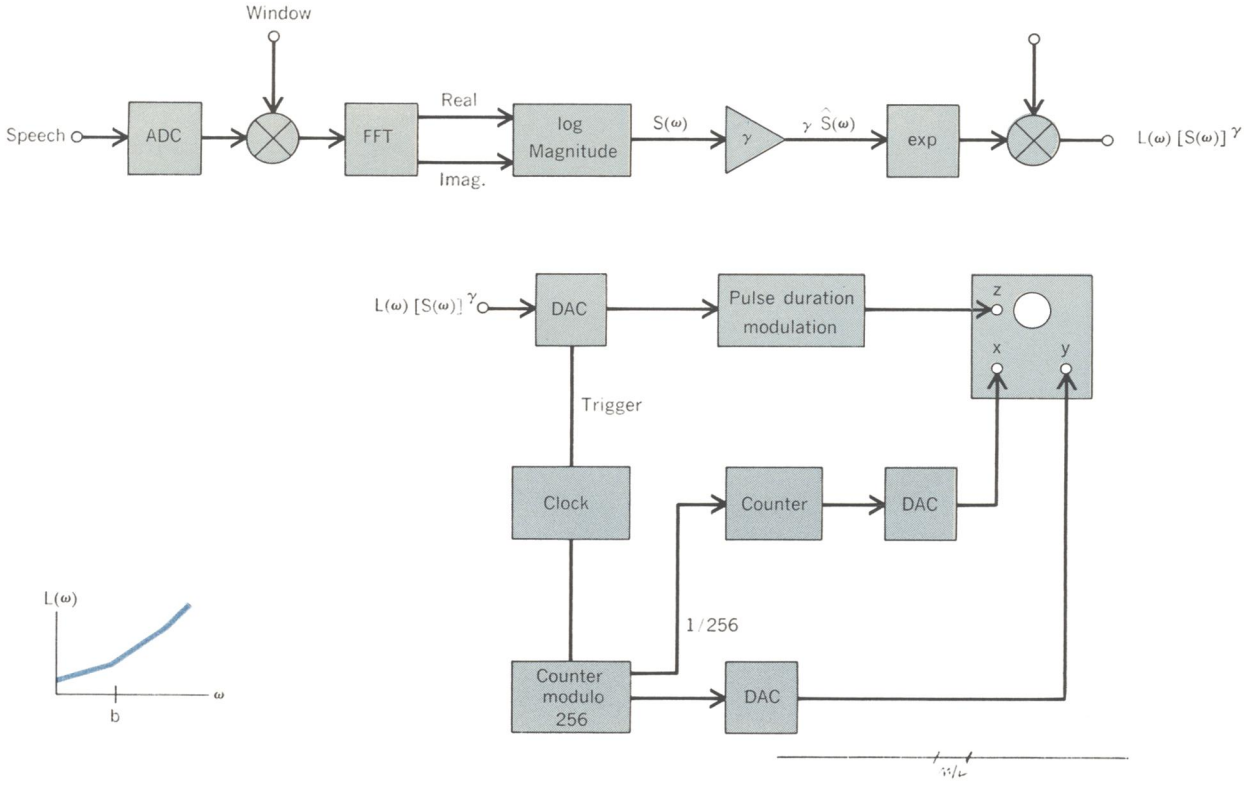
$$F_r(k) = \sum_{n=0}^{N-1} w(nT)\, f(nT + rMT)e^{-j\frac{2\pi}{N}nk} \tag{3}$$

Equation (3) introduces two changes. The first is to include a window $w(nT)$ to provide a better spectral characteristic. This is motivated by the fact that since a computation of the discrete Fourier transform as given by Eq. (1) is necessarily restricted to a computation on a finite length of data, there is implicit in (1) a time window imposed on $f(t)$, that is, $f(t)$ is multiplied by a rectangular window with a width equal to NT. It is that rectangular time window that leads to the spectral window shown in Fig. 3. By modifying the rectangular window with some new time window $w(nT)$, it is possible to modify the spectral shape shown in Fig. 3. The second modification incorporated in Eq. (3) corresponds to implementing a spectral analysis of successive sections of the waveform. In other words, the set of numbers $F_r(k)$ represents a computation of the discrete Fourier transform of a section of the analog time function starting at $t = rMT$ and ending at $t = rMT + (N - 1)T$. This corresponds to a filter bank output at time $t = rMT + (N - 1)T$. Successive sections (Fig. 4) are spaced in time by MT.

