# Speaker-identifying features based on formant tracks

Ursula G. Goldstein

*Department of Electrical Engineering and Computer Science and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*
(Received 21 July 1975; revised 17 September 1975)

The formant structure of three diphthongs, four tense vowels, and three retroflex sounds was examined in detail for possible speaker-identifying features. These sounds were spoken five times each in sentence context by ten speakers of General American on one day and by six of the speakers on a second day at least three weeks later. Formant tracks were computed for each sound under investigation using covariance-type pitch-asynchronous linear prediction together with a root-finding algorithm. The interspeaker variability of about 200 measurements made on these formant tracks was compared initially with intraspeaker variability through the calculation of $F$ ratios. Those with average $F$ ratios greater than 60 were evaluated further with a probability-of-error criterion. Features that are potentially most effective in identifying speakers are the minimum second-formant value in [ɑr], the maximum first-formant value in [ɑr], the maximum second-formant values of [o] and [ɔɪ], and the minimum third-formant value of [ɝ]. The individual differences apparent in these sounds presumably depend more on speaker habits than on vocal-tract anatomy. The error bound predicted for a speaker identification procedure based on these five features is 0.24%. An identification experiment using only the best two features gave 12 errors out of 80 identifications.

Subject Classifications: [43]70.65, [43] 70.40.

## INTRODUCTION

Methods of achieving automatic speaker recognition have generally fallen into two categories: template matching and acoustic feature extraction. The template-matching method requires the test utterance to be the same as the utterance used to create the sample. Several parameters may be extracted, generally as functions of time, without the help of segmentation. After undergoing some form of time and amplitude normalization, a distance is calculated between the unknown and the reference. If the distance is larger than a certain threshold, an answer of no match is given.[1-6] This approach is most applicable to speaker verification, where the speaker is cooperative and does not purposely introduce large variations into the speech sample.

The speech theoretic approach examines linguistic units from which it tries to extract an optimum set of features, thereby eliminating from further consideration some of the information in the speech signal that does not pertain to the speaker's identity. Several criteria have been suggested for selecting these features.[7] They should occur frequently in normal speech, vary widely between speakers but not for a given speaker, not change over time, not be affected by background noise or poor transmission, not be affected by conscious efforts to disguise the voice, and be easily measurable.

The list of possible identifying features is virtually endless, and has only been partially examined. A data base that has shown promise for providing a number of useful features is the set of formant tracks obtained from diphthongs, tense vowels, and r-colored sounds. As reported in the literature, these sounds have evidenced a large amount of variability from one speaker to another, especially across different dialects.[8-11]

The use of formant information in speaker identification systems has been limited almost exclusively to the measurement of formant frequencies inside a single window at the center of a vowel, leaving much of the formant structure unexplored.[7,12] One measurement that did include a larger amount of formant-structure information was the slope estimate of the second formant of [ɑɪ], which Sambur[12] ranked as the tenth best of a large number of attributes he examined. The success of this measurement gave further evidence that a closer examination of formant tracks might reveal some useful speaker-identifying features.

The purpose of this study was to investigate the formant trajectories of the diphthongs [ɔɪ], [ɑɪ], and [ɑʊ], the tense vowels [o], [e], [i], and [u], and retroflex sounds in three phonetic environments [re], [ɝ], and [ɑr], in order to find and statistically evaluate features that could be relevant to speaker identification.

## I. PROCEDURE

### A. Data base

Ten adult male speakers of American English with no noticeable foreign accents, strong regional dialects, or speech defects were chosen to make recordings of the sounds to be studied. In order to facilitate comparison of one sound with another, all sounds were recorded in the context, "Say b_d again." Each person spoke five repetitions of each of the ten sentences in one recording session. Six of the original ten speakers returned at least three weeks after the first recording session to make another set of recordings, again with five repetitions of each of the ten sentences. This second set of recordings was made in order to include the effects of changes in a speaker's voice over time.

