Received 30 March 1970; revised 18 March 1971

Efficient Acoustic Parameters for Speaker Recognition*

JARED J. WOLF

Bolt Beranek and Newman, Incorporated, Cambridge, Massachusetts 02138

In a scheme for the mechanical recognition of speakers, it is desirable to use acoustic parameters that are closely related to voice characteristics that distinguish speakers. This paper describes an investigation of an efficient approach to selecting such parameters, which are movitated by known relations between the voice signal and vocal-tract shapes and gestures. Rather than general measurements over the extent of an utterance, only significant features of selected segments are used. A simulation of a speaker recognition system was performed by manually locating speech events within utterances and using parameters measured at these locations to classify the speakers. Useful parameters were found in fundamental frequency, features of vowel and nasal consonant spectra, estimation of glottal source spectrum slope, word duration, and voice onset time. These parameters were tested in speaker recognition paradigms using simple linear classification procedures. When only 17 such parameters were used, no errors were made in speaker identification from a set of 21 adult male speakers. Under the same conditions, speaker verification errors of the order of 2% were also obtained.

INTRODUCTION

The problem of speaker recognition, like most problems in pattern recognition, may be considered to be divided into two parts: measurement and classification. In the first part, a number of parameters are abstracted from the pattern under test. These parameters (ideally) characterize the pattern. The resulting set of numbers in turn acts as the input to a classification scheme, which compares them with stored information on known reference patterns and makes a decision as to the class membership of the tested pattern.

Early work in speaker recognition confirmed that speaker-dependent information could be readily extracted from a prespecified speech signal by making regularly sampled spectrum or formant measurements.[1–5] Because such parameters are only generally related to speaker differences, much of their speaker-characteristic information content is likely to be encoded in a complex form not ideally suited to the subsequent classification process. Later speaker recognition work has shown a trend toward specific parameters which are more directly related to differences between speakers.[6–9] The benefits of efficient characterizing parameters are obvious. They may enable the storage of less data, the use of faster and less complex classification procedures, the achievement of lower error, or a fortuitous combination of these.

This paper reports an investigation of speaker recognition procedures in which the primary purpose was to attempt to use acoustic and phonological theory to find acoustic parameters which are both efficient in discriminating speakers and amenable to automatic implementation. A simulation of a speaker recognition system was performed by manually locating speech events within utterances and extracting parameters at these places. These parameters were evaluated and subsequently tested in two speaker recognition paradigms by means of elementary classification procedures.

I. GENERAL MEASUREMENT PRINCIPLES

The function of the measurement phase of a speaker recognition system is to perform a number of characterizing measurements on the voice pattern under test. Ideally, the speech characteristics measured should:

- occur naturally and frequently in normal speech,
- be easily measurable,
- vary as much as possible among speakers, but be as consistent as possible for each speaker,
- not change over time or be affected by the speaker's health,
- not be affected by reasonable background noise nor depend on specific transmission characteristics, and

- not be modifiable by conscious effort of the speaker, or, at least, be unlikely to be affected by attempts to disguise the voice.

In practice, the simultaneous fulfillment of all these criteria is probably beyond the present state of the art. Partial or complete relaxation of some of these standards is reasonable for some research purposes and for limited practical speaker recognition. Specifically, in the research described here, the last three factors were not investigated, but were controlled.

Differences in voices stem from two broad bases: organic and learned differences.[10] Organic differences are the result of variations in the sizes and shapes of the components of the vocal tract: larynx, pharynx, tongue, teeth, and the oral and nasal cavities. Since the resonances of the vocal tract and the characteristics of the sound energy sources depend upon just these anatomical factors, organic differences lead to differences in fundamental frequency, laryngeal source spectrum, and formant frequencies and bandwidths. Learned differences are the result of differences in the patterns of coordinated neural commands to the separate articulators learned by each individual. Such differences give rise to variations in the dynamics of the vocal tract such as the rate of formant transitions and coarticulation effects. Naturally, many speaker-dependent characteristics are affected by both of these factors.

One class of measurement schemes that has been used in the past performs a set of measurements at 10- to 20-msec intervals throughout an entire utterance.[1–5,11] There are three outstanding difficulties with this approach. First, regular and rapid sampling of the voice signal with the characterizing measurements produces large sets of data that have a high degree of redundancy. Secondly, a given set of parameters is not optimally suited to every segment of an utterance. Some of them will be useless during many speech sounds, as in the case of fundamental frequency during voiceless intervals. Finally, because of the normal detailed differences in timing of each utterance, corresponding articulatory events do not occur at exactly the same times, even if the utterances are registered at a particular point, such as the beginning or the energy peak. Therefore, comparisons of parameter values at points where the utterances are out of alignment are comparisons between somewhat different events. It may be argued that these misalignments are reflections of temporal patterns associated with learned characteristics of different speakers. This is indeed so, but in this form the temporal variations interfere with the comparisons of similar articulatory events. It would be useful to separate these effects.

Another class of measurement schemes uses characteristics averaged over time, such as long-term spectra.[2–4,12,13] Such parameters may be measured over specific contexts or over long enough intervals to be context independent. This approach leads to much smaller data sets, and it virtually eliminates the effects of timing differences, but it excludes a large class of probably useful speaker-dependent effects. These basically short-term effects include all phoneme-specific effects and virtually all learned characteristics.

A third and possibly still more efficient approach to measurement is to perform some degree of segmentation and recognition of the linguistic component of the speech signal. In most speaker recognition applications, it is reasonable to assume that a known phrase or sentence is used for the speech sample. In this case, the necessary segmentation and speech recognition would not be difficult. Ideally, the system designer would be free to specify the utterance, and he could tailor it both to contain an advantageous set of phonemes and to be easily segmented.

The ability to find its way about the utterance allows the system to locate certain speech events of interest and then to extract appropriate parameters at each of these points. Similar events can then be compared with a minimum of interference due to timing differences. Furthermore, the recognition of events and boundaries in the acoustic signal allows the separate measurement of relevant temporal patterns. This last approach is the one adopted in this investigation.

II. DATA BASE

In order to investigate specific speaker-characterizing parameters, examples of speech from 21 adult male American speakers were recorded. The speakers ranged in age from 22 to 42 years. None had a noticeable speech defect. Regional accent was not closely controlled; two speakers had mild southern accents. All speakers were staff or students at the Massachusetts Institute of Technology. They were told of the nature of the experiment and were asked to speak normally. All speakers were reasonably free from colds or other respiratory inflammations.

Six short sentences were devised to provide a generally wide variety of speech segments. These sentences are given below.

(1) Cool shirts please me.
(2) Pay the man first, please.
(3) I cannot remember it.
(4) Papa needs two singers.
(5) A few boys bought them.
(6) Cash this bond, please.

Ten repetitions of each of the six sentences were recorded in mixed order from each speaker. The data were recorded in a single session under low noise conditions with high-quality wide-bandwidth equipment.