In a filter-bank implementation of the spectral analysis, the time window $w(nT)$ corresponds to the low-pass prototype of the impulse response of each of the filters. One observation from this is that the spectral analysis described by (3) corresponds to a filter-bank analysis for which the spectral shape of each of the filters in the filter bank is approximately the same. For example, Eq. (3) could not represent a filter bank having constant-Q filters, for which the bandwidth is proportional to the frequency. If a constant-Q analysis were desired, a direct implementation of the filters would be used.[11]

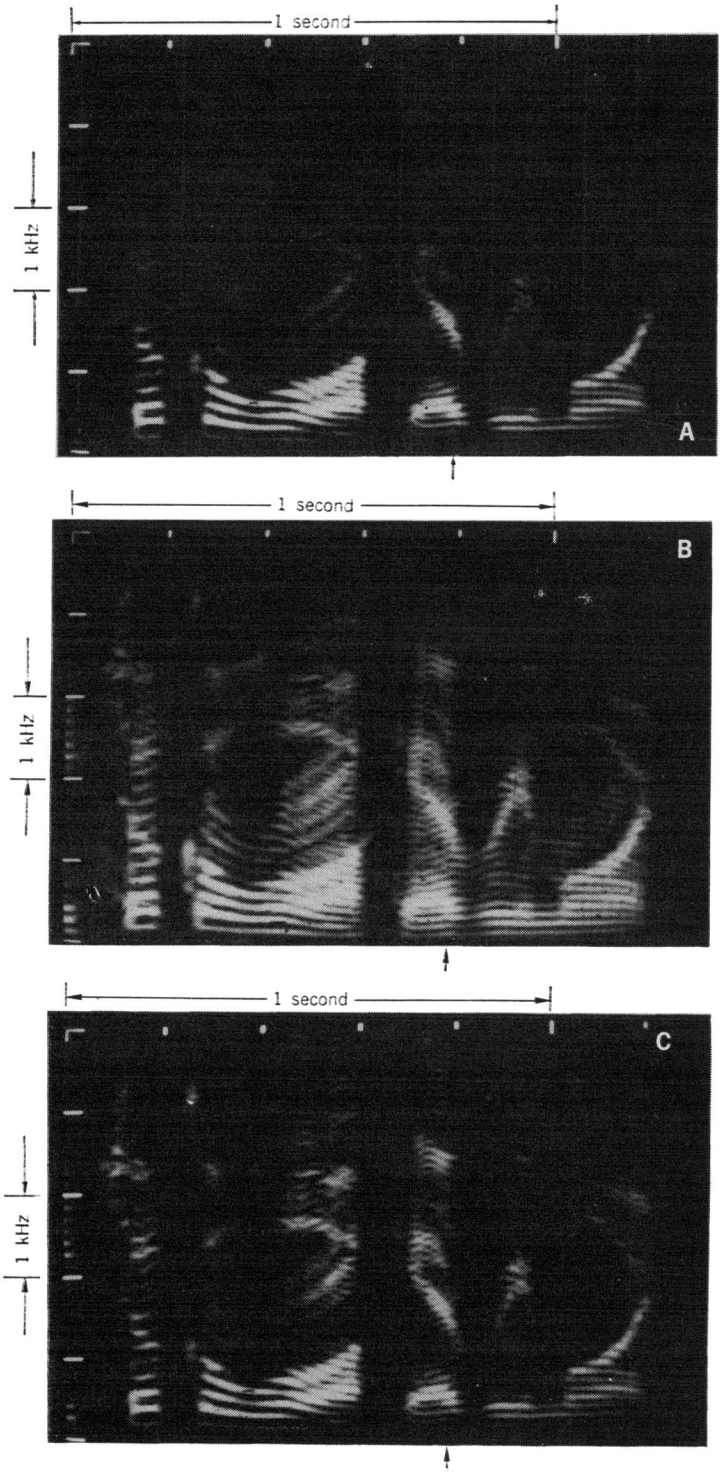
As an example of speech spectrograms obtained by

FIGURE 6. Block diagram for computation and display of spectrograms.