### B. Formant tracking procedure

Recordings were processed semiautomatically using linear prediction analysis on a PDP-9 computer specially set up for speech analysis. The software written for this purpose seeks to minimize the amount of time


Copyright © 1976 by the Acoustical Society of America

needed to compute a highly accurate set of formant tracks, both in terms of computer time and operator time. However, for several reasons, no attempt was made to fully automate this procedure. Currently, it is not possible to devise an automatic formant-tracking program which gives four formants with 100% reliability. The best systems giving three formants generally have very complicated decision algorithms, and even then cannot guarantee 100% reliability.[13] Since an error in formant identification can produce serious errors in a speaker identification scheme based on specific formant frequencies, and since the major emphasis of the study is on the use of formant tracks rather than on their computation, it was decided that an interactive system should be constructed. With such a system, the operator can observe the results of a first, automatic stage of tracking and can intervene manually to correct errors. Informal observations by the program user during the process of trying to identify missing formants or eliminate extraneous ones could be useful in the construction of a more automatic system at some later time.

Audio input was preemphasized 6 dB/octave, band-limited to 5 kHz, and sampled at 10 kHz. The sampled signal was displayed on a cathode-ray tube, and then marked by hand to indicate the beginning and end of the portion to be processed. The beginning was defined at 20 msec after the noise burst indicating the release of [b]. The end was marked when a sudden drop in amplitude and an obvious loss of high frequencies indicated the closure for [d].

The first main processing loop of the formant-tracking program calculates 12 predictor coefficients for each 10 msec frame, using the covariance method pitch asynchronously.[14,15] The second loop calculates pole locations of the transfer function for each frame and then transforms them to formants and bandwidths.[16] Formants with bandwidths greater than 700 Hz are removed from the general formant array and placed in a temporary location for extraneous formants. The last phase of the program displays a set of formant tracks, as shown in Fig. 1, and allows corrections to be entered manually by the operator, according to continuity considerations and his knowledge of acoustic theory. Formants can be restored from temporary locations if they are judged as not extraneous. Formants can also be removed if they seem extraneous but were not automatically eliminated. On very rare occasions, the root-finding subroutine could not find the roots in ten iterations. In this case, the formant frequencies were set to zero by the program and were later filled in by averaging the formants of the previous and the next frame. During the course of the measurements, this situation only occurred twice. Corrected formant tracks were appropriately labeled and stored on digital tape.

Occasionally, a situation arose when the user had trouble deciding which poles to choose as formants. To help him make a decision, the program calculated a log-magnitude plot of the spectrum and a pole plot in the $z$ plane for any frame indicated by the user. In actual practice, however, these plots rarely gave more insight than the actual formant frequencies and bandwidths.

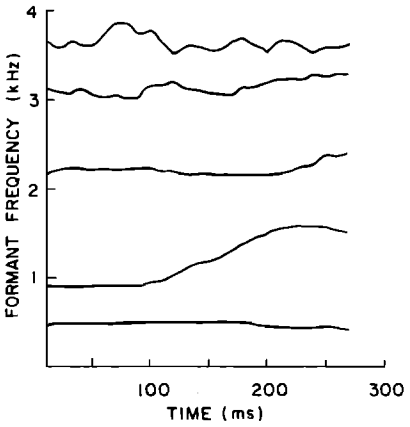


FIG. 1. Example of a formant-track display for the sound [aɪ]. The lines represent time plots of formant frequencies, the lowest being the first formant and the highest line being the fifth formant. The line representing a particular formant is formed by connecting the estimates of that formant frequency from one frame to the next with straight-line segments.

For about five sentences, it was not possible to mark the beginning and end of processing from a simple visual examination of the time waveform, due to imprecise pronunciation by the subject. In these cases, analysis was started well before the expected beginning and stopped after the expected end. The two end markers were then readjusted to the two points where gross discontinuities in the formants indicated the presence of a consonant.

## II. FEATURE SELECTION AND EVALUATION

The formant tracks, as determined by the methods indicated above, contain a mixture of noise, linguistic information, and personal information. The next phase of a speech-theoretic speaker identification system would be to extract the personal information pertaining only to the identity of the speaker. In this study, 199 measurements were made and evaluated in terms of effectiveness in speaker identification. A list of the names given to the features, a short description of each one, and a list of the sounds to which each was applied is given in Table I.