III. EQUIPMENT AND GENERAL PROCEDURE

The speech analysis and pattern recognition were performed with the aid of a digital computer laboratory

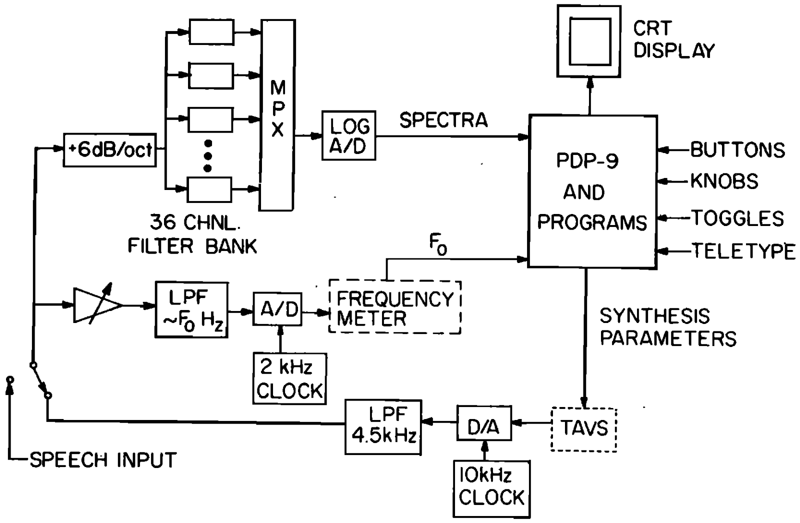

FIG. 1. Block diagram of the computer facility used for on-line speech analysis. Functions enclosed in broken lines are performed by subprograms.

facility built around a Digital Equipment Corporation PDP-9 computer. The computer is coupled to peripheral equipment designed to facilitate on-line speech research. This highly flexible arrangement makes it possible, through convenient interconnections and proper programming, to create in effect a special-purpose on-line laboratory instrument.

A set of programs was written to enable this facility to function as a versatile spectrum analysis and display device. Figure 1 illustrates the features of this system. The principal speech-analysis tool is a 36-channel filter bank spectrum analyzer. The filters use a single complex pole pair; their center frequencies are spaced linearly between 150 and 1650 Hz and logarithmically thereafter to 7025 Hz; the skirts of adjacent filters cross at their −3-dB points. Each filter output is rectified and smoothed with a 10-msec time-constant filter. A multiplexer and logarithmic analog-to-digital converter perform a scan of the 36 channels in 1.3 msec, which is small compared to the averaging time of the smoothing filters.

Since the first harmonic was present in the recorded speech data and the signal-to-noise ratio was high, fundamental frequency could be measured by the rudimentary scheme shown in Fig. 1. The speech was low-pass filtered (cutoff frequency adjusted for each speaker to be slightly above his maximum fundamental frequency), and F_0 was estimated by a subprogram that measured the intervals between zero crossings. This scheme was quite satisfactory as a research tool, but any prospective automatic speaker recognition system would certainly use one of the more effective pitch extraction techniques that have appeared in the literature.[14,15]

A variation on the manual analysis-by-synthesis procedure described by Bell *et al.*[16] was implemented for vowels on the PDP-9. In this version, a natural vowel spectrum is analyzed by generating a 30-msec synthetic vowel and adjusting the synthesis parameters to minimize the difference between the natural and synthetic spectra, as produced by the filter bank spectrum analyzer. The synthesis is performed by a five-formant cascade 10-kHz sampled data speech synthesis program developed by W. L. Henke.[17] Not only is this procedure faster than the original implementation, but it is also more accurate, since the synthesized spectrum is derived using the measured value of F_0 instead of an assumed value, and there is no error in calculating the filter bank response. The principal limitation on the accuracy of the present system is the quality of the glottal source, which is approximated by a single complex pole-pair.

The results of most of the programs are displayed on a 16-in. cathode-ray-tube display. An arrangement of pushbuttons, knobs, and toggle switches provides a convenient interface for the user to control the operation of the program.

The speech data were kept in analog form on tape. When an utterance was read in, the short-time spectrum and fundamental frequency were sampled every 10 msec and stored in the core memory. A 2.5-sec utterance could be stored at once. A typical display generated by the program is shown in Fig. 2. The two graphs in the lower half represent functions of time, from 0 to 2.5 sec. The upper one, called the "energy function," is the average of the outputs of several low-frequency filters, useful as a "low-frequency energy map" of the utterance. The lower graph is fundamental frequency. The vertical cursor, which is controlled by a knob, denotes a point in the utterance. The short-time spectrum corresponding to that point is displayed in the upper part of the screen.

For the purpose 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.

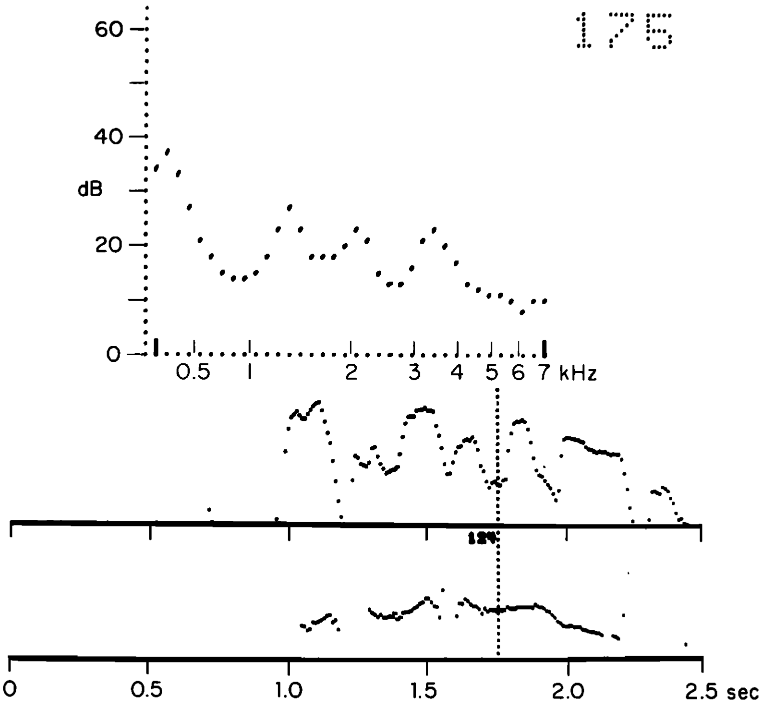

FIG. 2. Photograph of CRT display. The two graphs in the lower half represent low-frequency energy and fundamental frequency as functions of time. The vertical cursor shows the point in the utterance corresponding to the spectrum displayed above. The points on the vertical axis of the spectrum represent 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 175th spectrum in the data buffer, and the value of F_0 at that point is 124 Hz.)

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.[3,9] For a given parameter, the values obtained from the repetitions by each speaker may be regarded as samples from a probability distribution associated with that speaker. For speaker recognition, a good parameter is one for which these individual speaker distributions are as narrow and as widely separated as possible. The F ratio is given by

$$F=\frac{n}{m-1}\sum_{j=1}^{m}(\mu_j-\bar{\mu})^2 \Big/ \frac{1}{m(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\mu_j)^2, \quad (1)$$

where x_{ij} is the parameter value on the ith repetition by the jth speaker, $i=1, \cdots, n$, $j=1, \cdots, m$;

$$\mu_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij} \quad (2)$$

is the estimated mean for the jth speaker; and

$$\bar{\mu}=\frac{1}{m}\sum_{j=1}^{m}\mu_j \quad (3)$$

is the estimated over-all mean. Thus F is proportional to the ratio of the variance of the speaker means to the mean of the speaker variances. The farther apart the individual speaker distributions are or the narrower they are, on the average, the more suitable is the parameter and the higher is the value of F. However, it is not optimal in the sense of minimizing any error probability, and it takes no account of possible dependencies between parameters.

