Computer-based sound spectrograph system

William J. Strong and E. Paul Palmer

Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602
(Received 8 January 1975; revised 17 June 1975)

Sound spectrographs based on digital computers are of interest because of the versatility they offer in generating and displaying sound spectrograms. A particular computer-based sound spectrograph is described in which digitized time waveforms stored on disk memory are spectrally analyzed and the resulting data stored back on disk for later display. Analyses are also performed to obtain the voicing fundamental frequency and the overall-intensity level of the signal which are displayed time registered with the spectrogram. While viewing the spectrographic display, the user is able to tailor the display before making hardcopy. This permits the user to specify a spectral threshold level, the dynamic range represented in the display, spectral contrast, and spectral emphasis.

Subject Classification: 70.62, 70.40; 85.32.

INTRODUCTION

The sound spectrograph has been and continues to be a very significant tool for work in the acoustical aspects of speech. Sound spectrographs based on digital computers are of interest because computers are being widely used for speech processing and small computers capable of sophisticated analyses of speech form the basis of many modern speech research facilities.[1,2] In many instances, it is desirable to be able to generate sound spectrograms on line because of the inconvenience of doing it off line or because a sound spectrograph is not available. Computer-controlled generation and display of spectrograms offers versatility and flexibility, and furthermore, other information, such as voicing fundamental frequency and overall intensity, can be time registered with the spectrogram and included as part of the hard copy.

A particular computer-based sound spectrograph is described in which digitized time waveforms stored on disk memory are spectrally analyzed and the resulting data stored back on disk for later display. Analyses are also performed to obtain the voicing fundamental frequency and the overall-intensity level of the signal which are also stored. While viewing the spectrographic display produced from the stored data, and before making hard copy, the user is able to tailor the display by specifying spectral emphasis, spectral threshold level, spectral contrast, and the effective dynamic range represented in the display.

I. SYSTEM DESIGN

A. Basic system operation

Speech signals to be processed are low-pass filtered at 4.5 kHz, sampled at 10 kHz, and stored on disk. The digitized speech samples can then be edited under computer control. Consecutive sectors (256 samples each) are viewed on the CRT under console switch control and stepped through until the beginning of the desired speech sample is found. The operator then specifies "analysis mode" (either narrow- or wide-band) under console switch control. The spectral analyses are then carried out via the fast Fourier transform and the resulting log spectra stored on disk. The fundamental-frequency and overall-intensity analyses are also carried out and the results stored.

When the calculations are complete and the data stored, it is possible for the user to tailor the spectrographic display by specifying certain display parameters, viewing the results, and then modifying the parameters until the display is optimized as desired. This is all accomplished without recomputing the spectra. The display can be run in single-sweep mode for photographing or in continuous-sweep mode for studying. The user can, at this point, continually modify the display parameters or can return to the edit mode and select a new speech sample. Separate calculations must be run for narrow- and wide-band spectrograms.

B. Computer system

In addition to the basic computer, the peripheral equipment necessary to realize a computer-based spectrograph system are an analog-to-digital converter, a mass storage device, and a graphic display unit. The analog-to-digital converter should be capable of operating at rates of at least 10 000 12-bit samples per second. The mass storage device should be capable of data transfer rates several times that of the converter so that it can receive input data without missing any. (An actual design must account for the memory buffer size and the mass storage device latency times.) The graphics unit should be capable of producing several display intensity levels and several thousand grid points.

The results reported here were obtained using a Digital Equipment Corporation PDP-15 computer with 16 k of 18-bit memory, an RP-15 disk-pack unit, a VT-15 graphics unit, and an analog-to-digital converter. Several of the most-used subroutines were written in machine language for integer arithmetic to provide increased computational speeds, because this system does not have floating-point hardware.

C. Supplementary data and time registration

In addition to information in a sound spectrogram, spectral sections and data on voicing fundamental frequency and overall-intensity level are often important. Furthermore, it is desirable that all of the parameters be time registered so that they can be studied relative to one another. Because the data base is digitized and

time registered on the disk, it is possible to calculate all parameters time synchronously. Spectral sections are calculated via a separate spectral analysis program operating on the same data base for any time position in the spectrogram.