The data can be divided into three groups according to when they were recorded. Recordings from the first day for all ten speakers form one group, recordings from the second day for six of the ten speakers form the second group, and recordings from both days for the six speakers form the third group. A preliminary statistical evaluation consisted of computing the ratio of between-speaker variance to within-speaker variance, known as an $F$ ratio, for each feature for each of the three time groupings. The three $F$ ratios of each feature were then averaged, and only the 29 features having average $F$ ratios greater than 60 were retained for further consideration.

A higher $F$ ratio indicates a feature that exhibits larger interspeaker variation in relation to the intra-

TABLE I. Feature names, descriptions, and sound segments to which descriptions apply.

| Name | Sound segments | Description |
|---|---|---|
| AVEF3 | ɔɪ, aɪ, aʊ e, i, o, u | average third formant not including last 20 msec of formant track |
| AVEF4 | all | average fourth formant not including last 20 msec of formant track |
| AVE3M2 | ɛr, ar, ɝ | average third minus second formant omitting last 20 msec of formant track |
| DUR | all | total duration |
| DUR1 | ɔɪ, aɪ, aʊ | duration 1, measured from beginning of formant tracks to middle of second-formant glide |
| DUR2 | ɔɪ, aɪ, aʊ | duration 2, measured from middle of second-formant glide to end of formant tracks |
| FINAL1 | ɛr, ar | F1 measured 20 msec before end of formant tracks |
| FINAL2 | ɔɪ, aɪ, aʊ, ɛr, ar, e, i, o, u | F2 measured 50 msec before end of formant tracks on [aʊ], [o], and [u]; 20 msec before end on all other sounds |
| FINAL3 | ɛr, ar | F3 measured 20 msec before end of formant tracks |
| FINAL4 | ɛr, ar | F4 measured 20 msec before end of formant tracks |
| F1MAX2 | ɔɪ, aɪ, aʊ, e, i, o, u | F1 at point in time when F2 reaches a maximum |
| F1MAX3 | ɛr, ar | F1 at point in time when F3 reaches a maximum |
| F1MIN2 | ɔɪ, aɪ, aʊ, e, ɝ i, o, u | F1 at point in time when F2 reaches a minimum |
| F1MIN3 | ɛr, ar, ɝ | F1 at F3 minimum |
| F2MAX3 | ɛr, ar | F2 at F3 maximum |
| F2MIN3 | ɛr, ar, ɝ | F2 at F3 minimum |
| F4MAX3 | ɛr, ar | F4 at F3 maximum |
| F4MIN3 | ɛr, ar, ɝ | F4 at F3 minimum |
| INITL1 | ɔɪ, aɪ, aʊ, ɛr, ar, e, i, o, u | initial F1, measured 20 msec after beginning of formant track for [ɛr] and [ar]; at beginning of track for all other sounds |
| INITL2 | ɔɪ, aɪ, aʊ, ɛr, ar, e, i, o, u | initial F2, measured 20 msec after beginning of formant track for [ɛr] and [ar]; at beginning of track for all other sounds |
| INITL3 | ɛr, ar | initial F3, taken 20 msec after beginning |
| INITL4 | | initial F4, taken 20 msec after beginning |
| LOCALS | ɔɪ, aɪ, aʊ, ɛr, ar | local slope, measured by fitting a straight line over 110 msec around point of maximum slope of F2 for [aɪ] and [ɔɪ]; measured on F3 for [ɛr] and [ar]; measured on F2 from beginning to MINF2 for [aʊ] |
| MAXF1 | all | maximum first formant |
| MAXF2 | all | maximum second formant |
| MAXF3 | ɛr, ar | maximum third formant |
| MAXF4 | ɝ | maximum fourth formant |
| MIDF1 | e, i, o, u, ɝ | first formant at midpoint of formant tracks |
| MIDF2 | e, i, o, u, ɝ | second formant at midpoint |
| MIDF3 | ɝ | third formant at midpoint |
| MIDF4 | ɝ | fourth formant at midpoint |
| MID1AV | e, i, o, u, ɝ | MIDF1 averaged with the two first-formant values on either side of it |
| MID2AV | e, i, o, u, ɝ | MIDF2 averaged with the two surrounding values |
| MID3AV | ɝ | MIDF3 averaged with the two surrounding values |
| MID4AV | ɝ | MIDF4 averaged with the two surrounding values |

