TEXT-INDEPENDENT SPEAKER IDENTIFICATION FROM A LARGE LINGUISTICALLY UNCONSTRAINED TIME-SPACED DATA BASE

John D. Markel and Steven B. Davis

Signal Technology, Inc.
15 W. De La Guerra Street
Santa Barbara, CA 93101

Haskins Laboratories
270 Crown Street
New Haven, CT 06511

Abstract

A very large data base consisting of over thirty-six hours of linguistically unconstrained extemporaneous speech, from seventeen speakers, recorded over a period of more than three months, was analyzed to determine the effectiveness of long-term average features for speaker identification. The results were strongly dependent on the voiced speech averaging interval, or L_V. Monotonic increases in the probability of correct identification were obtained as L_V increased, even with substantial time periods between successive sessions. Speaker identification performance in open tests improved if features with small between-class to within-class variance ratios were eliminated. For L_V corresponding to approximately thirty-nine seconds of speech, true text-independent results (no linguistic constraints embedded into the data base) of 98.05% for speaker identification were obtained.

Introduction

In recent years, there has been an increasing interest in computer-based techniques for text-independent speaker identification (1-4). The term "text-independent" has been used in several different contexts. For example, Atal (1) has used the term in the sense of choosing independent randomized test frames from a single sentence to use against the remaining frames as a reference set. Sambur (4) has used the term in an experiment where the sentences in the test set were different from those in the reference set, even though each speaker read precisely the same list of sentences.

This study presents results for speaker identification with no linguistic constraints on the speech content (other than the ones implied when the speaker is cooperative, and English is used).

Part of this work was performed while the authors were at the Speech Communications Research Laboratory, Inc., 800A Miramonte Drive, Santa Barbara, CA 93109.

This work was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-73-C-0221.

Data Base and Processing Methodology

The data base was 170 recorded interviews obtained from seventeen speakers (eleven males, six females). The ten successive sessions were a minimum of one week apart and typically two or three weeks apart. The interviewee was recorded in an IAC sound room, and the interviewer was outside the sound room. Microphones and headphones were used to maintain communications. The interviewer posed a predefined topic to the interviewee, and twelve to thirteen minutes of an extemporaneous monolog by the interviewee were recorded.

The total duration of the data base was approximately 36.8 hours. The data were band-limited to 3250 Hz and sampled at a 6500 Hz rate for compatibility with future applications to telephone systems and narrowband vocoder systems. A linear prediction analysis with a frame rate of 50 Hz was performed. The analysis was performed in real-time under Fortran control using an SPS-41 processing system in conjunction with a PDP 11/45 computer (5). The analysis parameters were used to calculate feature vectors based on the long-term average mean, standard deviation and dispersion (standard deviation divided by the mean) of the fundamental frequency, gain and ten reflection coefficients (6). Each feature vector was the average of L_V successive voiced frames.

A summary of the number of feature vectors produced for all 170 interviews is given in Table 1. The data are partitioned into representative test and reference sets (7). Four choices of L_V were studied, namely L_V=30, 100, 300 and 1000. The average real-time interval (RTI) per feature vector is also given. The real-time interval for a long-term feature (seconds/feature) corresponds to a product of the following factors: 1) the number of voiced frames per feature vector (L_V), 2) the reciprocal of the voiced frame to total frame ratio (or the reciprocal of the voicing duty factor), and 3) the reciprocal of the number of analysis frames per second (or the reciprocal of the frame rate).

Sess.	L_V=30	L_V=100	L_V=300	L_V=1000
1-5	58,379	17,486	5,799	1,712
6-10	58,032	17,736	5,764	1,701
Total	116,411	34,862	11,563	3,413
RTI	1.14s	3.80s	11.47s	38.85s

Table 1. Number of feature vectors and average real-time interval (RTI) for each L_V condition.

CH1285-6/78/0000-0287$00.75©1978IEEE

Experiments

Intra-Speaker Variability

Using this data base, it was possible to study the intra-speaker variability for a large number of speakers and for cumulative sessions. If individual sessions are described by S(i), i=1,10, then cumulative sessions may be described by C(i), i=1,10, where C(i)=S(1)+S(2)+...+S(i).

The standard deviations of the long-term averages of the fundamental frequency and the first reflection coefficient, denoted as $s<F_0>$ and $s<k_1>$ respectively, as measured over the cumulative sessions C(i) for one representative speaker, are shown in Figure 1.

For each set of cumulative sessions, $s<F_0>$ decreases as L_v increases. This behavior demonstrates that over long real-time intervals, a speaker's average fundamental frequency approaches an "habitual" value, and for successive long real-time intervals, the deviation from the habitual value is small. For short real-time intervals, influences such as speech prosody may mask the habitual value, and successive short real-time intervals will deviate more widely from each other.

The behavior of $s<k_1>$ as L_v increases mirrors the behavior of $s<F_0>$ as L_v increases. As more sessions are included, however, the behavior of $s<k_1>$ differs from the behavior of $s<F_0>$. There is essentially no measurable increase in k_1 variability as the time period increases. This trend is observed for the other speakers and the other long-term reflection coefficient averages. Since the reflection coefficients are used to describe the vocal tract shape in an acoustic tube model (9), the result implies that the physical characteristics of a subject's vocal tract show no observable changes over at least several months.