The signal intensity level is calculated at 5 msec intervals by windowing the signal in a 256-point Hamming window and calculating the log of the mean-square signal.

The voicing fundamental frequency is calculated at 5-msec intervals using a method described by Gold and Rabiner.[4] To avoid displaying a frequency for low-level background noise, the system user can specify a sound intensity level (in decibels) below which display of the fundamental-frequency contour is inhibited. This is the parameter labeled FCO for "fundamental cutoff" on the photographs which follow. Display of the fundamental frequency is also inhibited when the sound is judged to be unvoiced.

D. Time and frequency resolution in the analysis

A fast Fourier transform spectral analysis is used to provide time resolution in the wide-band analysis and frequency resolution in the narrow-band analysis that are roughly comparable to the typical analog case. A 7.5-msec Hamming window with a half-power bandwidth of approximately 190 Hz is used in the wide-band analysis. A 40-msec window of half-power bandwidth 36 Hz is used for narrow-band analysis. The FFT employs 128 points or 512 points and is stepped 2.5 msec or 10 msec for wide- or narrow-band analysis, respectively. Additional details of analysis theory and procedures are available in Oppenheim[1] and Mermelstein[2] and the references cited by them.

E. Dynamic range of display

One difficulty with both analog sound spectrographs and digitally simulated sound spectrographs is a too-small dynamic range of the display medium. Contour plots have been developed to overcome the limited dynamic range of the facsimile paper of an analog spectrograph.[3] There are several ways of producing different apparent brightness levels on recording film from a CRT: by varying the intensity or the time duration of a displayed point or by varying the number of points turned on in a point matrix.

Of the seven nonzero display intensities available on the graphics unit used in this research, only four could be successfully distinguished on film. If the lowest intensity was made visible, then the four highest saturated the film, and all appeared equally bright. If the f stop of the camera was changed to make the upper intensities distinguishable, then the three lowest could not be seen.

Another scheme for controlling apparent brightness was tried in which each apparent point was represented by a matrix of eight display points. Apparent brightness was controlled by turning on from 1 to 8 points in the matrix, each at a fixed intensity. It was determined experimentally that there were seven distinguishable brightness levels with this scheme. A good rule of thumb seemed to be that the number of matrix points turned on had to increase by at least one-third in order for a new distinguishable level to be produced. This led to a selection of 0, 1, 2, 3, 4, 6, or 8 matrix points turned on to produce the seven brightness levels. An annoying problem resulting from this approach was the apparent graininess in the pictures. The spatial resolution of the CRT and camera and film combination was good enough so that the 8-point matrix did not appear as a single point with variable gray level, but the internal structure of the matrix was apparent to some extent.

The scheme finally adopted uses as 8-point matrix with all points turned on at the same intensity. Seven different nonzero brightness levels are produced by using four different display intensity settings either singly or in combination as in Table I. (For example, 5+4 indicates that the 8-point matrix is displayed with the intensity setting at 5 and then the same 8 points are displayed with the intensity at 4.)

A matrix of 8 points vertical (frequency) by 1 point horizontal (time) was used for the wide-band displays where good time resolution was the primary goal. A matrix of 2 points vertical by 4 points horizontal was used for the narrow-band displays where better frequency resolution was desired.

F. Spectral levels, shaping, and contrast

With seven distinguishable brightness levels, it is possible to capture much detail in a spectrogram especially if simple level manipulation and spectral shaping are used.

The user is given the option of assigning the number of decibels (in the speech spectrum) per brightness level of the display, somewhat like the γ factor of Oppenheim.[1] This is controlled by the parameter labeled STP (for step size) on the spectrograms that follow. As an example, if 5 dB per brightness level is the specification, then levels up to 4 dB above threshold will be assigned brightness zero, levels from 5 to 9 dB above threshold will be assigned brightness one, and so on. This feature does not increase the dynamic range of the display medium, but rather serves to incorporate a greater or lesser dynamic range of the acoustical signal within the available dynamic range of the display medi-

TABLE I. Nonzero brightness levels produced by using four different display intensity settings either singly or in combination.