TABLE I (*Continued*)

| Name | Sound segments | Description |
|---|---|---|
| MINF2 | all | minimum second formant |
| MINF3 | ɛr, ar, ɝ | minimum third formant |
| MINF4 | ɝ | minimum fourth formant |
| MIN3AV | ɝ | MINF3 averaged with the two surrounding values |
| PCTDUR | ɔɪ, aɪ, aʊ | percent duration one = (DUR1/DUR) ×100 |
| RANGE2 | ɔɪ, aɪ, aʊ, e, i, o, u | range of second formant = MAXF2 – MINF2 |
| RANGE4 | ɛr, ar | range of fourth formant = MAXF4 – MINF4 |
| STDF1 | ɔɪ, aɪ, aʊ, e, i, o, u | standard deviation of F1 over its entire duration |
| TOTALS | ɔɪ, aɪ, aʊ, e, i, o, u, ɝ | total slope, measured by fitting a straight line to F3 for [ɝ] and to F2 for all other sounds over the total duration of the formant in question |
| 1MIN3A | ɝ | F1MIN3 averaged with the two surrounding points |
| 2MIN3A | ɝ | F2MIN3 averaged with the two surrounding points |
| 4MIN3A | ɝ | F4MIN3 averaged with the two surrounding points |

speaker variation and is, therefore, usually more useful for speaker identification than a feature with a lower $F$ ratio. The calculation of $F$ ratios has the advantage of being fast and easy to implement—a very desirable advantage when dealing with a large number of features. However, the $F$ ratio does not give any information about possible dependence between features, and also tends to give overly high values for features where one or two speakers are very different from the rest. Therefore, average $F$ ratios were used as a means of eliminating the least promising features, with the final selection being done by other methods.

The 29 features having average $F$ ratios greater than 60 are listed in Table II. Several sets of features are obviously redundant, since they represented essentially the same acoustical property but with slightly different measurement procedures. Therefore, it is best to keep only the one feature in each set with the highest $F$ ratio. This procedure eliminates MAXF2 for [aɪ], MID3AV for [ɝ], MINF3 for [ɝ], MIDF3 for [ɝ], and MINF2 for [e].

The final evaluation was done using the probability-of-error method described by Sambur,[12] using all data available for the 24 remaining features. The probability-of-error method assumes that the underlying distributions of feature values for each speaker are multivariate Gaussian. The probability of falsely identifying one speaker as another when only two speakers are present is determined by the amount of overlap between distributions of feature values of the two speakers. The probability of error with all speakers present can be estimated from the union error bound.

Instead of the "knock-out" procedure of feature ordering used by Sambur, an "add-on" procedure similar to the sequential forward selection of Ichino and Hiramatsu[17] was used. This procedure involves computing the probability of error associated with each of the $N$ features taken separately, and then assigning a

TABLE II. Features having average F ratios greater than 60.

| Feature | Sound | F ratio |
|---|---|---|
| MAXF1 | ɑr | 119.3 |
| MAXF1 | e | 107.4 |
| MAXF1 | o | 103.6 |
| MINF2 | ɑr | 97.2 |
| AVEF4 | ɑʊ | 97.1 |
| MIN3AV | ɝ | 92.2 |
| FINAL2 | ɑɪ | 90.8 |
| AVEF4 | ɑɪ | 88.1 |
| MAXF1 | u | 86.2 |
| MAXF2 | ɑɪ | 86.0 |
| MID3AV | ɝ | 80.7 |
| MINF3 | ɝ | 77.3 |
| AVE3M2 | rɛ | 74.0 |
| AVEF4 | ɑr | 74.0 |
| INITL2 | e | 73.4 |
| MIDF3 | ɝ | 73.4 |
| MINF2 | e | 73.4 |
| F1MAX2 | o | 72.5 |
| MAXF2 | o | 71.8 |
| AVEF3 | u | 71.0 |
| MAXF1 | ɝ | 70.6 |
| MAXF2 | ɑʊ | 69.0 |
| MAXF2 | ɔɪ | 65.8 |
| F1MAX2 | u | 65.4 |
| MID2AV | i | 63.7 |
| MID1AV | o | 63.2 |
| FINAL2 | e | 62.5 |
| INITL4 | ɑr | 61.5 |
| INITL2 | i | 60.9 |

