Copyright © 1982 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 5, May–June 1982
Printed in U.S.A.

Graphic Displays of Combined Presentations of Acoustic and Articulatory Information

By J. E. MILLER and O. FUJIMURA

(Manuscript received August 17, 1981)

Articulatory data have been collected by a computer-controlled microbeam system. The locations of lead pellets placed on a speaker's articulators are recorded as x, y coordinates along with the acoustic signal and a train of timing pulses which synchronize the frames of X-ray data with the acoustic signal. Some 830 utterances have been recorded to date, resulting in a database of nearly 120 million bytes. We describe the graphic techniques by which those data are examined; namely, a simultaneous display of articulatory features such as velum lowering, lip and jaw closure, and tongue motion, together with the spectral information of the corresponding speech. These displays are also annotated with phonetic transcriptions obtained automatically by pattern-matching techniques. Acoustic representations of speech signals, in a simplified spectrographic format, are contrasted with articulatory measures, and implications are discussed concerning the difficulty of the automatic recognition of speech via conventional input processing.

I. INTRODUCTION

Articulatory data specified by x,y coordinate positions of lead pellets located on the velum, tongue, lip and jaw, all in the midsagittal plane, have been collected by a computer-controlled X-ray microbeam system. The corresponding speech signal and a sequence of timing pulses which synchronize the frames of articulation data with the acoustic data are also recorded. This system, which operates at the University of Tokyo, has been described in detail by Fujimura, Kiritani, and Ishida[1] and Kiritani, Itoh, and Fujimura.[2] In a typical data collection session an utterance lasts about six seconds, and approximately 700 frames of coordinate data on eight pellets are obtained. If fewer pellets are used, the frame rate will be higher since pellet positions are determined sequentially for each frame. The waveform of the acoustic signal is quantized into 12-bit samples at a 10-kHz sampling rate, and each frame of pellet data is marked by an integer pointing to the corresponding acoustic waveform sample. All quantities, i.e., pellet coordinates, samples of the acoustic wave, and frame pointers, are stored as 16-bit quantities, and thus a 6-second utterance requires approximately 144,000 bytes. To date there have been some 830 utterances in two languages—American English and Japanese by a total of six different speakers (one female). The resulting database totals nearly 120 million bytes.

II. ON-LINE CRT DISPLAYS

It is important in dealing with a large database that it be possible to retrieve and examine the data easily. The initial methods employed have been of an on-line, interactive nature. Utterances are identified by a 5-digit number, which indicates the data-collection session and position within the sequence of utterances recorded. Specification of this ID number at the computer console results in the retrieval from disk storage of all data pertinent to the utterance. Then a time window is selected under knob control for a CRT display of the time traces of the pellets, together with the RMS envelope of the speech wave. A copy of such a display is shown in Fig. 1. The ID number and the first and last frames of the window are specified on the top line. The text of the utterance is printed below, and average values of the pellet traces, together with their two-letter designators, such as BY for the y (up-down) coordinate of the pellet on the tongue blade, are listed on the right for up to six coordinates. The RMS envelope of the acoustic signal is shown at the bottom. Alternatively, an x-y coordinate map showing trajectories of movement in the pellet positions is displayed for the frames included in the time window. (See Fig. 2.) The window size has been cut down to the 45 frames spanning the portion of the text indicated by the bracket. The trajectory of each pellet begins at the point marked by an arrow and the final position is marked by an X. This type of display very clearly reveals the velum lowering for the nasalization, as well as the complex activity of the tongue associated with this portion of the sentence. The vertical dots represent the position of a cursor, which can be moved by a knob through the windowed frames. On-line digital-to-analogue facilities permit listening to the associated sound for time intervals of varying lengths. Generally, these display features have proved very useful in studying the data, but there is an inherent inadequacy in the representation of the speech signal by waveform envelopes. To combat this shortcoming, we have supplemented the output facilities with a spectrographic display of the acoustic information.