Apparent brightness level	Display intensity setting
0	0
1	4
2	4+4
3	5
4	5+4
5	6
6	6+5
7	7

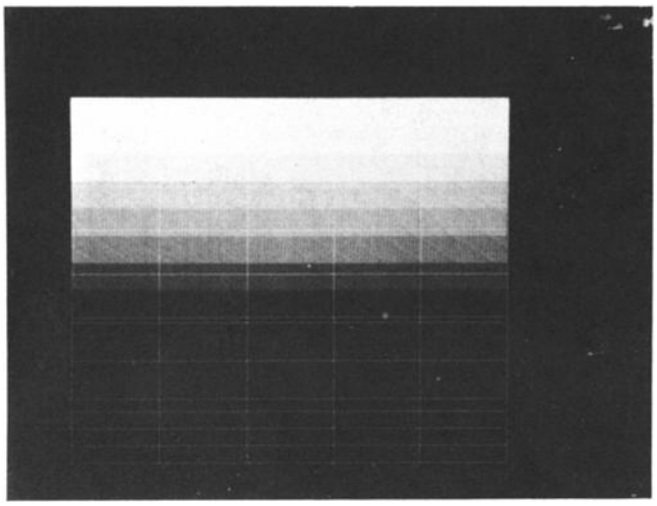

FIG. 1. Display test pattern.

um much as is done with contour plots in modern analog spectrographs.

In order to make the best use of the dynamic range available, a sound-intensity-level threshold specification is provided. This threshold defines the level relative to which the dynamic range begins and permits elimination of background noise without sacrificing part of the dynamic range of the display. This parameter is labeled THR (for threshold) in the spectrograms that follow.

A feature which provides a higher-contrast is the option to inhibit some of the lower brightness levels of the display. This makes certain features stand out more strongly at the expense of giving up some of the subtle details. This is controlled by the parameter SKP (for number of lower brightness levels to skip) shown on the spectrograms.

The spectral shaping method adopted is specified in terms of a starting frequency (in units of 100 Hz) and a high-frequency emphasis (in units of decibels per octave). In all cases, the spectrum is calculated from a waveform without pre-emphasis and the spectral shaping takes place only at the time of display. This shaping is controlled by the parameter SEF (spectral emphasis frequency) and SEA (spectral emphasis amount) as listed on the photographs.

II. RESULTS AND DISCUSSION

A. Details of spectrograph display

By means of computer console switch control, a test pattern is displayed on the CRT so that the lens opening of the camera can be adjusted to make the full dynamic range of the display available. The test pattern shown in Fig. 1 illustrates the seven nonzero brightness levels of the display. As noted above, the operator can specify either a wide- or narrow-band analysis and then continually modify the display parameters to optimize the display results. The spectrograms shown in Figs. 2–7 each represent 2.5 sec of speech over a frequency range of 0–5 kHz. A computation time of 2.5 min is required to compute the spectral data and 2.0 min for fundamental frequency and intensity data. About 5 sec is required to write the data on the CRT. The spectrograms were recorded on Polaroid type 47 film at $f/11$ using time exposure.

The label above the upper left-hand corner of the spectrogram is an arbitrary nine-character label input via the teletype for each speech sample. The numerical values assigned to the six display parameters appear below the parameter labels above the spectrogram. The vertical lines on the spectrograms are placed at 0.5-sec intervals and the horizontal lines at 1-kHz intervals. The display grid was 512 points vertical by 1000 points horizontal with 8-by-1-point matrices arrayed 64 in the vertical and 1000 in the horizontal for wide-band displays and 2-by-4-point matrices arrayed 256 in the vertical and 250 in the horizontal for narrow-band displays. The overall intensity level display at the bottom of the picture has horizontal lines placed 20 dB apart.

The fundamental frequency display just above the intensity has horizontal lines whose spacing depends on the average fundamental frequency. The average fundamental frequency is calculated by averaging the fundamental frequency values for all samples in the 2.5-sec utterance that are judged to be voiced and that correspond to intensity levels above the cutoff value specified by FCO. The labelling number in the display is the average frequency and it is also the range of the fundamental-frequency display. The middle line of the display is placed at the average frequency, and the upper and lower lines are placed at 1.5 and 0.5 times the average, respectively. Fundamental frequencies greater than 1.5 times the average or less than 0.5 times the average are not displayed.