rank of 1 to the feature giving the lowest error. Next, a set of $N-1$ pairs is evaluated, each including feature No. 1. The feature that together with feature No. 1 gave the lowest probability of error is then ranked as No. 2. The procedure is continued until as many features as desired have been ranked. The "knock-out" procedure, on the other hand, starts by computing error bounds for all combinations of $N-1$ features and then eliminating the one feature not present in the most effective combination. The procedure is repeated until only one feature is left.

The "add-on" procedure has two main advantages: it is computationally more efficient, and allows one to rank order a subset of best features without ordering the entire set of $N$ features. The computational savings of the "add-on" method over the "knock-out" method are dependent on the number of features and the number of speakers. Based on the number of multiplications involved, the "add-on" method can be expected to produce an ordered list of features about three times as fast as the "knock-out" method for 24 features and ten speakers.

A list of the ten best features is given in Table III in order of decreasing usefulness for speaker identification. Next to each feature is the total error bound expected for that feature together with all features listed above it. This list is very similar to an earlier one which had been prepared from Table II without the help of the probability-of-error computation.[18] In the earlier analysis, correlation coefficients were used to eliminate redundant features and a visual examination of speaker means was used to eliminate features with artificially high $F$ ratios.

An identification experiment was performed using only the two best features: MINF2 of [ɑr] and MAXF1 of [ɑr]. A scatter plot of the data involved is shown in Fig. 2. Eighty identifications were made by using each data point in turn as a test sample, with the rest of the data forming the design set. Classification was performed using the optimum linear classifier described by Sambur. Twelve errors were made, giving an error rate of 15% as predicted in Table III.

## III. DISCUSSION

One of the original motivations for studying diphthongs and r-colored sounds was the large dialect variation shown by these sounds in American English.[9,11] As can be seen by examining Table III, features derived from these sounds, which presumably depend more upon speaker habits than on vocal-tract anatomy, in fact showed large individual differences. Also, the first-formant measures were generally uncorrelated with the second-formant measures, indicating that one or both of these types of measures contain more information than just the overall length of the vocal tract.

One feature that was particularly uncorrelated with any of the others was MINF2 for [ɑr]. This feature gives an indication of how strongly a speaker's [ɑ] is affected by the adjacent [r] and may also depend on the way he shapes his tongue blade for the retroflex [r]. This notion is supported by the low correlation coefficients between MINF2 for [ɑɪ] and MINF2 for [ɑr]: 0.56 for day 1 and 0.26 for day 2. Five other features of Table III are related to retroflex sounds. MIN3AV for [ɝ] is a direct indication of the degree of retroflexion while AVE3M2 for [rɛ] takes into account both the degree of retroflexion and the duration of [r] relative to [ɛ]. MAXF1 for [ɑr] and [ɝ] show the influence of the retroflex on F1 of the vowel.

Although AVEF4 of [ɑr] is also a measurement taken from a retroflex sound, it probably reflects mainly vocal-tract shape and size, as does AVEF3 of [u].