using the FFT, the procedure was implemented on the Univac 1219 computer facility at the M.I.T. Lincoln Laboratory. This computer, which is similar in size and speed to those in many speech laboratories, has a memory cycle time of 2 μs, an 18-bit register length, and generally utilizes fixed-point arithmetic. The implementation to be illustrated was programmed in assembly language. One of the objectives was to have an analysis time comparable to that required with modern analog spectrographic equipment. For a narrow-band spectral analysis, the number of spectral points computed, corresponding to the parameter N in Eq. (3), is larger than that required for a wide-band analysis. On the other hand, since the time resolution is worse in the narrow-band than in the wide-band case, spectral sections need be computed less frequently for narrow-band spectrograms, that is, the value of the parameter M can be larger. In the system implemented, M was chosen as a fixed percentage of N. Since the computation of the fast Fourier transform algorithm requires a time proportional to $N \log_2 N$ (assuming N is a power of 2), the analysis time for an utterance is essentially the same for either a narrow-band or a wide-band analysis. On this computer the analysis time was approximately three minutes for a two-second utterance. (This will vary somewhat, depending on the speed of the computer.) With the values used for M and N, a sufficient number of points were obtained to define the short-time spectrum with the appropriate time and frequency resolution. Linear interpolation in time and frequency between these samples was utilized to provide a smooth display. Hard copy is obtained photographically, with a time exposure.

FIGURE 7. Narrow-band spectrograms; the sentence is the same as in Fig. 2.



The time window $w(nT)$ in Eq. (3) was chosen to be a Hanning window, defined as

$$w(nT) = \frac{1}{2}\left[1 - \cos\frac{2\pi}{NT}nT\right] \qquad 0 \le n \le N \qquad (4)$$

The corresponding spectral window is shown in Fig. 5. For both the wide-band and narrow-band analysis, the input speech was preemphasized at 6 dB per octave starting at 350 Hz, low-pass-filtered to 5 kHz, and sampled at 10 kHz. For a wide-band analysis, the parameter N was chosen as 128. This means that the half-power filter bandwidths were approximately 112 Hz and the separation between successive spectral samples was 78 Hz. For the narrow-band analysis, N was chosen as 512, corresponding to half-power filter bandwidths of 28 Hz and a difference between the center frequencies of successive filters of 20 Hz.* The parameter M was chosen as 24 for the wide-band analysis and 96 for the narrow-band analysis. This corresponds to obtaining spectral sections every 2.4 ms in the wide-band case and every 9.6 ms in the narrow-band case. The assumption is that values in between can be obtained by interpolation.

In this system the spectrograms are displayed on a conventional oscilloscope with a z-axis input (Hewlett-Packard 1200AR). Each amplitude value to be displayed is used to modulate the duration of an unblanking pulse applied to the z-axis of the oscilloscope. In this way a point is displayed for a time proportional to the intensity of the spectral point. The spectrogram is reproduced, using a time exposure, on Polaroid type 52 film. Thus, intensities are recorded by utilizing the integrating time of the film, that is, the brightness of a point as it appears on the film is (approximately) proportional to the duration for which it is displayed on the oscilloscope, which in turn is proportional to the spectral amplitude to be recorded. The spectrograms are displayed on a grid with 256 points on the vertical (frequency) axis and 512

* The value of N was chosen to result in what was considered to be the best spectral display. In general, there is no constraint on the value of N; if it is not a power of 2, then the sequence to be transformed is augmented with zeros and thus an efficient computation of the transform will result.