B. Specific examples

An example of a narrow-band spectrogram produced without spectral emphasis is shown in Fig. 2. The utterance is, "The second planet was inhabited by a con-

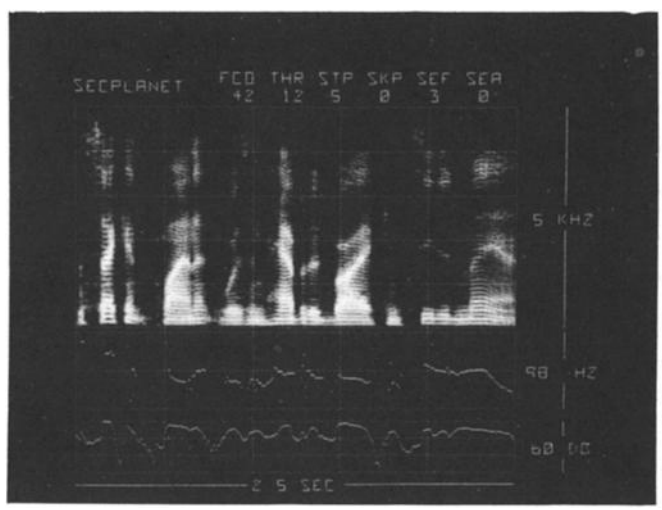

FIG. 2. Narrow-band spectrogram of the utterance, "The second planet was inhabited by a conceited man." No spectral emphasis was used. Step size (STP) is 5 dB per brightness level.

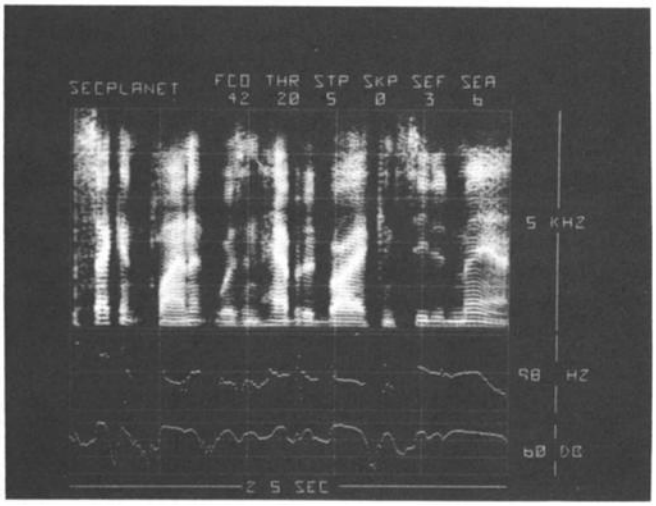

FIG. 3. Same as Fig. 2 except a spectral emphasis (SEA) of 6-dB octave staring at 300 Hz (SEF) was used.

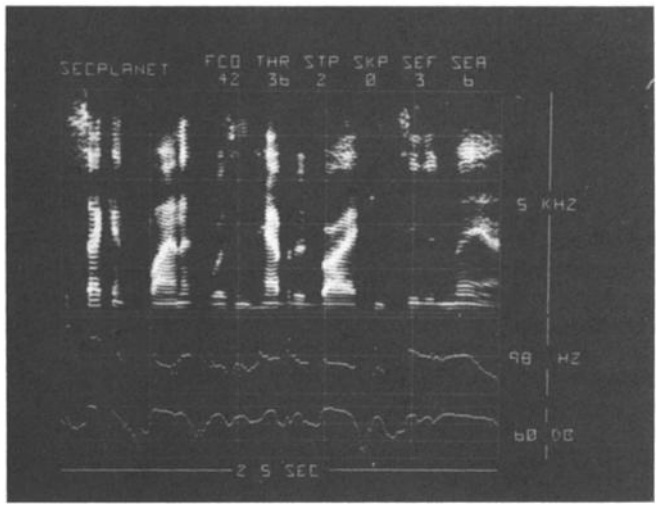

FIG. 5. Same as Fig. 3 except STP = 2 dB per brightness level.

ceited man." With a step size (STP) of 5 dB per octave, a dynamic range of 40 dB (eight brightness levels at 5 dB) is represented. It is apparent that much detail is captured even without spectral emphasis.