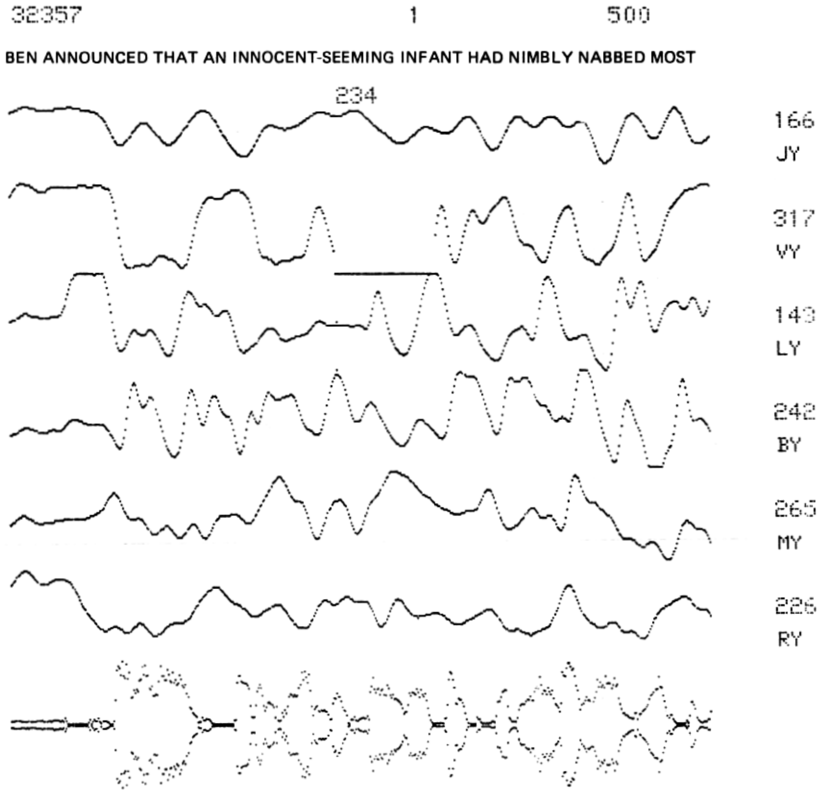


Fig. 1—On-line display of time traces.

III. HARD COPY DIGITAL SPECTROGRAMS

To obtain the resolution necessary to generate an eight-level gray scale, we have employed a Versatec plotter having 200 rasters per inch. We define a super pel or picture element as a matrix of 4 by 4 rasters and let the number of black rasters out of the sixteen in each super pel be a nonlinearly increasing function of the gray level. Specifically, for levels 0 through 7 the number of black rasters is 0, 1, 2, 3, 4, 7, 10, 16. (See Fig. 3.) The original resolution of 200 rasters per inch is thus decreased to 50 pels per inch, but the resulting gray scale proves fairly uniform and is quite sufficient to display signal intensity in the frequency bands of the digital spectrograms. Figure 4 illustrates an example that adopts the time-frequency proportions of the traditional spectograms.[3]

These displays are produced by using a raised cosine on a time window of 100 samples, or 10 milliseconds, and computing a spectral slice by fast Fourier transform techniques every 40 samples, or 4