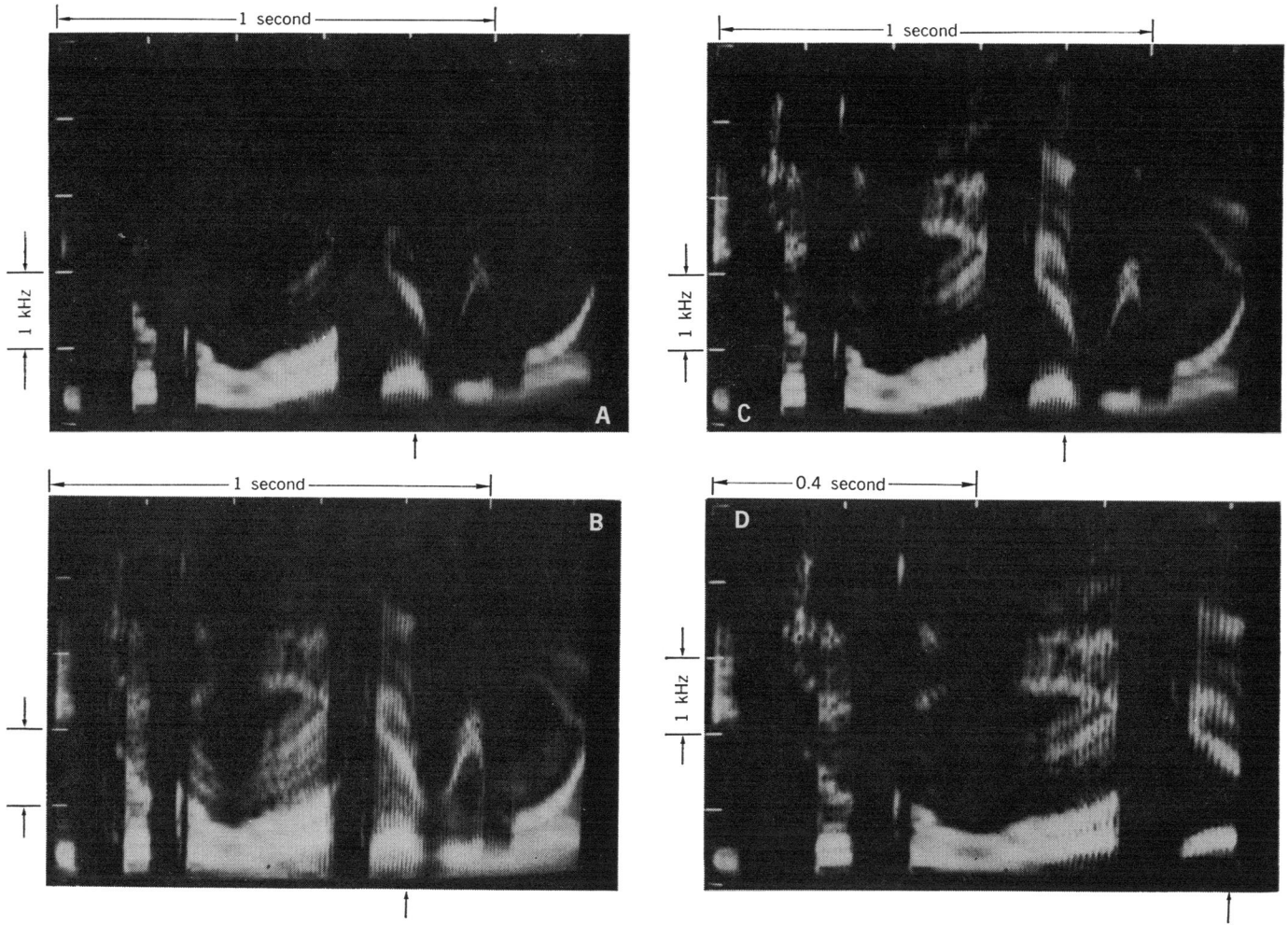
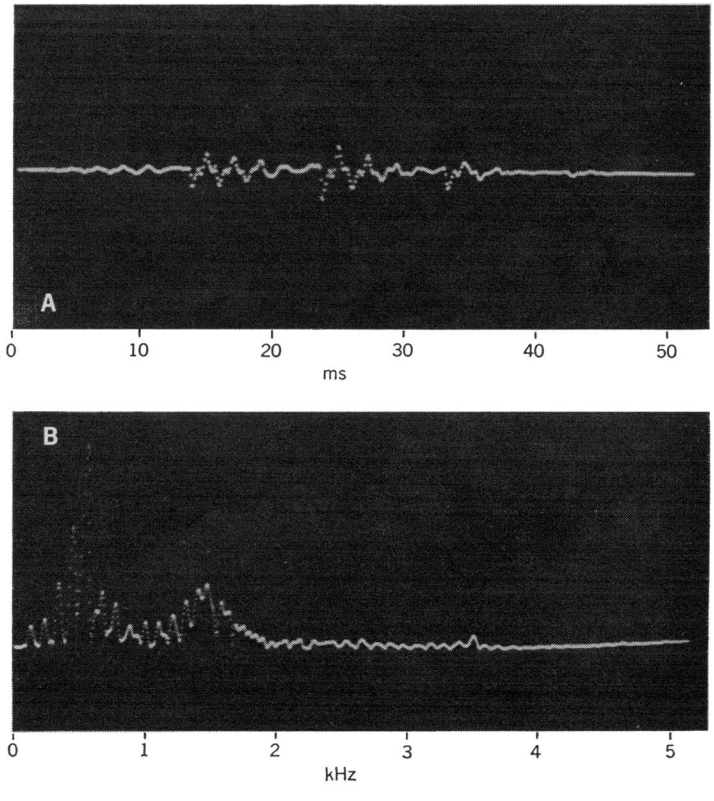


FIGURE 8. Wide-band spectrograms; the sentence is the same as in Fig. 2.