Figure 3 illustrates the effect of a simple spectral emphasis of 6 dB per octave (SEA) beginning at 300 Hz (SEF). Note that the threshold (THR) has also been raised to compensate for the increased background at the higher frequencies due to the spectral emphasis.

A too-low specification of threshold (THR) is illustrated in Fig. 4 which has the same parameter specifications as Fig. 3 except for threshold. Note that part of the dynamic range is wasted with this too-low value for threshold.

Figure 5 shows a spectrogram in which the step size (STP) is set to 2 dB per brightness level for a dynamic range of 16 dB, which presumably approximates the typical analog situation where contour plots are not used.

Without changing other settings it is possible to produce a more contrasty result by inhibiting some of the lower brightness levels. This is illustrated in Fig. 6, where three lower brightness levels are inhibited via parameter SKP as compared with Fig. 3.

Figure 7 illustrates a wide-band spectrogram with spectral emphasis and with STP set to 5 and corresponds to the narrow-band spectrogram of Fig. 3. The vertical striations that appear well defined in analog-produced wide-band spectrograms of many male talkers are not so well defined in the computer-produced wide-band spectrogram. This is due in part to the lack of resolution in the CRT display and in part to the discrete time steps of 2.5 msec in the computation of the individual spectra.

Figures 8–10 show three narrow-band sections produced by a separate program from the same digitized waveforms. The frequency range of a section is 0–5 kHz. The labelled dynamic range of the section displays is 100 dB; however, the actual dynamic range of the system is probably more like 50–60 dB. A 50-msec, Hamming-windowed sample was analyzed for each sec-

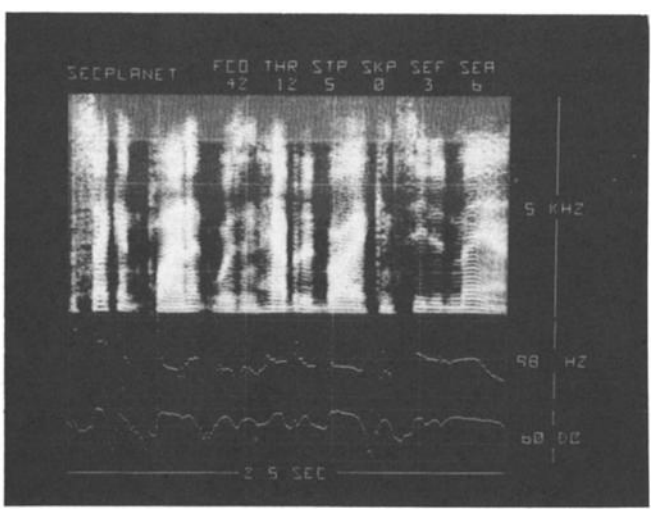

FIG. 4. Same as Fig. 3 except a lower value was specified for threshold (THR); 12 dB compared with 20.

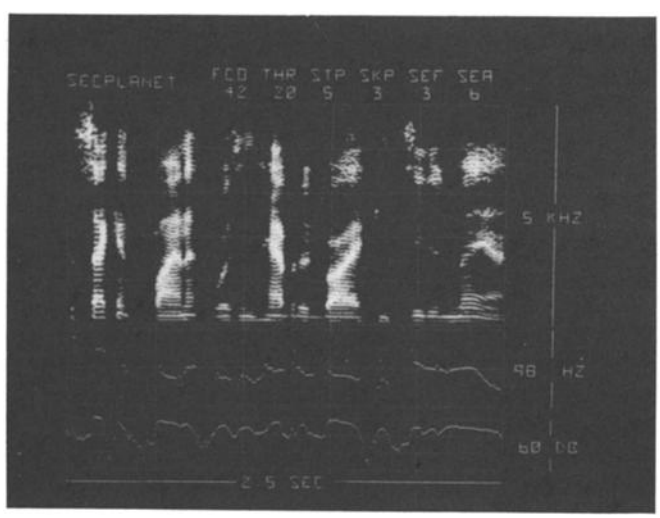

FIG. 6. Same as Fig. 3 except lower three brightness levels have been inhibited (SKP = 3).