The estimation of interparameter dependence is less straightforward. The correlation coefficient is not really appropriate for this, since it is far better suited for confirming a linear relationship than for testing for the absence of any relationship. Furthermore, except for special cases, the absence of correlation implies nothing about independence. Pairwise interparameter dependence was roughly estimated by a statistic called ΔP which deals with the range overlap of the parameter samples from individual speakers. This procedure is briefly explained in Appendix A.

IV. PARAMETERS EXAMINED

This section describes the acoustic parameters that were examined for potential use in speaker recognition systems. In the course of the investigation, each parameter was given a mnemonic symbolic name, which was also used to identify that set of data in the file system of the computer's tape and disk storage. Further details of the parameters are given in Appendix B. F-ratio evaluations of the parameters are given in Table I.

A. Fundamental Frequency

Fundamental frequency was recorded at several locations in each of two of the utterances. Such measurements are undoubtedly somewhat correlated, but in addition to average value, they also contain informa-

TABLE I. Parameters ranked by F ratio.

Name	F-ratio
5F02	84.9
5F01	81.0
AEF0	72.2
3F02	71.2
5F04	69.5
3F01	61.8
5F03	54.3
3F05	52.8
3F04	51.8
AEF2	46.6
UHF2	44.6
3M1	43.4
3N18	41.0
SSS	36.3
IU3	34.4
IS2	32.7
3N8	32.5
3F03	30.9
3M6	28.4
3M17	24.8
3N23	24.4
AF1	22.9
3M23	21.7
UHF1	21.1
BAWT	20.7
AF2	19.0
SH	17.5
AEF1	15.5
PREV	14.5
AS2	11.8
AU3	10.2

tion about the pitch contour, which has been used for speaker recognition by Atal.[11] We wished to find out if such data would be useful in the context of a small number of efficient parameters (or if it would be more efficient to measure fundamental frequency only once and to use other, less dependent, parameters). We also wished to find out if the increment in fundamental frequency due to stress is useful as a speaker-specific characteristic.

Increments in F_0 due to stress in *cannot* and *remember* were estimated by subtracting 3F03 from 3F02 and 3F06 from 3F04 (i.e., the values at the F_0 peak and in the preceding unstressed syllable). Inspection of these results showed rather large intraspeaker variations (comparable to the total range in many cases), so the possibility of using these increments in F_0 was abandoned without further analysis.

With the exception of 3F03, the individual fundamental frequency parameters had the highest F ratio of all the parameters investigated. The exception occurred at the F_0 peak at the end of the word *cannot.* Its especially high variability was attributed to the proximity of a sudden voicing and articulation change.

B. Nasal Consonants

As pointed out by Glenn and Kleiner,[7] the articulatory configuration of the nasal consonants makes them particularly appropriate for speaker recognition measurements. A large portion of the acoustic system is fixed and not subject to articulatory intraspeaker variation. The analyses and experiments of Fant[18] and Fujimura[19] suggest that certain poles of the transfer functions of the nasal consonants are closely tied to the nasal cavity alone. However, the measurement of the frequencies of spectral peaks corresponding to these poles is complicated by the presence of zeros in the transfer function and the generally higher formant damping of the nasal consonants. Figure 3 shows spectra of four examples of [m] by each of six speakers. The speaker represented by the top row (and in Fig. 2) has a rather classic [m]-spectrum. Clear peaks are evident around 0.25, 1.3, 2, and 3 kHz, and a pole around 0.9 kHz is cancelled by a zero. The spectra in the lower five rows show how some of the peaks can be obscured for some speakers. With the exception of the first formant, it is impossible to locate and consistently identify any of these spectral features.

The possibility was investigated that individual filter outputs in the regions of the formants may be sensitive to formant frequencies even in cases where the spectrum peaks are not clear. Such measurements were performed in one example of [m] and [n] in "I cannot remember it." Since such data are subject

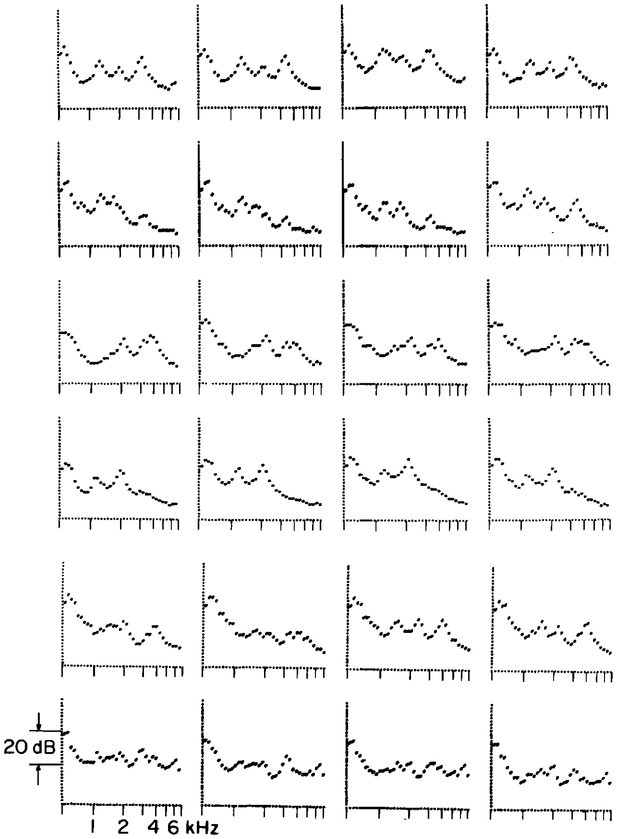

FIG. 3. Spectra of [m]. Each row contains four examples by one speaker, and different speakers are represented by different rows.