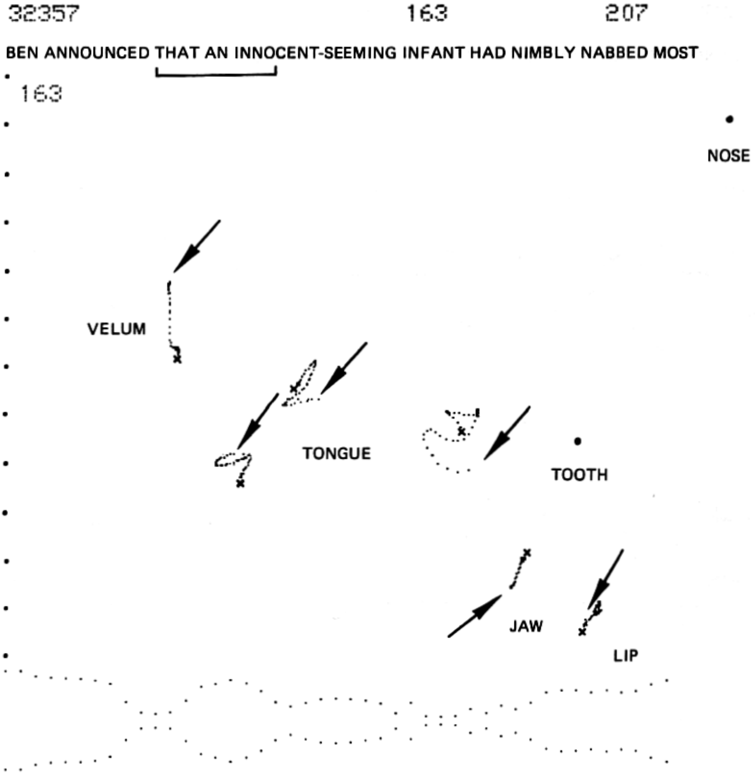
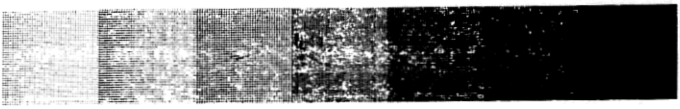

32357
163
207
BEN ANNOUNCED THAT AN INNOCENT-SEEMING INFANT HAD NIMBLY NABBED MOST
163
NOSE
VELUM
TONGUE
TOOTH
JAW
LIP

Fig. 2—On-line display of *x*-*y* trajectories (lateral view).



LEVEL:	0	1	2	3	4	5	6	7
BITS:	0	1	2	3	4	7	10	16
	0000 0000 0000 0000	0000 1000 0000 0000	1000 0000 0001 0000	0001 0000 1000 0010	0010 1000 0001 0100	1001 0110 0010 1001	0110 1011 1101 0110	1111 1111 1111 1111

GRAY-LEVEL RESOLUTION – 5Ø PELS PER INCH.

ONE PEL = 4 x 4 RASTERS.

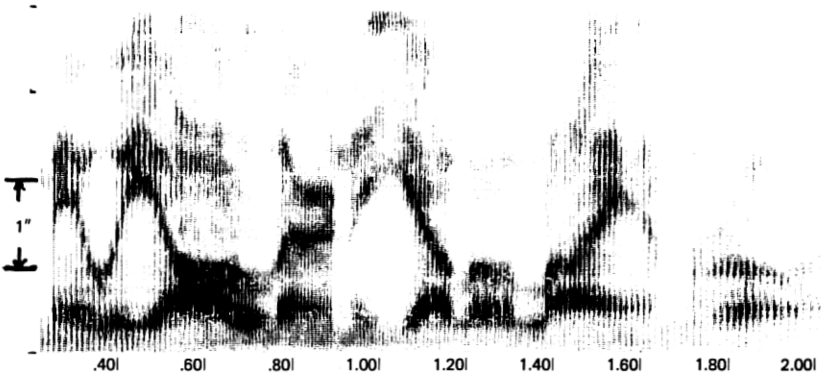
Fig. 3—Gray scale.

milliseconds. Each slice is "boosted" by a linear function to raise the high frequencies, and all amplitudes above an adjustable threshold are linearly mapped onto the eight gray levels. Having a resolution of 50 pels per inch, a sampling rate of 10 kHz and taking slices every 40 samples results in a time scale of 0.2 second per inch. In Fig. 4, one inch of height (on the actual computer output) represents 1000 Hz of frequency. It has been found, however, that great economies in paper space can be achieved by shrinking the frequency (vertical) scale, while leaving the time (horizontal) scale unchanged, and this deviation from the standard visible speech format does not adversely affect the usefulness of the spectrogram. (See Fig. 5 in which there are four 2-second lines of spectral data placed on one page as the result of such frequency scaling.) We find that if we cut the height down to as little as 0.8 inch, we are then able to display a speech spectrogram, together with as many as twelve time functions of pellet coordinates on a hard copy output. (See Fig. 6.) Such a display is substantially more useful than the RMS envelop (as in Fig. 1) for identifying phonetic events in the utterance being examined.