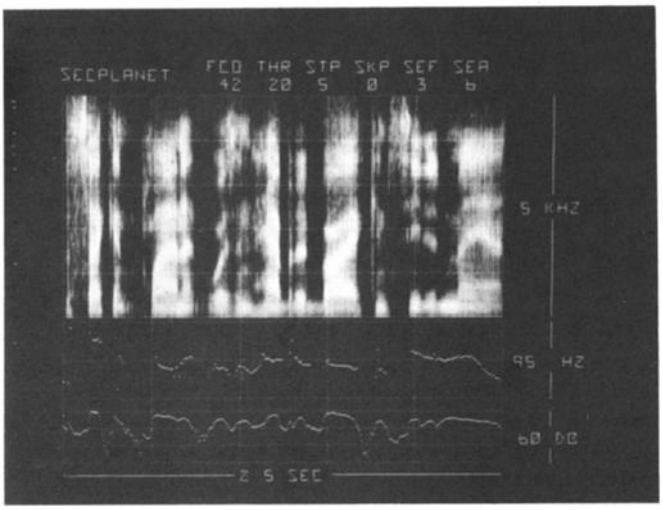

FIG. 7. Wide-band spectrogram with 6 dB per octave spectral emphasis starting at 300 Hz. Compare with Fig. 3.

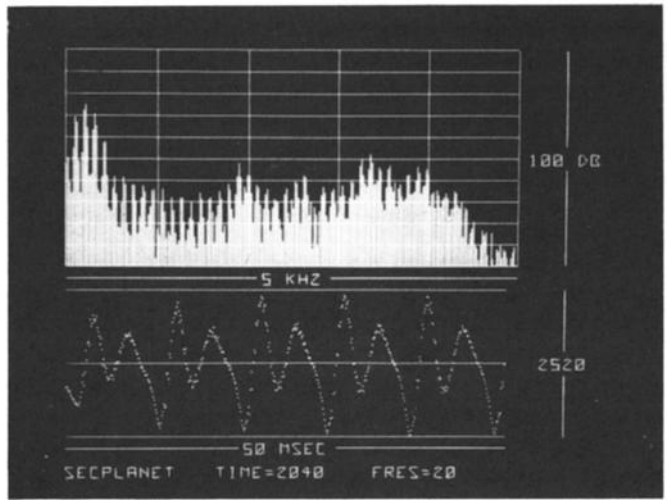

FIG. 9. Narrow-band section of /i/ in "conceited."

tion shown here but this is variable and under user control. The numerical label at the right-hand side of the time waveform in each case is a maximum peak-to-peak relative amplitude for the sample being analyzed.

Figure 8 is a section of the /s/ in "second" taken at a time of 100 msec from the beginning of the spectrogram. Figure 9 is a section of /i/ in "conceited" taken at a time of 2040 msec. Figure 10 is a section of /æ/ in "man" taken at a time of 2400 msec with some nasal influence apparent.

C. Applications

The system has been used to study several different data bases of digitized speech. Applications have included the following: (1) demonstration of spectral and prosodic features of speech for students in a descriptive acoustics class; (2) illustration of spectral differences between similar Spanish and English vowels for English-speaking students in a Spanish class; (3) study of prosodics to determine appropriate control paramaters to be used in junction-phonology speech synthesis-by-rule; (4) illustration of different spectral and rhythmic characteristics of Korean as spoken by a native Korean, a native English speaker who had learned to speak Korean, and a native English speaker who was trying to speak Korean for the first time.

Users have reacted favorably to the time-registered fundamental frequency and intensity-level displays. In some instances users have preferred to use these displays to the exclusion of the spectrogram. However, in most applications the ability to study these things in conjunction with the spectrogram has proven very useful.