FIGURE 9. A—Speech segment with window applied for narrow-band analysis. B—Spectral section obtained from the speech segment (A).



points on the horizontal (time) axis. Values on this grid are obtained by linear interpolation from the computed values, which lie on a less dense grid.

Although the foregoing display procedure involving photographic recording results in a more dynamic range than can be obtained with the paper used on an analog spectrograph machine, the dynamic range is generally not sufficient to provide an adequate display of the higher frequencies without some additional preemphasis of the speech or the application of frequency shaping to the spectrum. Moreover, it is sometimes desirable to increase or decrease the contrast and so the facility for accomplishing both linear frequency shaping and modified contrast was included. If $S(\omega)$ represents the spectral section to be displayed, the high-frequency emphasis is accomplished by multiplying $S(\omega)$ by a function $L(\omega)$, which is constant to some frequency and thereafter increases linearly with a specified slope. Contrast is modified by raising $S(\omega)$ to the γ power. With γ greater than unity, contrast is increased; with γ less than unity, contrast is decreased. The system is summarized in Fig. 6.

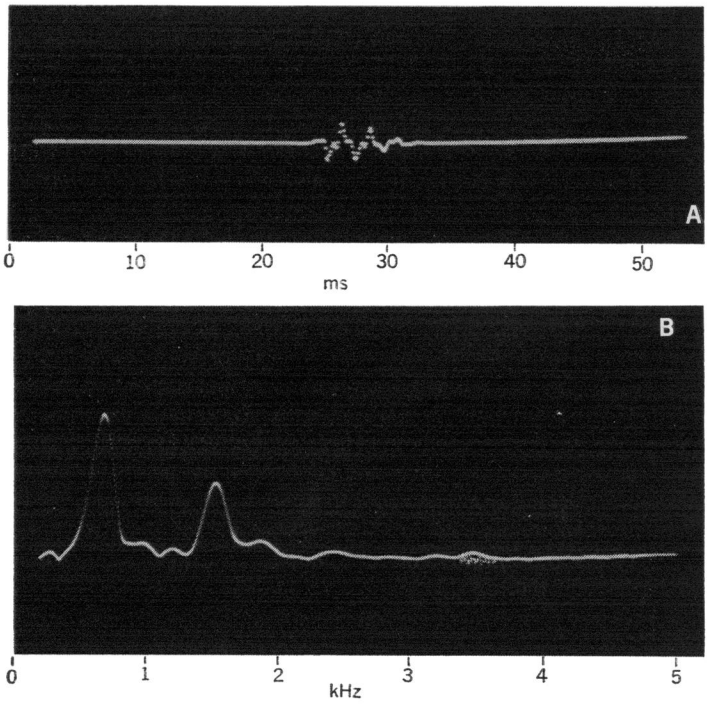


FIGURE 10. A—Speech segment with window applied for wide-band analysis. B—Spectral section from (A).

Figures 7 and 8 show some typical spectrograms obtained with the foregoing system. The sentence again is "He took a walk every morning," as spoken by a male. Figure 7 corresponds to narrow-band spectrograms: In Fig. 7(A), the parameter γ is unity and no frequency shaping has been applied; Fig. 7(B) corresponds to no frequency shaping but a γ of 2/3; Fig. 7(C) represents a γ of unity and linear shaping starting at 1.25 kHz with a slope of 1.6 (1/kHz). Figure 8 represents wide-band spectrograms: Fig. 8(A) corresponds to a γ of unity and no frequency shaping; Fig. 8(B) to no frequency shaping but a γ of 2/3; and Fig. 8(C) to a γ of unity and linear shaping starting at 1.25 kHz with a slope of 1.6 (1/kHz). Figure 8(D) is identical to Fig. 7(C), except for an expanded time scale. In Fig. 9(A) is a typical section of input after the time window has been applied; Fig. 9(B), which shows the magnitude of the DFT of that section, using the narrow band, represents a vertical pass across Fig. 7 at the abscissa marked by the arrows. Similarly, the time function and spectral cross section shown in Fig. 10 are for the wide-band spectrogram, and correspond to the arrows in Fig. 8.