A display of this type can handle at maximum a time window of two seconds, which, at typical frame rates of 120 per second, is approximately 240 frames. The time scale is marked off both by frames (every tenth) and by milliseconds (every 200). Although computed while running the display program on-line (at about 100 times real time with our present CPU), this output is generated off-line on the Versatec plotter.

IV. OFF-LINE OVERVIEW DISPLAYS

It is frequently of considerable convenience to both prepare and examine the data away from the computer, and in so doing, it is advantageous to have a complete representation of as much as an entire utterance on the same sheet of output. The example shown in Fig. 7 meets these requirements. As an off-line or "background job," we are able to prepare on the Versatec plotter an 8-1/2 by 11 inch



/usr5/bsa/spdir/spi10 "May we all learn a yellow lion roar."

Fig. 4—Digital spectrogram in traditional format.

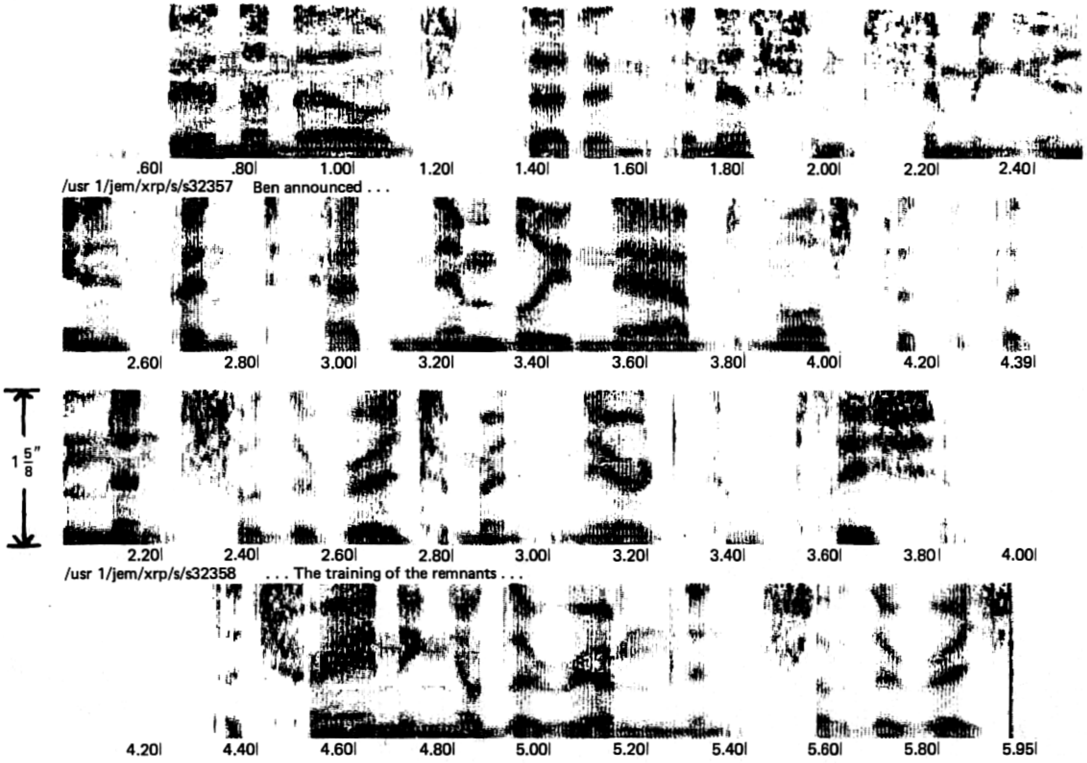


Fig. 5—Digital spectrograms with reduced frequency scale.