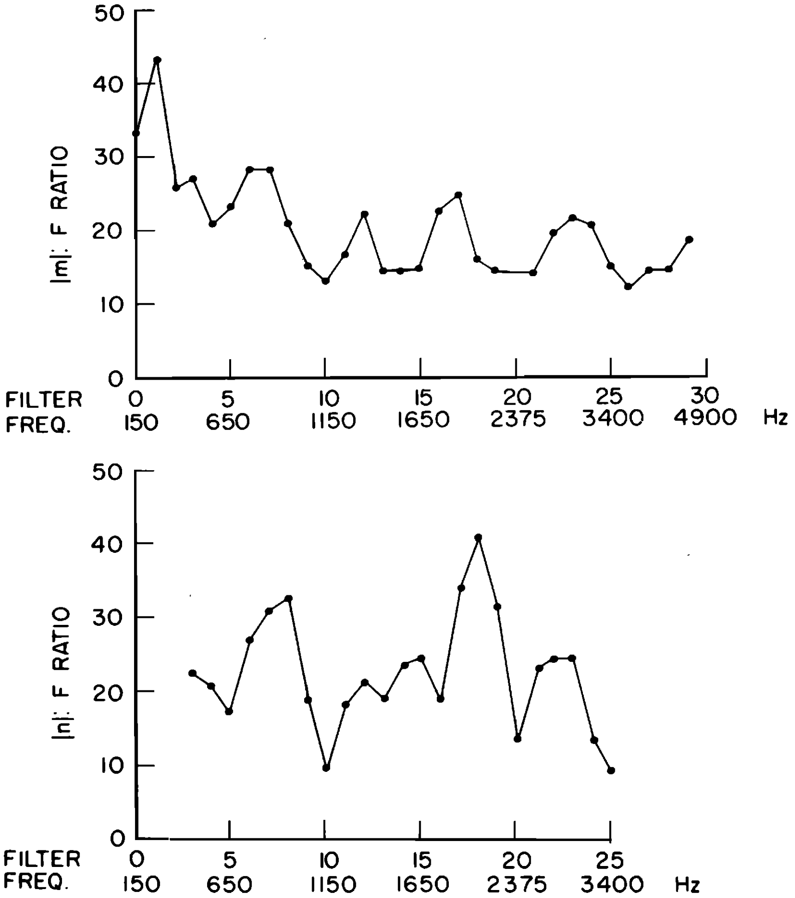

FIG. 4. *F* ratio versus filter number (and hence frequency) for one example of [m] and [n], parameters 3Mi and 3Ni.

to variation due to differences of voice level, they must be suitably normalized. This was done by subtracting from each filter output the value of the energy function measured in the middle of the following vowel (both values in decibels). (An energy estimation from the same nasal spectrum would probably also have been effective.) Figure 4 shows the *F* ratio versus filter center frequency for these parameters. Broad peaks occur in this "goodness" evaluation in regions that correspond 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]. This suggests that amplitude-normalized filter outputs in these regions of nasal spectra may be useful as speaker-characterizing parameters. Such an approach is certainly less satisfying than directly measuring formant frequencies theoretically associated with the structure of the speaker's nasal cavity, but in the absence of the ability to do so, it at least takes cognizance of the locations of the poles and zeros underlying nasal spectra.

C. Vowels

The frequency range of a speaker's formants, which has been found to be a correlate of voice quality,[20] is determined by the size and shape of his vocal tract. It has also been found that the identification of a vowel can be strongly influenced by changing the formant range of the surrounding context.[21] This suggests that the listener effectively applies some kind of normalization for each speaker to the important formants.

Some efforts have been made in speech recognition to normalize formant frequencies by speaker-specific factors. The experiment of Gerstman[22] in particular suggests that the extreme values of F_1 and F_2, corresponding to extremes of articulation [i], [a], and [u], for example, may act as reference points. It has been found that the first two formants of these vowels are the least sensitive to context,[23] and their stability is supported by Stevens's theory of the quantal nature of certain vowel articulations.[24] Unfortunately, the formants at their extremes are the most difficult to measure accurately because of their proximity to other formants. Three techniques were brought to bear on the 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 and identification of individual formants, and they may be useful as

TABLE II. ΔP for pairs of parameters. (Decimal point omitted. Read −03 as −0.03).

	3F01	3F05	5F02	5F03	5F04	3M1	3M6	3N8	3N23	IS2	IU3	UHF1	UHF2	AEF1	AEF2	AF1	AF2	SSS	PREV	BAWT
3F01		52	36	41	58	−08	−04	03	−07	−14	00	−06	−14	05	−01	−04	−01	−01	−01	−07
3F05	52		85	41	92	−05	02	16	01	−17	06	04	−05	05	29	01	01	−04	01	−07
5F02	36	85		15	93	−06	−03	07	00	−02	01	−03	−16	−03	21	−02	00	−07	−02	−07
5F03	41	41	15		24	−15	00	02	04	−04	04	−03	11	01	06	−06	−02	−00	01	−12
5F04	58	92	93	24		−02	05	09	−08	−05	03	07	−11	01	12	06	−03	05	−03	−01
3M1	−08	−05	−06	−15	−02		31	01	04	06	06	04	−03	02	03	02	−01	−05	−02	04
3M6	−04	02	−03	00	05	31		13	02	−04	01	04	−01	02	−06	−08	05	−01	−02	−01
3N8	03	16	07	02	09	01	13		13	−14	−02	−04	−01	11	−03	05	03	02	−00	−02
3N23	−07	01	00	04	−08	04	02	13		−01	−02	−00	04	01	11	07	−02	−02	−02	−02
IS2	−14	−17	−02	−04	−05	06	−04	−14	−01		09	14	−04	−10	−01	21	−02	−07	00	−03
IU3	00	06	01	04	03	06	01	−02	−02	09		07	−05	−00	04	−01	−04	−04	−01	05
UHF1	−06	04	−03	−03	07	04	04	−04	−00	14	07		−01	07	03	00	−06	−03	−03	03
UHF2	−14	−05	−16	11	−11	−03	−01	−01	04	−04	−05	−01		−04	23	−05	−03	−01	05	06
AEF1	05	05	−03	01	01	02	02	11	01	−10	−00	07	−04		−05	09	−06	05	−01	01
AEF2	−01	29	21	06	12	03	−06	−03	11	−01	04	03	23	−05		03	02	−07	02	−03
AF1	−04	01	−02	−06	06	02	−08	05	−07	21	−01	00	−05	09	03		−02	07	07	−01
AF2	−01	01	00	−02	−03	−01	05	03	−02	−02	−04	−06	−03	−06	02	−02		−03	02	−04
SSS	−01	−04	−07	−00	05	−05	−01	02	−02	−07	−04	−03	−01	05	−07	07	−03		−03	−00
PREV	−01	01	−02	01	−03	−02	−02	−00	−02	00	−01	−03	05	−01	02	07	02	−03		03
BAWT	−07	−07	−07	−12	−01	04	−01	−02	−02	−03	05	03	06	01	−03	−01	−04	−00	03	

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 on the $F_2-F_3-F_4$ peak in [i] (parameters IS2=2nd moment, IU3=3rd moment) and on the F_1-F_2 peak in [a] (parameters AS2 and AU3).

In the case of vowels with sufficiently widely spaced formants, formant frequencies can usually be estimated from the filter bank representation of the spectrum. This was done for the schwa vowel, patterned after the normalization techniques of Hemdal.[25] Only the first and second formants were estimated (parameters UHF1 and UHF2). The third and fourth formants were also attempted, but their peaks were frequently absent or weak in the filter bank spectrum.

The method of analysis-by-synthesis has the potential of measuring the frequencies of close formants in the range of validity of the synthesis model. A spectrum based on hypothesized formant locations is compared with the original spectrum and the synthesis parameters are varied until the two spectra match. The parameter adjustments may be done manually, as in this case, or by a strategy algorithm.[26] This technique was used to analyze examples of [æ] and [a]; only F_1 and F_2 in these vowels were examined (parameters AEF1, AEF2, AF1, AF2). Analysis-by-synthesis also yields values for formant bandwidths, but these were not felt to be as reliable as the formant frequencies, so they have not yet been examined.