One feature in the list in Table III was a measurement made on a mid vowel. Since mid vowels are not produced with an extreme high or low tongue position,

TABLE III. The ten best features listed in order of decreasing effectiveness.

| Rank | Name | Error bound in % |
|---|---|---|
| 1 | MINF2 ɑr | 70. |
| 2 | MAXF1 ɑr | 15. |
| 3 | MAXF2 o | 3.5 |
| 4 | MIN3AV ɝ | 0.93 |
| 5 | MAXF2 ɔɪ | 0.24 |
| 6 | AVEF3 u | 0.061 |
| 7 | AVE3M2 rɛ | 0.015 |
| 8 | MAXF2 ɑʊ | 0.0040 |
| 9 | AVEF4 ɑr | 0.0012 |
| 10 | MAXF1 ɝ | 0.00026 |

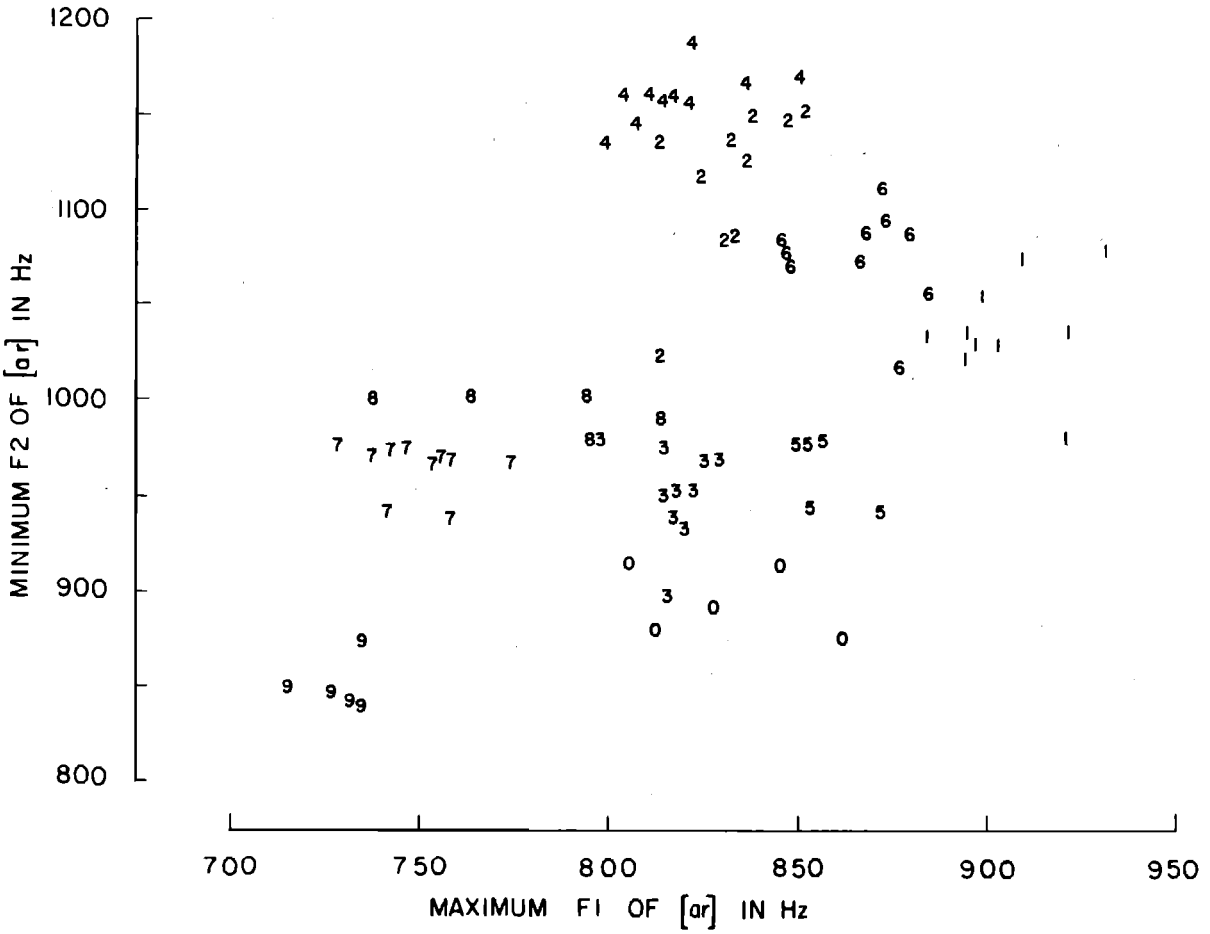


FIG. 2. Scatter plot of data for the two most effective features. The ten speakers are represented by ten different digits.

they may be subject to more individual variation than [i] or [u]. Their acoustic characteristics might also reflect the shape of the palate, since these sounds are produced with the sides of the tongue in contact with the lower edges of the palate.[19] Two features in the table are taken from vowel targets of diphthongs. Since the main acoustic cues distinguishing the three diphthongs are the rate and direction of formant-frequency changes,[20] one might expect the target frequencies of the first and second formants to reflect a person's individual speaking habits.