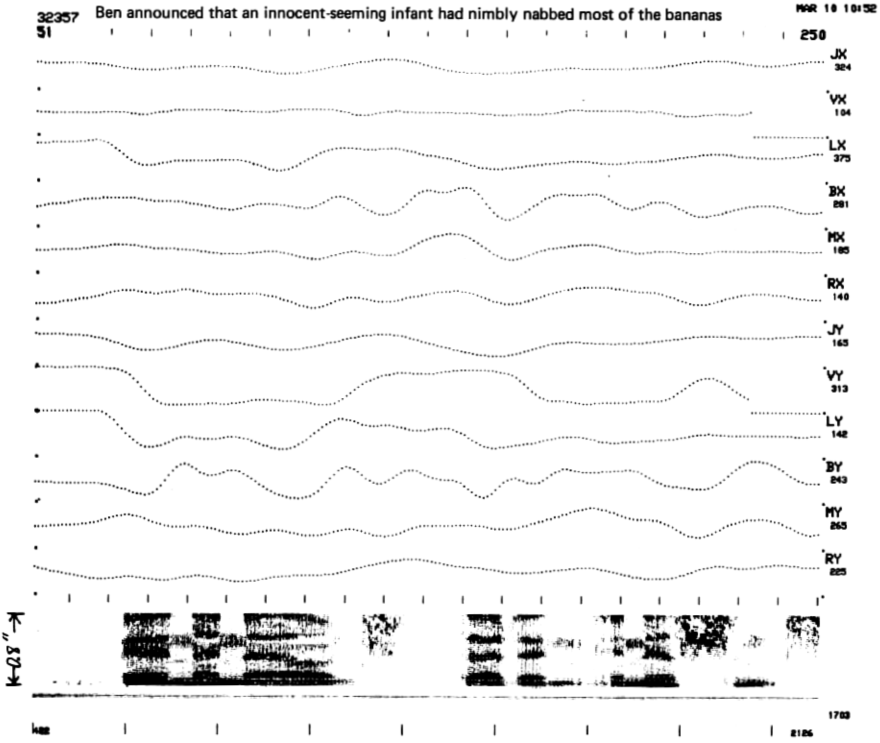


Fig. 6—Time traces and digital spectrograms.

page divided into three horizontal sections, each of which spans two seconds, plus 0.1 second of overlap into the next section. The sections contain ten stripes representing articulatory time functions coded in the eight-level gray scale and plotted in time synchrony with a spectrographic display of the acoustic data. Timing marks at 200-millisecond intervals are also indicated.

The ten gray-level stripes which describe the articulatory events are arranged as follows: There is a top set of four stripes indicating front-back or horizontal motion for, from the top, the lower lip and three tongue pellets (blade, mid and posterior position on the tongue surface), and a lower set of six which portray vertical motion for lower lip, mandible, three tongue pellets, and the velum. The correspondence between stripes and articulators may be determined by associating the stripes ordered from top to bottom on the page with the articulators ordered from front to back in the speaker's head. The range of coordinate values for each pellet is linearly mapped into eight levels of gray such that the minimum is the first level or white, the average value is level five or medium gray, and large values are level eight or