D. Other Parameters

The structure of the larynx affects not only the laryngeal pulse repetition rate, but also the pulse shape, which is reflected in the laryngeal source spectrum. Unfortunately, this spectrum is not directly accessible to measurement in the speech signal, since it is modified by the transfer function of the upper vocal tract. Mártony, using inverse filtering, has found significant differences among speakers in the high-frequency slope of the source spectrum.[27]

The instrumentation problem of inverse filtering was avoided by roughly estimating the source spectrum slope from a vowel spectrum. The parameter SSS is the difference (in decibels) between the amplitudes of the spectrum peaks corresponding to F_1 and F_3 in an example of [u], normalized by their frequency separation on a logarithmic scale. This yielded a surprisingly good result, but the approximations involved here are so crude that this parameter may be heavily affected by other factors.

The spectrum of fricatives like [ʃ] or [s] depends mainly on the anatomical details of the region around and forward of the alveolar ridge. Hence parameters derived from these sounds are not influenced by the entire vocal tract, but only by a small portion of it. The fricative [ʃ] was chosen over [s] because its principal features occur farther below the 7-kHz upper limit of the spectrum analyzer. It was found that the high-frequency portion of the spectrum was difficult to characterize in terms of the frequency of a peak, but its shape seemed to lend itself more readily to the problem. In this case, moments of the spectrum in the region 2.5–7.0 kHz had rather low F ratios, and an algorithmic classification of the high-frequency spectrum into four classes, *single narrow peak*, *wide or double peak*, *flat*, and *very low-frequency major peak* proved more useful. The parameter SH was given one of the arbitrary values 1, 2, 3, or 4, corresponding to its shape class.

The preceding parameters have emphasized the structural aspects of individual voice characteristics. Learned voice characteristics probably deal mainly with timing; in this case it is somewhat harder to know what to measure. As a single example of a gross timing parameter (BAWT), the duration of the word "bought" was examined. The definition of the duration was simplified because the word began and ended with stops. The 10-msec quantization intervals proved to

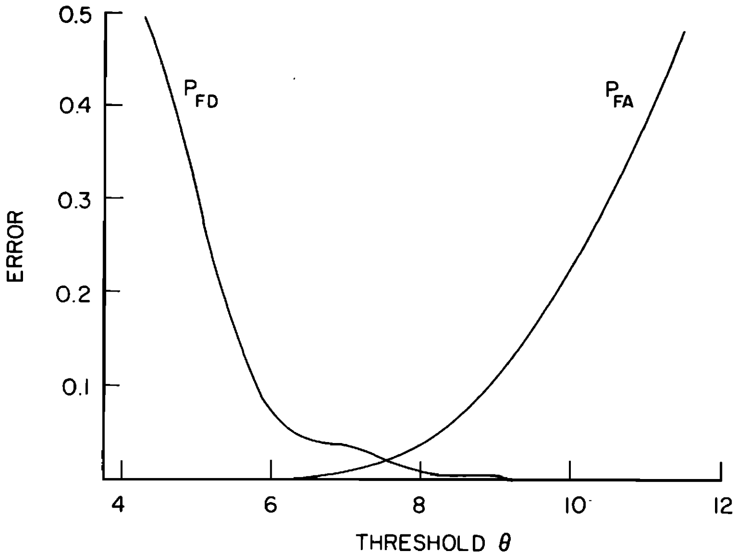

FIG. 5. Average speaker verification errors versus comparison threshold: 210 "utterances" by 21 speakers, all possible utterance-speaker comparisons.

be rather coarse for an event of less than 200-msec duration, but the parameter did prove to be useful.

In voiced stops following an unvoiced segment, the onset of voicing before the release of the stop is not used for phonemic distinction in English, but it is not uncommon, and its speaker-specific tendency has been noted.[28] This phenomenon was observed in the context "this bond." A binary distinction (prevoiced or not) was used; the stop was termed prevoiced if voicing preceded the burst by 20 msec or more. Six of the 21 speakers prevoiced more than half the time, four occasionally did, and 11 never did. Although this parameter (PREV) can yield only limited information, it is appealing because it concerns a rapid event which is not likely to be consciously modifiable, and it is an event of such specificity that it is probably independent of most other parameters.

E. Comparison of Parameters

The F ratios of the parameters described above are given in Table I. In the case of nasal consonant parameters, only those which correspond to F-ratio peaks in Fig. 4 are included. The significance of these values may be roughly demonstrated by the following examples. If all the speakers had distributions with equal variances of σ, and if half of them had means of $-\sigma$ and half had means of $+\sigma$, the resulting value of F would be 10; if they fell into four groups with adjacent means separated by 2σ, the F ratio would be 50.

With the one exception noted earlier, the fundamental frequency parameters lead all others in terms of F ratio. Given the examples at hand, there seems to be no particular advantage to measuring F_0 in stressed or unstressed syllables. Below the F_0 parameters appears a rather mixed bag of all the rest. The number of vowels investigated was so small that no conclusions can be drawn as to which vowels or vowel features are best suited to speaker recognition. The estimation of source spectrum slope proved to be quite useful, as did the better nasal consonant parameters. The small number of categories in the [ʃ]-shape and prevoicing parameters clearly limits their F-ratio values.

Values of the interparameter dependence estimator ΔP for some parameters are given in Table II. As expected, the several F_0 parameters are generally quite dependent. Nasal parameters from adjacent filters in the same consonant are also quite dependent, but this decreases as the two filters are chosen farther apart. This means that the several nasal parameters shown in Table I are generally not shown to be highly dependent by ΔP. The values of ΔP for the parameters omitted from Table II show no departure from the pattern shown in the table: generally high values for comparisons of F_0 parameters and low values otherwise.

V. RECOGNITION PROCEDURES AND RESULTS

In order to test the usefulness and efficiency of these parameters, simple linear classification algorithms were programmed for the PDP-9 computer. These procedures 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 of the repetitions of a speaker is the centroid of those points. The square of the distance between a datum $\mathbf{x} = (x_1, x_2, \cdots, x_r)$ and the centroid of the jth speaker $\boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2}, \cdots, \mu_{jr})$ is given by:

$$d^2(\mathbf{x}, \boldsymbol{\mu}_j) = \sum_{k=1}^{r} \frac{(x_k - \mu_{jk})^2}{v_k}, \qquad (4)$$

where

$$v_k = \frac{1}{m} \sum_{j=1}^{m} \sigma_{jk}^2 \qquad (5)$$

is the average estimated speaker variance for the kth parameter. Dividing each distance component by the average speaker variance weights it according to the average narrowness of the individual speaker distributions for that parameter.

The parameter data, which consisted of ten 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 was then used to test the effectiveness of these references in characterizing the speakers. In order to make full use of the available data, a "leave one out" procedure was used in the recognition trials. Each of the ten repetitions was used in turn as the test set, while the remaining nine formed the design set.

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. Parameters were selected from Table I in order of F ratio, but those whose ΔP comparisons with parameters already chosen exceeded 0.25 (an arbitrary threshold) were omitted. With only nine parameters, an average identification error of 1.5% was achieved for 210 "utterances" by the 21 speakers. When the number of parameters was increased to 17, zero identification error was achieved for this set of speakers.

