XII. SPEECH COMMUNICATION*

Academic and Research Staff

Prof. K. N. Stevens
Prof. M. Halle
Prof. W. L. Henke
Prof. A. V. Oppenheim
Dr. Mary C. Bateson
Dr. Margaret Bullowa
Dr. A. W. F. Huggins
Dr. D. H. Klatt
Dr. Paula Menyuk
Dr. T. H. Shriner
Dr. J. J. Wolf

Graduate Students

R. E. Albright
Kay Atkinson-King
M. Bilofsky
A. J. Goldberg
J. S. Perkell
J. E. Richards
R. M. Sachs
V. C. Shields, Jr.
V. W. Zue

A. CHOICE OF SPEAKER RECOGNITION PARAMETERS

1. Introduction

In order to recognize an individual from a sample of his voice, a system must measure a number of parameters of that speech sample. A decision, based on the similarities (or lack of them) between the measurement results and stored information on known speakers, can then be made. A previous report[1] discussed the importance and nature of these characterizing parameters. Its main points were the following.

1. It is not enough that the parameters characterize the speaker sufficiently; they should do so efficiently.

2. Efficient parameters would enable us to use simpler classification procedures, or obtain lower classification error, or both.

3. We should base these parameters on known relations between the voice signal and vocal-tract shapes and gestures.

These considerations suggested an approach of measuring only significant aspects of certain segments of an utterance, rather than making general measurements over the full extent of the utterance.

This report summarizes the results of an investigation of speaker-recognition procedures in which the primary purpose was to use acoustic and phonological theory to find acoustic parameters that are both efficient in discriminating speakers and amenable to automatic implementation. The parameters were extracted and studied with the aid of a laboratory digital computer under real-time operator control. They were tested in two speaker-recognition paradigms by means of elementary classification procedures.

2. Data Base, Equipment, and General Procedure

Ten repetitions of 6 short sentences were recorded under good conditions from each of 21 American adult male subjects. The sentences were devised to contain a wide

*This work was supported by the U. S. Air Force Cambridge Research Laboratories, Office of Aerospace Research, under Contract F19628-69-C-0044.

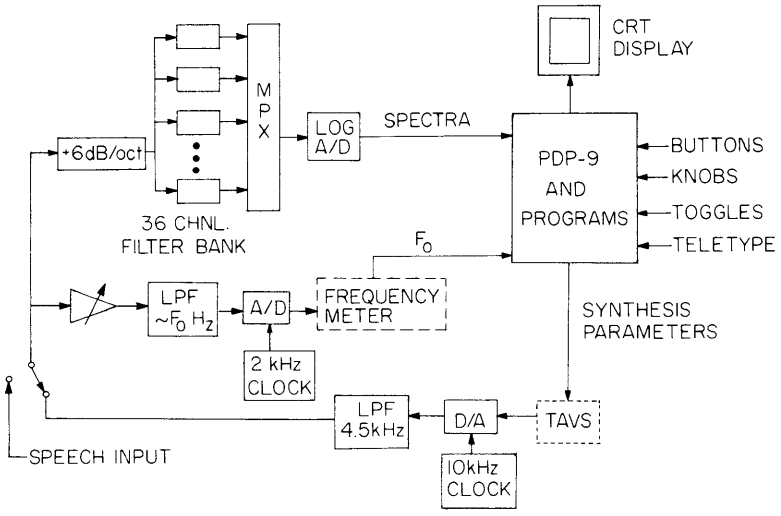

Fig. XII-1. The computer facility used for on-line speech analysis.

variety of potentially useful speech segments. The subjects were asked to speak normally. Recording was done in a single session, so the stability of the parameters with respect to time, health, or stress was not investigated.

The analysis and measurements performed on the utterances were carried out on a highly flexible digital computer laboratory facility designed for on-line speech research.[2] Figure XII-1 depicts the computer system as configured for this research. The boxes drawn in broken lines are functions performed by subprograms rather than by physical devices.