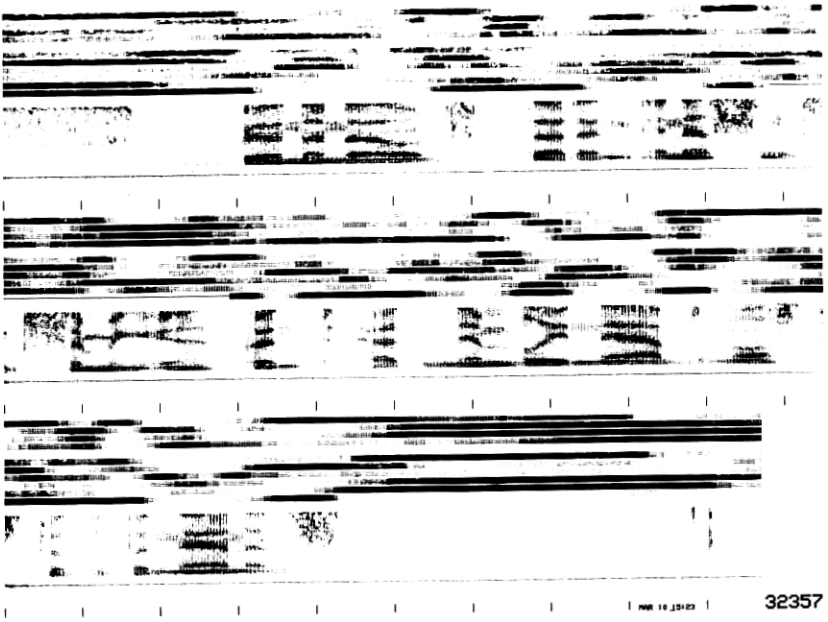


Fig. 7—Off-line display of articulatory and acoustic data.

black. That is, in Fig. 7 black corresponds to up for vertical motion and front for horizontal motion.

V. FEATURE-ENHANCED DISPLAYS

As an extension of the display of the type shown in Fig. 7, which serves as a record and check on the database, we produce others of the type shown in Fig. 8, designed to emphasize articulatory features and contrast articulatory and acoustic events. In these, we reverse the mapping onto the gray scale for vertical motion of both the velum and jaw pellet and horizontal motion of the lip. As a result of these reversals, for example, the nasalization produced by a lowered velum, stress correlated with the minima in the jaw position, and the consonantal gesture indicated by lip constriction and a raised or advanced position of the tongue are easily spotted as black regions in the stripes (see below). It is an option of the program to produce the output with the gray scale for the stripes used in reverse so that when the displays are reproduced as transparencies the stripes can be individually colored. Thus, the features can be readily identified by a color coding.

It will be immediately noticed that the spectrographic representation of the acoustic information has been substantially modified. It has a checker-board appearance due to the frequency range having been