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 than the threshold value, the speaker is verified; if farther away, he is rejected. This comparison was performed for each "utterance" and every speaker in the set of 21. Figure 5 shows the average verification errors as a function of the distance threshold. The same 17 parameters mentioned above were used. P_{FD} (false dismissal) is the chance of rejecting a true speaker, and P_{FA} (false acceptance) is the chance of verifying an imposter. The curves cross at 2%.

VI. DISCUSSION

This investigation has been directed toward the improvement of speaker recognition techniques by means of finding efficient characterizing parameters to be extracted from the voice signal. The approach adopted here uses several different types of parameters taken from specific speech events which have been segmented and located in the utterance. The choices of the speech events and the parameters derived from them are guided by considerations of vocal tract structure and the ways in which the various speech sounds are produced.

The usefulness of this approach is shown by the good results obtained in identification and verification paradigms with only a small number of such parameters and elementary classification procedures. It should be noted that the output of the measurement phase is a set of numbers, which may be applied to any of the numerous classification procedures that have been described in pattern recognition literature, including those which abstract optimal "features" from combinations of the parameters provided as inputs. In this investigation, classification techniques were of secondary interest, and the elementary one described gave good results.

The set of parameters described here cannot be called optimum, since only a relatively small number of possible parameters were investigated. The use of vowels was hindered by the lack of a rapid and reliable formant analysis technique, and the characterization of nasal consonants and of source spectrum slope can undoubtedly be improved. Other parameter types, notably those reflecting learned characteristics, also bear investigation.

The usefulness of the specific parameters reported here should be interpreted in light of the controlled conditions under which the speech data was obtained. For example, fundamental frequency is very susceptible to stress on the speaker,[29] and it is perhaps the easiest and most obvious acoustic correlate to modify for the purpose of voice disguise. Measurements on nasal consonants may be particularly sensitive to the state of health of the speaker. Furthermore, the speech samples were recorded at a single session, so the extent of parameter variation over time is not represented in the data.

ACKNOWLEDGMENTS

The author wishes to thank Professor Kenneth N. Stevens for his advice and criticism during the course of this study. This work was supported in part by a fellowship from the National Science Foundation, and in part by the U. S. Air Force (Office of Aerospace Research) under Contract F19628-69-C-0044. The preparation of this report was further supported by a visiting fellowship to the Department of Psychology, University of Edinburgh, given by the Leverhulme Trust.

APPENDIX A: A MEASURE OF INTERPARAMETER DEPENDENCE

This technique is intended to provide a rough estimate of pairwise interparameter dependence. It is based on a method of estimating the extent of overlap of the parameter distributions of individual speakers.

Consider measurements of two parameters for n examples by each of m speakers. These values may be represented by points in a two-dimensional space, as

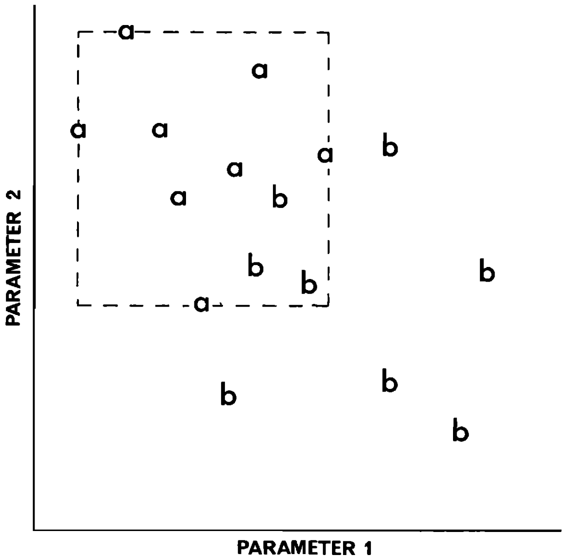

FIG. A-1. Hypothetical two-parameter data points for two speakers, A and B. The box delineates the range of the samples of speaker A in both dimensions.

illustrated in Fig. A-1 for the case $n=8$, $m=2$. Perhaps the most elementary measure of the overlap of the individual speaker clusters is the extent to which their ranges overlap each other, or the extent to which data points from one speaker fall inside the ranges of the other speakers in the set. A datum point of one speaker is termed discriminable from another speaker if it falls outside the range of the data points of that other speaker. Thus, in Fig. A-1, the five speaker B samples outside the box are discriminable from speaker A; the remaining three are not. If this comparison is performed for each of the mn data points against the $(m-1)$ other speakers, an average measure of speaker discriminability is given by

$$\hat{P}_d = d/[mn(m-1)], \tag{A1}$$

where d is the total number of such datum-speaker discriminations that were found. This quantity is a relative frequency; it may be regarded as an estimate of P_d, the average probability of discrimination of the samples of one speaker from all the other speakers in the set, according to the criterion stated above. Similarly, $(1-\hat{P}_d)$, the proportion of indiscriminable comparisons, may be regarded as an estimate of the probability of confusion.

The quantity P_d may also be estimated for each parameter (or dimension) separately. Let P_1, P_2, and P_{12} represent the values of P_d for parameter 1 alone, parameter 2 alone, and both parameters jointly. Since a datum which is discriminable in one dimension alone must be discriminable in both dimensions jointly, then for independent parameters:

$$1-P_{12} = (1-P_1)(1-P_2). \tag{A2}$$

We may expect that the estimates approximate this relationship, or that the quantity

$$\Delta P = \frac{1-\hat{P}_{12}}{(1-\hat{P}_1)(1-\hat{P}_2)} - 1 \tag{A3}$$

will be close to zero for the case of independent parameters.

Unfortunately, the distribution of this statistic under the hypothesis of independent parameters is not known, so it is not possible to assign a critical region and significance level to a test using ΔP. For the purposes of practical pattern recognition, however, it is not necessary to have strictly independent parameters, and an ad hoc procedure of comparing ΔP values to some (possibly experimentally determined) threshold probably tends to avoid strong dependencies. Likewise, pairwise independence does not guarantee mutual independence among more than two parameters, but in practical situations, it probably helps.

APPENDIX B: PARAMETER NAMES AND LOCATIONS

Figures B-1 and B-2 contain spectrograms of one example of each of the six sentences in the data set. The locations of the parameters given below refer to the points marked and labeled on the spectrograms.

A. Fundamental Frequency

These parameters are simply the values of F_0 at the locations specified. The first six occur in sentence 3, an example of which is also shown in Fig. 2.

3F01: At the middle of "I" (point 3a). This syllable is clearly shown by the energy function and by the absence of voicing on either side.

3F02: At the middle of the first vowel in "cannot" (point 3b). The energy function drops sharply at the beginning of the [n].

3F03: At the peak of F_0 in the syllable "not" (point 3d). The word "cannot" was stressed on the second syllable.

3F04: At the middle of the first vowel (peak of energy function) in "remember" (point 3e).

3F05: At the middle of the second vowel in "remember" (midway between the energy function drops caused by the [m]'s on either side, point 3g).

3F06: At the peak in F_0 in "remember" after 3F05 (point 3h). If there was no such peak, 3F06 was set equal to 3F05; this occurred often, so this parameter was subsequently discarded.