The list of "best" features includes three measures involving the third formant and one involving the fourth. Of these four features, the only one for which the formant extraction appeared completely reliable was MIN3AV for [ɝ], as evidenced by the consistently narrow bandwidths and good continuity from one frame to the next for F3 of [ɝ]. This was not always the case with AVEF4. For example, the $F$ ratio of AVEF4 for [o] was extremely low because of measurement difficulties encountered with one speaker. On two repetitions, a resonance that had been identified as the fourth formant during the other three repetitions, was too weak to be detected, possibly due to a zero in the spectrum of either the glottal source or the vocal-tract filter. Spectrographic analysis confirmed the problem. This finding might indicate that the fourth formant makes a rather unreliable feature, and that the high ranking of AVEF4 for [ɑr] was coincidental. Similar measurement problems were encountered with the third formant of another speaker.

Considering the trouble given by the higher formants during the formant-tracking procedure, it might be more reasonable not to even try to measure them. A system where the speech waveform was low-pass filtered to 2500 Hz and then sampled at 5000 Hz would allow at most three formants. The linear-prediction program could then be run with eight predictor coefficients, and the root finder would have to solve only an eighth-order polynomial. With very little loss of information, this system could be expected to run about twice as fast as the one used in this study.

## IV. COMPARISON OF BEST FEATURES WITH PREVIOUS WORK

A direct comparison of this study with the work of Sambur[12] is not totally justified, since this study involved ten speakers and Sambur based his error analysis on 11 speakers. Also, this study did not make use of any consonant measurements, which Sambur found highly effective. However, some conclusions can be drawn from the results of two features that were evaluated both in this study and in that of Sambur.

The best vowel measurement found by Sambur was the third formant of [u], which he ranked as second best. A good approximation of this parameter is AVEF3 of [u], which was ranked as sixth best in this study. The

slope of the second formant of [aɪ], which was ranked as tenth best in Sambur's work, showed an intermediate degree of effectiveness in this work. Since its average $F$ ratio of 39.6 was well below 60, it was not included in the probability-of-error evaluation.

Sambur's work suggests that first-formant measurements are less important for speaker identification than second- or third-formant measurements, whereas in this study two of the ten most effective features were first-formant measurements. A closer examination of the data from F1 measurements shows a very large variation of $F$ ratios according to exactly where the measurement was taken. For example, $F = 119.3$ for MAXF1 of [ɑr], but $F = 13.8$ for FINAL1 of [ɑr]. The maximum value of F1 generally yielded the highest $F$ ratios and was characterized by the lowest intraspeaker standard deviations. Informal observation of formant tracks revealed that F1 rarely reached a maximum at the same time when F2 reached a maximum or a minimum.

Ten measurements found in this study had higher $F$ ratios than the best feature found by Wolf,[7] which had an $F$ ratio of 84.9. Wolf's nine best features were fundamental frequency measures, which Sambur downrated because of their variability from one recording session to another. The feature with the tenth largest $F$ ratio in Wolf's study was the second formant frequency of [æ], having an $F$ ratio of 46.6. This value is considerably lower than the $F$ ratios of the better features of this study.

## V. THE LIMITATIONS OF THE STUDY

The most obvious limitation of this study is the small data base that was used, and, in particular, the very limited amount of data concerning speaker variation over time. Another problem is the somewhat artificial nature of the recordings used. The recordings were made under ideal laboratory conditions with the subjects reading a prearranged set of sentences. In practice, one will probably have to cope with background noise, distortion, or band limiting of transmission equipment, and different sentence contexts for the sounds under investigation. Other complications include the possibility of an alteration in a person's voice due to emotional or physical stress, or an uncooperative speaker, i.e., a person who is trying to disguise his voice or mimick another person. All of the speakers in the current study were cooperative and under no particular stress.[21]

On the other hand, the speakers for this study did not represent a complete cross section of all dialects of American English. The speakers that were originally chosen for the study had no noticeable foreign accents and no strong regional dialects. Therefore, the interspeaker variations of some of the features are probably not as high as they might have been with a more varied group of subjects.