Advantages of FFT-generated spectrograms

This article has discussed and illustrated a procedure for generating and displaying speech spectrograms on a digital computer. In the present system, the analysis time is roughly comparable to that obtainable with modern analog spectrographic equipment. Because of the inherent capabilities of digital computers, there is the potential for considerable flexibility with this method.

The primary advantages of obtaining spectrograms digitally are (1) the increased flexibility and (2) the ability to carry out on-line spectrographic analysis of speech that is being synthesized or processed digitally for other reasons. At present, it does not seem to be efficient or advantageous to carry out *routine* spectrographic analysis of *large* amounts of speech data on a sophisticated computer facility. However, as the cost of small computers and digital hardware decreases, it may eventually be practical and economical to reserve a small facility for preliminary analysis of speech signals, including displays of spectrums and time waveforms.

The author wishes to acknowledge the assistance of Don Johnson and Mrs. Ann Fried.

The work reported in this article was carried out in part at the M.I.T. Lincoln Laboratory (sponsored in part by the U.S. Air Force), and partly at the M.I.T. Research Laboratory of Electronics [supported in part by the U.S. Air Force (Office of Aerospace Research, under contract F19628-69-C-0044)].

REFERENCES

1. Koenig, W., Dunn, H. K., and Lacey, L. Y., "The sound spectrograph," *J. Acoust. Soc. Am.*, vol. 18, pp. 19–49, 1946.

2. Potter, R. K., Kopp, G. A., and Green, H. C., *Visible Speech.* Princeton, N.J.: Van Nostrand, 1947.

3. Kersta, L. G., "Amplitude cross-section representation with the sound spectrograph," *J. Acoust. Soc. Am.*, vol. 20, pp. 796–801, 1948.

4. Denes, P. B., and Mathews, M. V., "Laboratory computers: their capabilities and how to make them work for you," *Proc. IEEE*, vol. 58, pp. 520–531, Apr. 1970.

5. Schroeder, M. R., "Computers in acoustics: symbiosis of an old science and a new tool," *J. Acoust. Soc. Am.*, vol. 45, pp. 1077–1088, May 1969.

6. Flanagan, J., *Speech Analysis, Synthesis and Perception.* New York: Academic Press, 1965.

7. Cooley, J. W., and Tukey, J. W., "An algorithm for the machine computation of complex Fourier series," *Math. Comput.*, vol. 19, pp. 297–301, Apr. 1965.

8. Gold, B., and Rader, C., *Digital Processing of Signals.* New York: McGraw-Hill, 1969.

9. Bergland, G., "A guided tour of the fast Fourier transform," *IEEE Spectrum*, vol. 6, pp. 41–52, July 1969.

10. Rothauser, E., and Maiwald, D., "Digitalized sound spectrography using FFT and multi-print techniques (abstract)," *J. Acoust. Soc. Am.*, vol. 45, p. 308 ,1969.

11. Arnold, C. E., "Spectral estimation for transient waveforms," presented at IEEE Arden House Workshop, Jan. 1970.

Alan V. Oppenheim (M) received the S.B. and S.M. degrees in 1961 and the Sc.D. degree in 1964, all in electrical engineering, from the Massachusetts Institute of Technology. From 1961 to 1964 he was a member of the M.I.T. Research Laboratory of Electronics and an instructor in the Department of Electrical Engineering. During this period his main activities centered on system theory and communication theory. His doctoral research involved the application of modern algebra to the characterization of nonlinear systems. In 1964 he became an assistant professor in the M.I.T. Department of Electrical Engineering and a staff member of the Research Laboratory of Electronics; in 1969 he was appointed to an associate professorship, after spending 1967–69 as a staff member of Lincoln Laboratory. His present research activities are in the areas of speech communication and digital waveform processing. Dr. Oppenheim is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, and the Acoustical Society of America. With coauthors R. W. Schafer and T. G. Stockham, he received the 1969 IEEE G-AE Senior Award.