The principal analysis tool is a 36-channel filter bank spectrum analyzer covering 150-7025 Hz. Fundamental frequency is estimated by lowpass-filtering the speech above the first harmonic and measuring the intervals between zero crossings. The lower branch that is illustrated is a vowel synthesizer used in the analysis-by-synthesis scheme that will be described. The operating position of the computer system is generously provided with switches, pushbuttons, and knobs to facilitate interaction with and control of the programs. Most of the output of the analysis programs is by means of a cathode-ray tube display and teletype.

The speech data were kept in analog form on tape. When an utterance was read in, the spectrum and fundamental frequency were sampled every 10 ms and stored in core. A typical display generated by the program is shown in Fig. XII-2. The two graphs in the lower half represent functions of time, from 0 to 2.5 s. The upper one is the sum of the outputs of several low-frequency filters, useful as a "low-frequency energy map" of the utterance; the lower one is fundamental frequency. The vertical cursor, manually controllable by a knob, shows the point in the utterance at which

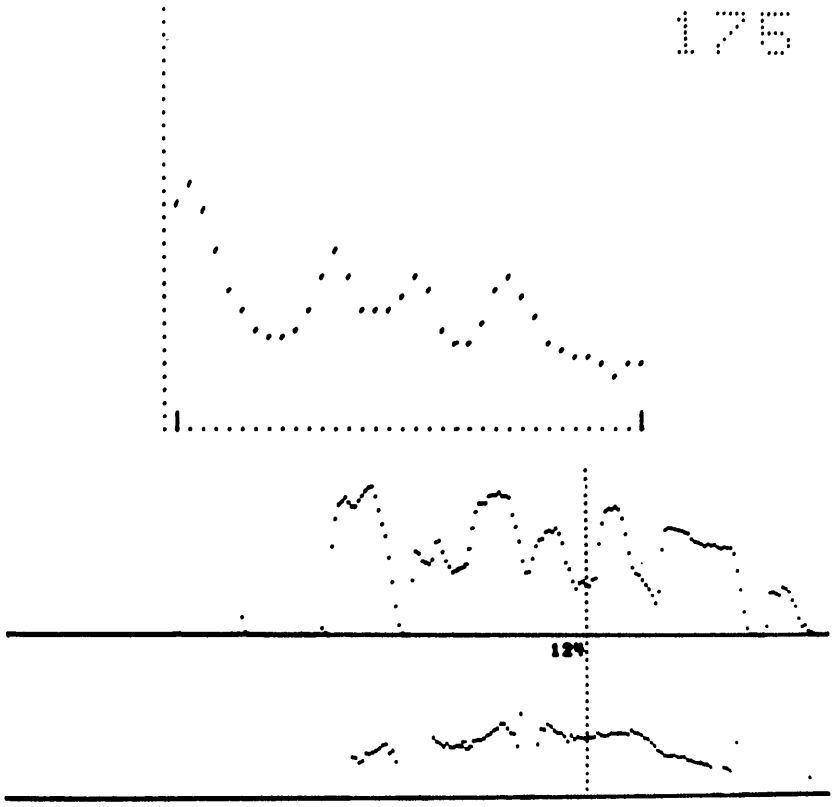

Fig. XII-2. Photograph of a CRT display. The two graphs in the lower half represent low-frequency energy and fundamental frequency as functions of time. The vertical cursor shows the point in the utterance corresponding to the spectrum above. The points on the vertical axis of the spectrum represent 2-dB steps in amplitude; each horizontal point represents one of the 36 filter outputs. The spectrum shown occurs in the first /m/ in I cannot remember it. It is the 175^{th} spectrum in the data buffer, and the value of F_o at that point is 124 Hz.

the short-time spectrum displayed above was measured.

For the purposes of this research, the speech events at which speaker-recognition parameters were measured were located manually. An effort was made to systematize the location process in order to simulate procedures that an automatic segmentation and location program would have to perform.

Individual speaker-recognition parameters were evaluated in terms of their ability to discriminate speakers and their dependence on other parameters. For the former purpose, the F-ratio of the analysis of variance was used.[3] Parameter values for the repetitions by each speaker may be regarded as samples from a probability distribution associated with that speaker. For purposes of speaker recognition, it is desirable that these individual speaker distributions be as narrow and as widely separated as possible. For the case in which the number of repetitions is the same for each speaker and equal to n, the F-ratio is given by

$$F = \frac{n(\text{variance of speaker means})}{(\text{average of speaker variances})}.$$

(The ratio F/n would be more general.) The farther apart the individual speaker distributions spread, or the narrower they become, on the average, the more suitable is the parameter and the higher is the value of F. It is not optimal in the sense of minimizing any error probability, however, and it takes no account of possible dependencies between parameters.