5F01: At the middle of "few" in sentence 5 (point 5a).

5F02: At the peak of F_0 in "few" (point 5b). This was often the same as 5F01.

5F03: At the middle of the diphthong in "boys" (point 5c).

5F04: At the middle of "bought" (point 5e).

AEF0: At the middle of the voiced portion of "cash" in sentence 6 (point 6a).

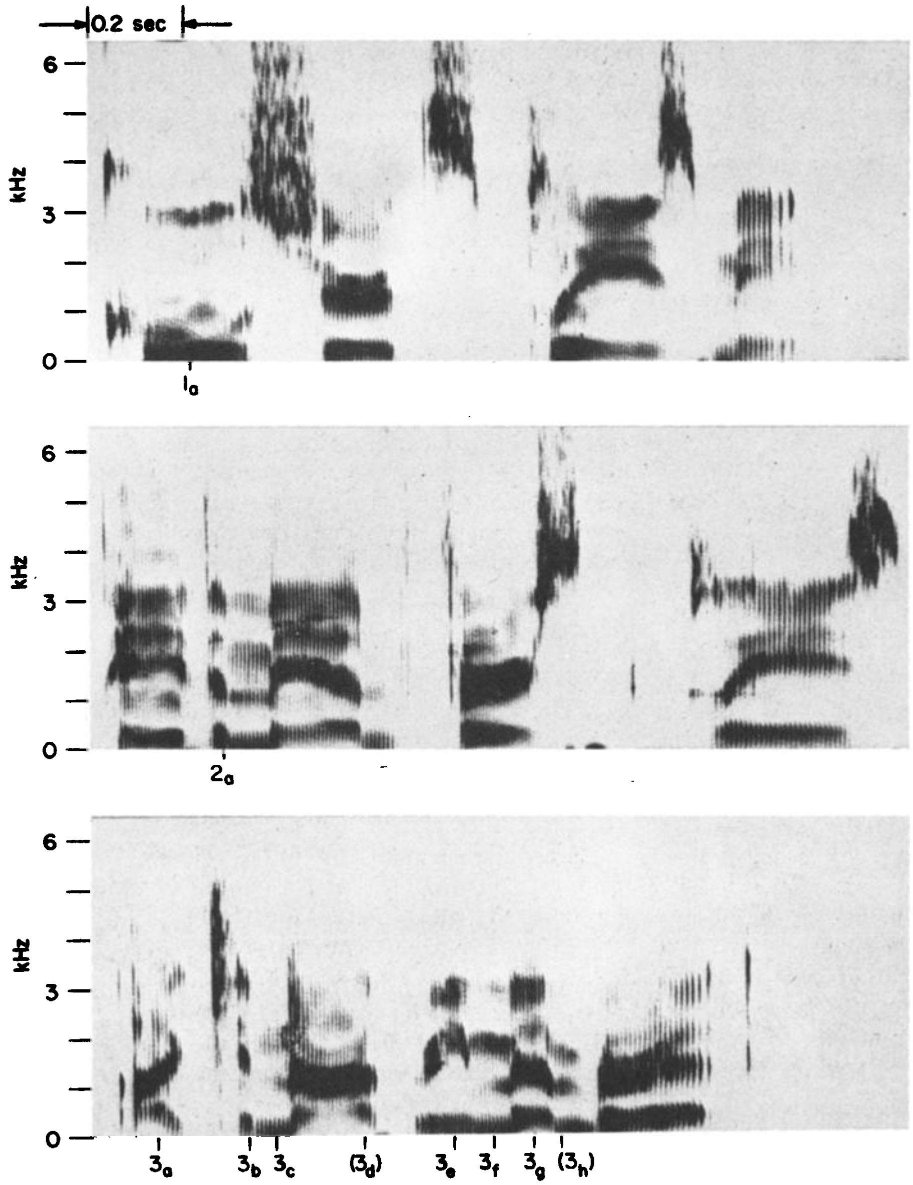

FIG. B-1. Spectrograms of sentences 1, 2, and 3. *Top*: "Cool shirts please me." *Middle:* "Pay the man first, please." *Bottom:* "I cannot remember it."

B. Nasal Consonants

These parameters are individual filter outputs, normalized by the amplitude of the following vowel, in nasals in sentence 3. Owing to their lower amplitudes, the nasals show clearly as depressions in the energy function (see Fig. 2). The parameter names are 3Mi and 3Ni, where i is the filter number (see Fig. 4). Parameters 3Mi are located at the middle of the first [m] in "remember" (point 3f), and 3Ni are at the middle of [n] in "cannot" (point 3c).

C. Vowels

The moment calculations were performed on the spectra in the form in which they were displayed, i.e., simply using the filter number as abcissa and amplitude (in decibels, and including the +6 dB/oct preemphasis) as ordinate. In order that the over-all amplitude should not affect the result, each ordinate was expressed as a distance above the minimum amplitude within the range of filters specified.

IS2, IU3: Second and third central moments in the frequency range 1.5–4.5 kHz, at the middle of [i] in "needs" in sentence 4 (point 4b).

AS2, AU3: Second and third central moments in the range 0.5–1.5 kHz, at the middle of [a] in "papa" (point 4a).

UHF1, UHF2: Estimated F_1 and F_2 at the middle of the vowel in *the* in sentence 2 (point 2a).