III. CONCLUSIONS

The sound spectrograph system described above computes and displays sound spectrograms in times that are roughly comparable to those required with analog equipment. In addition, it provides voicing-fundamental-frequency and overall-intensity level plots that are time registered with the spectrogram. Spectral sections are

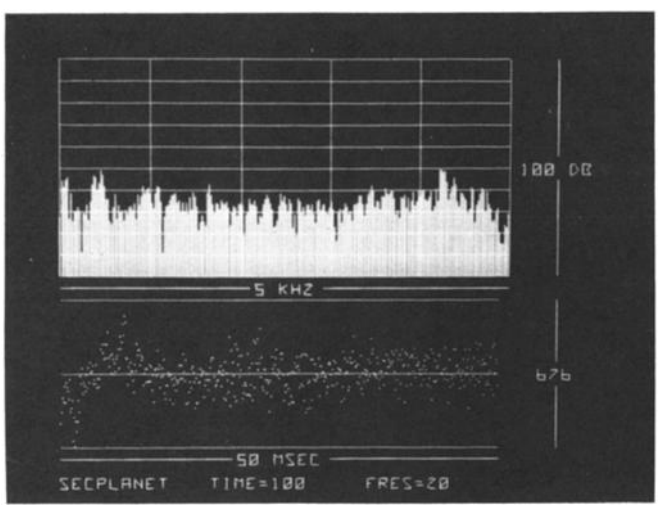

FIG. 8. Narrow-band section of /s/ in "second." Utterance is the same as above.

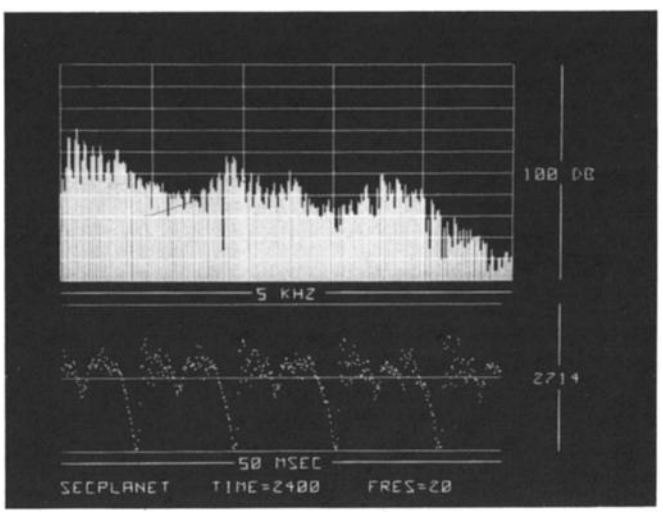

FIG. 10. Narrow-band section of /æ/ in "man."

computed for any given portion of the signal with a separate program. The variable effective dynamic range of the system is adequate to capture much of the high-frequency information in the spectra even when no spectral emphasis is used. However, spectral shaping brings out additional detail and is very useful in practice. Probably the prime virtue of the system, other than its variable effective dynamic range, is the capability it gives a user to tailor a spectrogram before making a hard copy. This tailoring permits specification of a spectral threshold, the effective dynamic range, a contrast effect, and the extent of spectral shaping.

The system is of interest because of the wide use of computers in speech processing and because of the flexibility they offer in computing and displaying speech parameters. The system can be easily implemented on computer-based speech processing systems with the requisite hardware. The primary software requirement is a FFT to provide reasonable efficiency in the spectral computations. The 5-kHz analysis range can be modified to provide greater flexibility.

[1]A. V. Oppenheim, "Speech Spectrogram Using the Fast Fourier Transform," IEEE Spectrum 7(8), 57–62 (1970).
[2]P. Mermelstein, "Computer-Generated Spectrogram Displays for On-Line Speech Research," IEEE Trans. Audio Electroacoust. AU-19, 44–46 (1971).
[3]A. J. Prestigiacomo, "Amplitude Contour Display of Sound Spectrograms," J. Acoust. Soc. Am. 34, 1684–1688 (1962).
[4]B. Gold and L. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am. 46, 442–448 (1969).