Interparameter dependence was roughly estimated by a technique dealing with the range overlap of pairs of individual distributions.[4] This procedure is, at present, more qualitative than quantitative, for the statistic has not been formally analyzed or derived. For the purposes of pragmatic pattern recognition, it may not be necessary to have strictly independent parameters. It may suffice to use parameters that are merely not strongly dependent, for which purpose such a rough estimation of dependence is appropriate. With the exceptions that will be described (specifically with respect to the F_o and to the nasal consonant parameters) most of the parameters investigated were not strongly dependent.

3. Specific Parameters Examined

We shall describe the acoustic parameters that were examined for potential use in speaker-recognition systems. Rather than an exhaustive inventory, it represents the results of a limited experiment in which several classes of parameters were explored. These classes are presented in approximate order of descending F-ratio.

Fundamental frequency was measured at several locations in two of the utterances. This was done to see if the increment in F_o attributable to stress was a personal

characteristic and to see if multiple F_o measurements, which would contain information on the pitch contour, would be more useful than an equal number of less dependent measurements. We found that the change in F_o caused by stress was not a clear personal characteristic, and that the dependence between F_o measurements makes it preferable to use only one or two such measurements. Fundamental frequency measurements in stressed and unstressed syllables seemed to be about equally effective, but we found that F_o measured close to a sudden articulatory and voicing change, as at the end of cannot, had an unusually high variance.

The previous report[1] illustrated the complexity of the spectra of nasal consonants and the difficulty of measuring the locations of the formants believed to be approximately related to the length of the nasal cavity. The possibility was investigated that individual filter outputs in the regions of these formants may be sensitive to formant frequencies (and hence speaker differences), even in cases for which the spectrum peaks are not clear. The filter output data were normalized for over-all amplitude and automatically recorded by a subprogram. Figure XII-3 shows the F-ratio evaluations vs frequency for individual filter outputs in the spectra of /m/ and /n/ taken from context. Broad peaks occur in this "goodness" evaluation in regions corresponding to the frequencies of the spectrum features. Specifically, these are the region of pole-zero interplay below 1 kHz and the formants around 0.25, 2, and 3 kHz in /m/ and the formants around 1, 2, and 3 kHz in /n/. Parameters measured from adjacent filters were, of course, highly correlated, but this dependence decreased as the interval between the filters increased. Consequently, parameters from the regions listed above were not strongly dependent on each other.

The frequency range of a speaker's formants, which has been found to be a correlate of voice quality,[5] is determined by the size and shape of his vocal tract. Some efforts have been made in speech recognition to normalize formant frequencies by speaker-specific factors. The experiment of Gerstman[6] in particular suggests that the extremes of F_1 and F_2, corresponding to extremes of articulation /i/, /a/, and /u/, for example, may act as reference points. Unfortunately, the extremes of the formants are the most difficult to measure accurately because of the proximity of other formants. Three techniques were brought to bear on characterizing vowels.

The shape of a multiformant spectrum peak is governed by the frequencies and bandwidths of the constituent formants. Properties of that shape, such as moments, do not require the isolation of individual formants, and they may be useful as speaker-characterizing parameters. The second central moment is related to the separation of the formants and the third central moment is related to the skewness of the peak. These moments were used with some success on the F_2-F_3-F_4 peak in /i/ and with less success on F_1-F_2 in /a/.