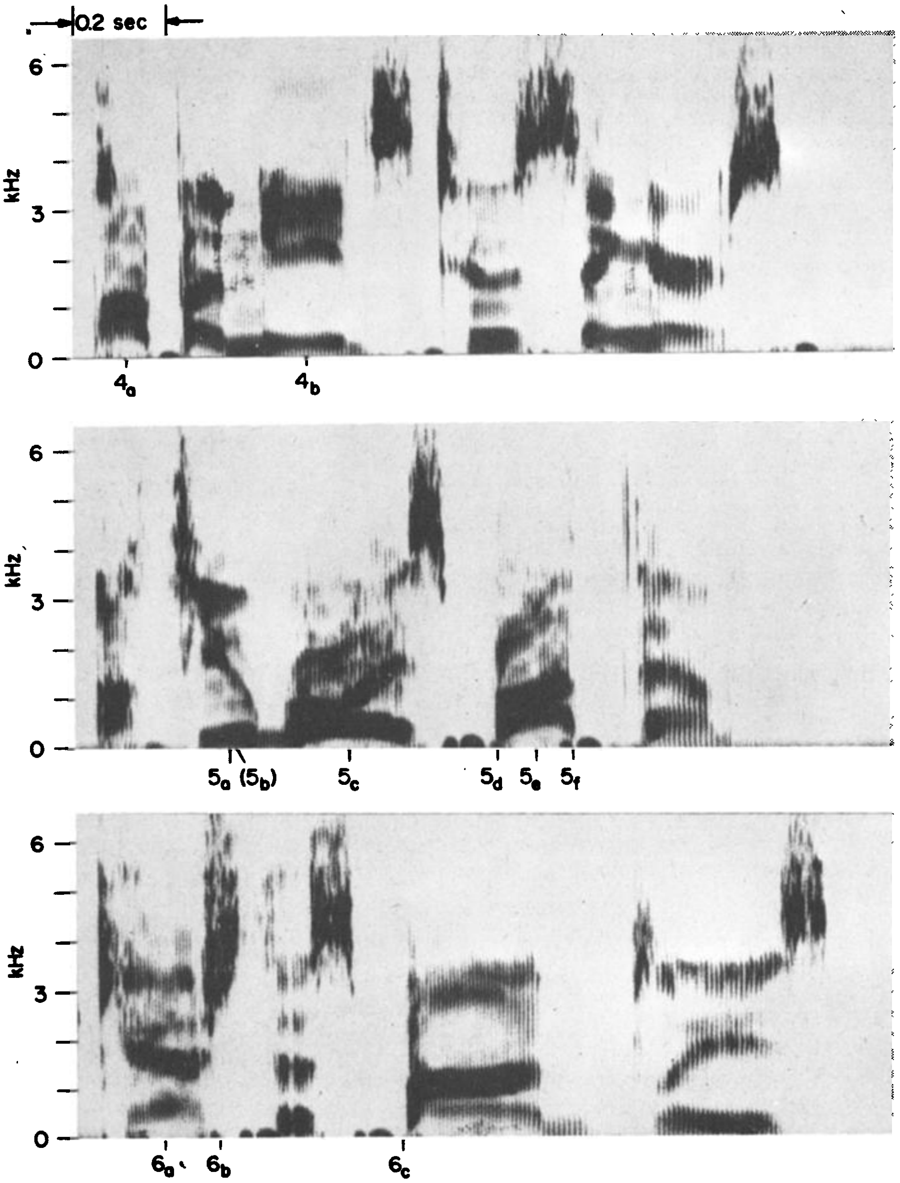

FIG. B-2. Spectrograms of sentences 4, 5, and 6. *Top:* "Papa needs two singers." *Middle:* "A few boys bought them." *Bottom:* "Cash this bond, please."

AEF1, AEF2: F_1 and F_2 at the same location as AEF0.

AF1, AF2: F_1 and F_2 at the same location as AS2.

D. Other Parameters

SSS: Source spectrum slope measurement during [u] in "cool," at a point one-third of the way through the voiced portion of this syllable (point 1a).

SH: High-frequency spectrum shape parameter at the middle of [ʃ] in sentence 6 (point 6b).

BAWT: Time interval between the half-amplitude points in the energy function corresponding to the beginning and end of "bought" (points 5d and 5f).

PREV: Prevoicing measurement in the [b] release in "bond" (point 6c).

* This paper is mainly based on "Acoustic Measurements for Speaker Recognition," PhD thesis, Dept. of Electrical Engineering, MIT (1969).

[1] J. Edie and G. Sebestyen, "Voice Identification General Criteria," RADC-TDR-62-278, Litton Systems, Inc. (1962).

[2] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Amer. 35, 354–358 (1963).

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

[4] M. H. Becker, R. Gnanadesikian, M. V. Mathews, R. S. Pinkham, S. Pruzansky, and M. B. Wilk, "Comparisons of Some Statistical Distance Measures for Talker Identification," J. Acoust. Soc. Amer. 36, 1988(A) (1964).

[5] K.-P. Li, J. E. Dammann, and W. D. Chapman, "Experimental Studies in Speaker Verification, Using an Adaptive System," J. Acoust. Soc. Amer. 40, 966–978 (1966).

[6] W. F. Meeker, T. B. Martin, M. B. Herscher, D. Phyfe, and M. Weinstock, "Automatic Speaker Recognition Using Speech Recognition Techniques," J. Acoust. Soc. Amer. 42, 1182(A) (1967).

[7] J. W. Glenn and N. Kleiner, "Speaker Identification Based on Nasal Phonation," J. Acoust. Soc. Amer. **43**, 368–372 (1968).

[8] J. E. Luck, "Automatic Speaker Verification Using Cepstral Measurements," J. Acoust. Soc. Amer. **46**, 1026–1032 (1969).

[9] S. K. Das and W. S. Mohn, "A Scheme for Speech Processing in Automatic Speaker Verification," IEEE Trans. Audio Electroacoust. **AU-19**, 32–43 (1971).

[10] P. Garvin and P. Ladefoged, "Speaker Identification and Message Identification in Speech Recognition," Phonetica **9**, 193–199 (1963).

[11] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," PhD thesis, Dept. of Electrical Engineering, Polytechnic Inst. of Brooklyn (1968).

[12] K.-P. Li, G. W. Hughes, and A. S. House, "Approaches to the Characterization of Talker Differences by Statistical Operations on Speech Spectra," J. Acoust. Soc. Amer. **47**, 66(A) (1970).

[13] C. Bordone-Sacerdote and G. G. Sacerdote, "Some Spectral Properties of Individual Speakers," Acustica **21**, 199–210 (1969).

[14] A. M. Noll, "Cepstrum Pitch Detection," J. Acoust. Soc. Amer. **41**, 293–309 (1967).

[15] B. Gold and L. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Amer. **46**, 442–448 (1969).

[16] 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. Amer. **33**, 1725–1736 (1961).

[17] W. L. Henke, "TASS—Another Terminal Analog Speech Synthesis System," Res. Lab. of Electron., MIT, Quart. Prog. Rep. No. 95 (1969), pp. 73–81.

[18] C. G. M. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960), Chap. 2.4, pp. 139–148.

[19] O. Fujimura, "Analysis of Nasal Consonants," J. Acoust. Soc. Amer. **34**, 1865–1875 (1962).

[20] J. E. Miller, "Decapitation and Recapitation, a Study in Voice Quality," J. Acoust. Soc. Amer. **42**, 2002(A) (1964).

[21] P. Ladefoged and D. E. Broadbent, "Information Conveyed by Vowels," J. Acoust. Soc. Amer. **29**, 98–104 (1957).

[22] L. J. Gerstman, "Classification of Self-Normalized Vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 73–77 (1968).

[23] K. N. Stevens and A. S. House, "Perturbations of Vowel Articulations by Consonantal Context: An Acoustical Study," J. Speech Hearing Res. **6**, 111–123 (1963).

[24] K. N. Stevens, "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," in *Human Communication: A Unified View*, E. E. David, Jr., and P. B. Denes, Eds. (McGraw-Hill, New York, in press).

[25] J. Hemdal, "Some Results from the Normalization of Speaker Differences in a Mechanical Vowel Recognizer," J. Acoust. Soc. Amer. **41**, 1594(A) (1967).

[26] A. P. Paul, A. S. House and K. N. Stevens, "Automatic Reduction of Vowel Spectra: An Analysis-by-Synthesis Method and Its Evaluation," J. Acoust. Soc. Amer. **36**, 303–308 (1964).

[27] J. Mártony, "Studies of the Voice Source," Speech Transmission Lab., Roy. Inst. Technol. (Stockholm), Quart. Progr. Status Rep. No. 1, 4–9 (1965).

[28] L. Lisker and A. S. Abramson, "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," Word **20**, 384–422 (1964).

[29] M. H. L. Hecker, K. N. Stevens, G. von Bismarck, and C. E. Williams, "Manifestation of Task-Induced Stress in the Acoustic Speech Signal," J. Acoust. Soc. Amer. **44**, 993–1001 (1968).