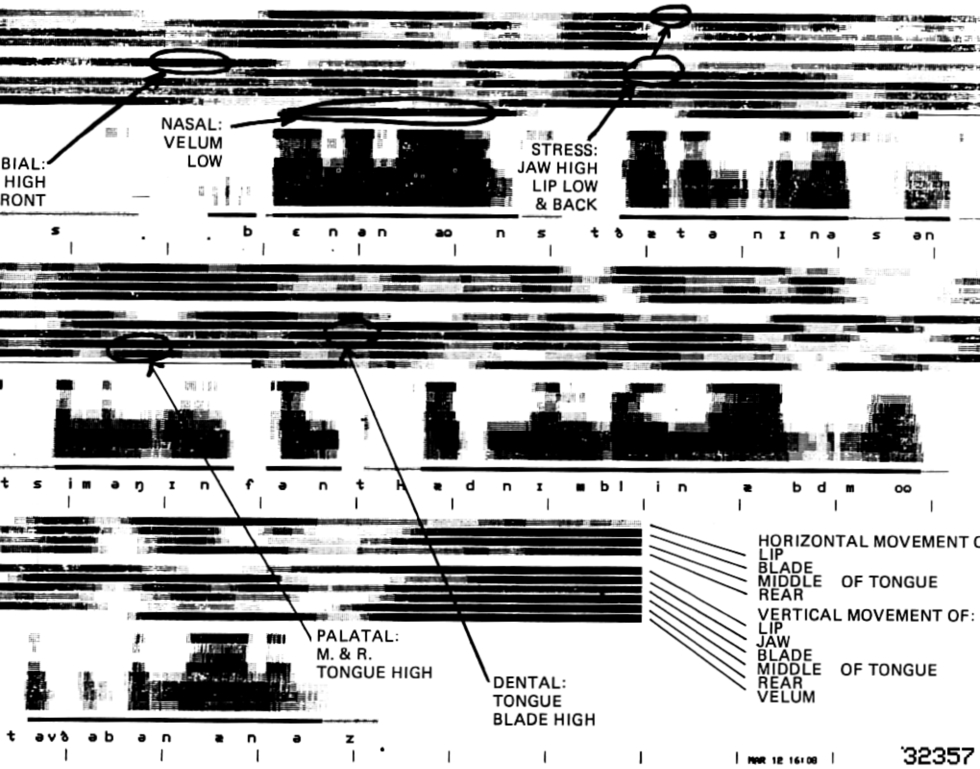


Fig. 8—Feature-enhanced display of articulatory and acoustic data.

grouped into eight bands. This output was produced by calculating a Fourier transform every two milliseconds and taking the maximum amplitude values obtained from three adjacent spectral slices. The maxima are used as an estimate of the spectrum at time points spaced every four milliseconds. The time window for the Fourier transform is five milliseconds; hence, the frequency resolution is 200 Hz, and the first twenty frequency components span a usable frequency range of 4000 Hz. These components are grouped into eight bands centered around 200, 400, 600, 900, 1300, 1800, 2500, and 3500 Hz, respectively. The average output magnitude in each of these bands is then mapped onto the gray scale. It will be noted that with such a grouping, the first formant and the nasal formant will be found in bands 1 to 4, the second in 4 to 6, and the third in 6 and 7. Voiceless fricatives and stops show up often in bands 7 and 8.

The black, gray, and white stripe plotted below the spectral display indicates the voiced, unvoiced, and silence (VUS) coding of the acoustic signal as determined by the Atal-Rabiner algorithm.[4]

Finally, a processing due to Nelson[5] utilizes VUS coding and information on jaw minima to locate phrase boundaries, and then identifies occurrences of specific articulatory events that are predicted from a feature interpretation of a sequence of phonetic symbols obtained from a manual transcription of the utterance material. Each symbol is positioned automatically, according to these pattern identifications, and the resulting symbol alignment is shown in Fig. 8. A VUS stripe and phonetic annotations will also appear on displays of the type shown in Fig. 6 and 7, provided the utterance has been processed with the coding algorithm and the Nelson pattern-matching scheme.

The display of the acoustic data in Fig. 8 represents a reduction by a factor of five over the traditional spectrogram on information to be stored; that is, from 40 harmonics to eight bands per spectral slice (albeit at the expense of more computation for the particular methods we use). But more important is its very "block-like" appearance, which in contrast to the smoothly fluctuating "grayness" of the articulation stripes, reveals apparent "segmentation" of the time events. These blocks make it easier to evaluate the success of the phonetic alignment. Also, since the frequency bands are selected considering the human auditory characteristics, the spectral representation may be similar to what the human perceives.

The marginal notations in Fig. 8 point out examples of the feature enhancement mentioned above. The utterance in this case is "Ben announced that an innocent seeming infant had nimbly nabbed most of the bananas." The nasalization occuring in "Ben announced" is indicated by the black on the bottom of the pellet stripe set. There are numerous other occurrences of nasalization in the sentence also marked by black in the bottom stripe, except where pellet tracking failed during the recording session. The missing pellet data are indicated by the thin line in an interval of approximately 600 milliseconds beginning about two seconds into the utterance (end of first section—beginning of second). Stressed words are revealed by black in the jaw stripe—second from top in a group of six. The feature of stress is also accompanied by a lowering of the lip which is indicated by white in the top stripe of the six, unless the vowel shows lip constriction. Hence, in most cases the white-black combination of these two stripes serves as a very clear cue for the stressed sounds.

Since the horizontal motion of the lip is plotted in the top stripe of the set of four, with black indicating retraction of the lip in this type of display, and since vertical motion is plotted in the top stripe of the set of six, labials such as /b/ and /m/ have black only in the set of six. However, the fricative /v/ in which the lower lip is retracted would be identified by black in both stripes. Color coding in which the same color is used for each articulator in both horizontal and vertical dimensions is helpful in directly identifying such articulatory features.