In the case of vowels with sufficiently widely spaced formants, formant frequencies

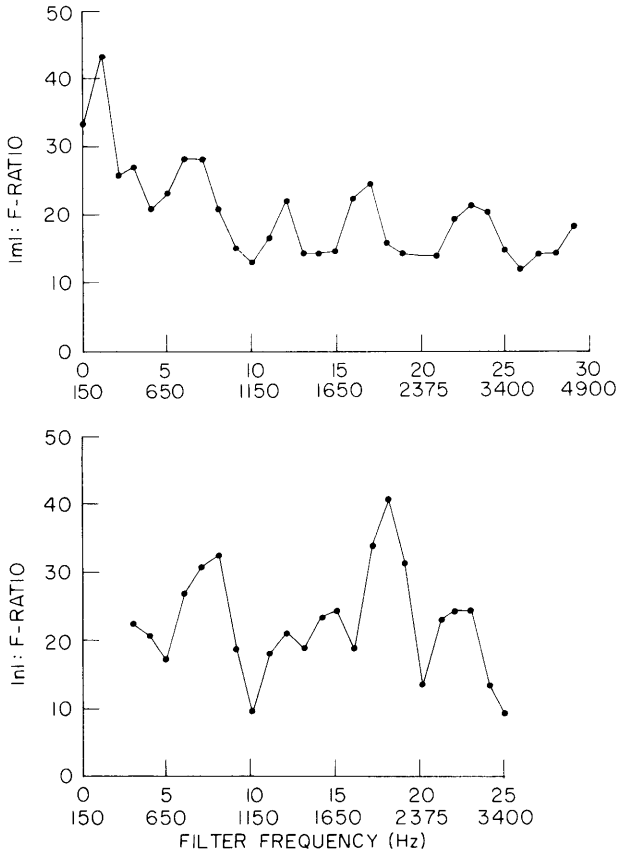

Fig. XII-3. F-ratio vs filter number (and hence frequency) for one example of /m/ and /n/.

can usually be estimated from the filter-bank representation of the spectrum. This was done for the schwa vowel, patterned after the normalization technique of Hemdal.[7]

Analysis-by-synthesis has the potential of enabling the measurement of close formants in the range of validity of the synthesis model.[8] A calculated spectrum based on hypothesized formant locations is compared with the original spectrum and the hypothesized parameters are varied until the spectra match. The parameter adjustments may be done manually, as in this case, or by algorithm.[9] In the present implementation, the hypothesized spectrum is obtained by synthesizing a short segment of vowel by means of a digital filter synthesizer and playing the synthesis output through the 36-channel filter bank. This technique was used to analyze examples of /æ/ and /a/. Only F_1 and F_2 were examined for these vowels. Analysis-by-synthesis also yields values for the formant bandwidths, but these have not yet been examined.

Only a few vowels were examined, so it is not possible to tell whether /i/, /a/, /u/ are in fact more stable than less extreme articulations. The limited data suggest that F_2 is better than F_1 for speaker characterization, but this, too, needs further confirmation.

Mártony[10] has found significant speaker differences in the laryngeal source spectrum obtained by inverse filtering. The instrumentation problem of inverse filtering was avoided by calculating an approximate source spectrum slope from a vowel spectrum. This technique yielded a surprisingly good result, but the approximations involved were so crude that this parameter may have been strongly affected by other factors.

Three other parameters were also investigated. Voice-onset time in voiced stops, and the shape of the high-frequency spectrum of the fricative /ʃ/ were described in the previous report.[1] One other parameter, the duration of a single-syllable word, was of minor consequence.

4. Recognition Procedures and Results

In order to test the usefulness and efficiency of these parameters, elementary linear classification algorithms were programmed for the PDP-9 computer. They used a weighted Euclidean distance metric similar to that used by Pruzansky and Mathews.[3] If r parameters are used, each datum is represented by a point in an r-dimensional space. The average over the repetitions of a speaker is the centroid of those points. The square of the distance between a datum $\bar{x} = (x_1, x_2, \ldots, x_r)$ and the centroid of the j^{th} speaker $\bar{\mu}_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jr})$ is given by

$$d^2(\bar{x}, \bar{\mu}_j) = \sum_{k=1}^{r} \frac{(x_k - \mu_{jk})^2}{\langle \sigma^2_{jk} \rangle_j},$$

where $\langle \sigma^2_{jk} \rangle_j$ is the average variance over all speakers for the k^{th} measurement. Dividing the squared distance in each dimension by the average speaker variance weights it according to the average narrowness of the individual speaker distributions for that parameter.