Besides extending the work of this study to include some of the additional variables mentioned above, it would be useful to test the more effective features together with some of the more successful features of other studies such as the second formant of [n], the voice-onset time of [k], the third and fourth formants of [m], and an F0 measurement. Next, one might run an identification experiment with this combined feature set, using speakers who had not been involved in the original feature evaluation.

Another area of interest might be the study of the higher formants; why they appear and disappear unexpectedly, and how to compensate for this problem. If the higher-formant measuring problems with two of the speakers in this study were caused by zeros in the spectrum, perhaps one could devise a system to indicate the presence of a zero, and then determine its frequency.[22]

## ACKNOWLEDGMENT

The author wishes to thank Professor Kenneth N. Stevens for his help and guidance during the course of this study and in the preparation of the manuscript.

*This work was supported in part by the Office of Naval Research.

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

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

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

[4]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).

[5]R. C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio Electroacoust. AU-21, 80–89 (1973).

[6]A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker Verification," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 169–176 (1975).

[7]J. J. Wolf, "Efficient Acoustic Parameters for Speaker Recognition," J. Acoust. Soc. Am. 51, 2044–2056 (1972).

[8]A. Holbrook and G. Fairbanks, "Diphthong Formants and their Movements," J. Speech Hear. Res. 5, 38–58 (1962).

[9]W. Labov, M. Yaeger, and R. Steiner, "A Quantative Study of Sound Change in Progress," Report on National Science Foundation Contract NSF-GS-3287, Univ. Pennsylvania, Philadelphia, PA (1972).

[10]D. Klatt, "Acoustic Characteristics of /w, r, l, y/ in Sentence Contexts," J. Acoust. Soc. Am. 55, 397(A) (1974).

[11]H. Kurath, *Handbook of the Linguistic Geography of New England* (The American Council of Learned Societies, Providence, RI, 1939).

[12]M. R. Sambur, "Selection of Acoustic Features for Speaker Identification," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 176–182 (1975).

[13]S. S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans. Acoust. Speech Signal Process. ASSP-22, 135–141 (1974).

[14]J. I. Makhoul and J. J. Wolf, "Linear Prediction and the Spectral Analysis of Speech," BBN Report 2304, Bolt Beranek and Newman, Cambridge, MA (1972).

[15]J. I. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE 63, 561–580 (1975).

[16]M. A. Jenkins and J. F. Traub, "A Three-Stage Algorithm for Real Polynomials Using Quadratic Iteration," SIAM J. Numer. Anal. 7, 545–566 (1970).

[17]M. Ichino and K. Hiramatsu, "Suboptimum Linear Feature

Selection in Multiclass Problem," IEEE Trans. Syst. Man Cybern. **4**, 28–33 (1974).
[18]U. G. Goldstein, "An Investigation of Vowel Formant Tracks for Purposes of Speaker Identification," S. M. Thesis, Dept. of Electrical Engineering, MIT, Cambridge, MA, (1975).
[19]K. N. Stevens, "Quantal Configurations for Vowels," J. Acoust. Soc. Am. **57**, Suppl. 1, 70(A) (1975).
[20]T. Gay, "A Perceptual Study of American English Diphthongs," Lang. Speech **13**, 65–88 (1970).
[21]T. H. Crystal, H. Gish, and R. F. Bloom, "Psychophysiological Factors Affecting Speaker Authentication and Identification," U. S. Army Elect. Command, Tech. Rep. ECOM-0161-F, Fort Monmouth, NJ (1973).
[22]J. M. Tribolet, "Identification of Linear Discrete Systems with Applications to Speech Processing," S. M. thesis, Dept. of Electrical Engineering, MIT, Cambridge, MA (1974).