Apical consonants, namely, /t/, /d/, and /n/, are produced with the tongue tip and blade high behind the teeth and, hence, are revealed by black in the stripe third from top in the set of six.

The palatal sounds produced with a high mid-position of the tongue are similarly marked by black in the fourth stripe. All these articulatory characteristics are utilized in Nelson's automatic annotation.

We can with these displays go beyond the mere identification and location of these articulatory events and observe interesting time relations between the movements of the articulators and the resulting temporal patterns of the acoustic output. For example, we observe the very early onset of the lip closure for /b/ at the beginning of the sentence. Nasalization for the nasal consonants in the beginning phrase extends continuously over several syllables. We can also see how the labialization trails the vowel glides that cause the gesture as in "announced" and "most," where the lip constriction extends well beyond the bounds of the vocalic segment. Palatalization can be seen to extend through all the front vowels in the words "seeming infant," unaffected by the intervening consonants. We note finally the state of the articulators at the conclusion of the sentence and contrast the slow but still changing activity during this silence with the fast, yet smooth, movement during the sound production.

VI. CONCLUDING REMARKS

In conclusion, we point out that the gray scale facilities have made it possible to reveal a great deal of information on a single page of output and to be freed in many cases from the necessity of on-line use of the computer as a means of data inspection. Furthermore, adequate use of the gray scale has made it possible to detect important features in the events and interesting relationships between the acoustic and articulatory events. Such spectrographic information as we observe in Fig. 8 is typical of the type of input used by an automatic speech recognition system; although, the coarse sampling used by most systems often eliminates the sharp discontinuities we see here. In this connection, it is interesting to note that the acoustic information as shown here is in many ways complementary to the articulatory information shown in comparison. Individual consonantal gestures, particularly in terms of place distinction (e.g., /p/ as opposed to /t/ or /k/) are obvious by articulatory measures, but often very difficult to identify via acoustic signals. We need more complex details, such as continuous formant transition patterns seen in Fig. 7, or some ad hoc and context-sensitive feature detection strategies based on such details, for acoustic identification of phonemic characteristics. While the articulatory gestures are sluggish and asynchronous, their phonemic correlates are, for the pertinent dimension, reproducible and often nearly invariant.[6]

After having identified the real invariant characteristics in the articulatory aspects that are only indirectly observable in the acoustic signals, we can ask proper questions as to how we might process such acoustic signals most effectively to derive the necessary phonetic information. To the extent we are impressed by an acoustic representation as in Fig. 8, we can identify the phonetic message by the present machines. To the extent we have to supplement with the other representations, such as direct articulatory or abstract pattern matching of more detailed acoustic information, effective speech recognition schemes remain to be devised.

REFERENCES

1. O. Fujimura, S. Kiritani, and H. Ishida, "Computer Controlled Radiography for Observation of Movement of Articulatory and Other Human Organs," Comp. Biol. Med. *3* (1973) pp. 371–84.
2. S. Kiritani, K. Itoh, and O. Fujimura, "Tongue-pellet Tracking by a Computer Controlled X-ray Microbeam System," J. Acoust. Soc. Am., *57,* No. 6, Part II (June 1975), pp. 1516–20.
3. R. K. Potter, A. G. Kopp, and H. C. Green, *Visible Speech,* New York: Van Nostrand, 1947.
4. B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-24* (June, 1976) pp. 201–12.
5. W. L. Nelson, "Automatic Alignment of Phonetic Transcriptions of Continuous Speech Utterances with Corresponding Speech-Articulation Data," Speech Communication Papers, J. J. Wolf and D. H. Klatt, eds., New York: Acoust. Soc. Amer. (June 1979), pp. 63–6.
6. O. Fujimura, "Temporal Organization of Articulatory Movements as a Multidimensional Phrasal Structure," Phonetica, *38,* Nos. 1–3 (1981), pp. 66–83.