The data, which consisted in 10 repetitions by each speaker, were partitioned into design and test sets. The design set was used to form the references (speaker means and parameter variances); the test data were then used to test the effectiveness of these references in characterizing the speakers. In order to make full use of the available data, each of the 10 repetitions was used in turn as the test set, while the remaining 9 were used to form the references.

In a speaker-identification paradigm, the distance from the test datum to the centroid for each speaker is computed, and the datum is associated with the speaker whose centroid is closest. Choosing parameters in order of F-ratio, but omitting those with strong dependence on the ones already chosen, we achieved an identification error of 1.5% for 210 "utterances" by the 21 speakers with only 9 parameters. When the number of parameters was increased to 17, zero error was achieved.

In a speaker-verification paradigm, the distance between the test datum and the centroid of the claimed speaker is compared with a threshold. If the datum is closer

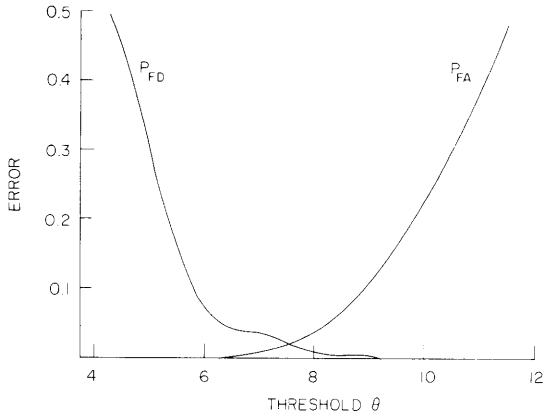

Fig. XII-4. Speaker verification errors vs comparison threshold: 210 "utterances" by 21 speakers.

than the threshold value, the speaker is verified; if farther away, he is rejected. Figure XII-4 shows the variation of the verification errors with the threshold θ. The same 17 parameters mentioned above were used. P_{FD} (false dismissal) is the chance of rejecting a true speaker, and P_{FA} (false alarm) is the chance of verifying an imposter. The curves cross at 2% error.

5. Discussion

This study was directed toward improvement of speaker-recognition techniques by means of improving the characterizing parameters extracted from the speech signal. The approach adopted here made specific measurements on speech events that had been segmented and located in the utterance. The choices of the phonetic segments and the parameters derived from them were guided by considerations of vocal-tract structure and the ways in which speech sounds are produced. The final selection of parameters was aided by evaluations of the speaker separating ability and the interdependence of the parameters. For the conditions of this experiment, the validity of this selective and efficient approach to measurement was demonstrated by the success achieved in speaker identification and verification with only a small number of measurements and a simple classification procedure.

J. J. Wolf

References

1. J. J. Wolf, "Acoustic Measurements for Speaker Recognition," Quarterly Progress Report No. 94, Research Laboratory of Electronics, M.I.T., July 15, 1969, pp. 216-222.

2. W. L. Henke, "Speech Computer Facility," Quarterly Progress Report No. 90, Research Laboratory of Electronics, M.I.T., July 15, 1968, pp. 217-219.

3. S. Pruzansky and M. V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance," J. Acoust. Soc. Am. 36, 2041-2047 (1964).

4. J. J. Wolf, "Acoustic Measurements for Speaker Recognition," Ph.D. Thesis, Department of Electrical Engineering, M.I.T., September 1969.

5. J. E. Miller, "Decapitation and Recapitation, a Study in Voice Quality," J. Acoust. Soc. Am. 36, 2002 (Abstract) (1964).

6. L. J. Gerstman, "Classification of Self-Normalized Vowels," IEEE Trans., Vol. AU-16, pp. 73-77, 1968.

7. J. Hemdal, "Some Results from the Normalization of Speaker Differences in a Mechanical Vowel Recognizer," J. Acoust. Soc. Am. 41, 1594 (Abstract) (1967).

8. C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Am. 33, 1725-1736 (1961).

9. A. P. Paul, "An Automatic Spectrum Matching Program," Quarterly Progress Report No. 66, Research Laboratory of Electronics, M.I.T., July 15, 1962, pp. 275-278.

10. J. Mártony, "Studies of the Voice Source," Quarterly Progress and Status Report 1/65, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, 1965, pp. 4-9.