Variance Ratio Analysis

One method of measuring the usefulness of a feature for speaker identification is the F-ratio or variance ratio (10,11). The variance ratio of a feature is the quotient of the intra-speaker variance and the inter-speaker variance. In general, the larger the variance ratio for a particular feature, the greater the contribution of the feature in identifying speakers.

Trends as a function of population

The variance ratios for the case L_v=1000 and cumulative sessions 1-10 are shown in Figure 2 for the male and female speakers separately. Only the variance ratios of the mean and standard deviation features are shown. The variance ratios of the dispersion features were consistently much smaller.

There are noticable differences in the variance ratios between the male and female populations. Based on relative magnitudes, the features $<s(k_9)>$, $<s(k_8)>$ and $<k_1>$ would be the most significant for identifying the male population, while $<s(k_7)>$, $<s(k_8)>$ and $<k_8>$ would be the most significant for identifying the female population. If the male population is arbitrarily divided into two equal-sized subsets, the dominate features are $<k_1>$, $<F_0>$ and $<k_2>$ for the first subset and $<k_4>$, $<k_8>$ and $<k_3>$ for the second subset. These results show the need to have a substantially larger speaker population in order to characterize the parameters of major importance.

Trends as a function of L_v and time-spacing

The variance ratios were determined for L_v=100 and cumulative sessions 1-10 (Figure 3), and for L_v=100 and cumulative sessions 1-2 (Figure 4). Comparing Figures 2 and 3, the variance ratios generally maintain their same relative relationships, i.e. the larger variance ratios remain larger, and the smaller variance ratios remain smaller. However, the magnitudes of the variance ratios are smaller for L_v=100 than for L_v=1000. Comparing Figures 2 and 4, the relative relationships and the magnitudes of the variance ratios are similar for two cumulative sessions and for ten cumulative sessions.

Speaker Identification

Speaker identification was based on a weighted Euclidean distance metric, where the mean vector and inverse covariance matrix for each of the seventeen speaker classes were estimated from feature vectors in the specified reference set. The method of cross-validation in both directions was used, where independent subsets of the data were cyclically treated as test and reference groups, and the probabilities of correct identification for each cycle were averaged for the final results (7).

Performance as a function of L_v and time-spacing

Rosenberg (11) has noted that one of the most important considerations in designing a data base is the time period over which utterances are collected and the methods for establishing reference patterns over time. In this study, an experiment in speaker identification was designed using cumulative time-spaced reference sets. Reference sets were composed of from two to five successive sessions (with a time interval of at least one week between sessions). For each case, the reference and test sets were composed of equal numbers of successive independent sessions, and all four values of L_v were evaluated for speaker identification performance.

The results are presented in Table 2. For all L_v conditions, the performance was better as the number of cumulative sessions increased. For five sessions, the probability of correct identification P(CI) monotonically increased from 60.5% to 92.5% as L_v increased from 30 to 1000. L_v=1000 was not used for two sessions since there were insufficient data on a per-speaker basis for inverting the covariance matrices.

Performance as a function of feature subsets

It was noted that the dispersion features had very small variance ratios, whereas the mean features as a group consistently had the largest variance ratios. How would performance change if the dispersion features were omitted, or if only the mean features were included? The identification test

Number of sessions	2	3	4	5
Reference	1-2	1-3	1-4	1-5
Reference	3-4	4-6	5-8	6-10
L_V=30	50.36%	54.29%	59.91%	61.20%
	53.45%	57.04%	59.26%	59.87%
Average	51.91%	55.67%	59.59%	60.54%
L_V=100	64.34%	70.03%	76.41%	78.65%
	67.95%	72.69%	74.62%	75.48%
Average	66.15%	71.36%	75.52%	77.07%
L_V=300	71.18%	79.12%	86.73%	88.20%
	75.31%	82.14%	83.45%	85.27%
Average	73.25%	80.63%	85.09%	86.74%
L_V=1000		80.58%	92.85%	93.34%
	---	84.94%	89.60%	91.56%
Average		82.76%	91.23%	92.45%

Table 2. Speaker identification performance as a function of L_V and the number of sessions of reference data.

for L_V=1000 and five sessions per reference and test set was repeated using several different feature subsets, based on an analysis of the magnitudes of the variance ratios. In one case, only the twelve mean features were used, and in a second case, only the twenty-four mean and standard deviation features were used. The average performances for the two cases were P(CI)=93.6% and P(CI)=96.8% respectively.

Not only did both of these new cases based on feature subsets perform better than the original thirty-six dimension feature set, but in the second case, the performance was better by more than 4%. This result is a significant practical illustration that the inclusion of some parameters which would hopefully improve performance, or at worst case would have no effect on performance, can sometimes actually degrade the performance in an open test.

This improved performance by eliminating features with relatively small variance ratios was the basis for one additional test with a feature subset. In considering the remaining twenty-four features, the gain-related features had very small variance ratios. Therefore, the last identification test was repeated with the gain-related features removed. The performance of this new test was better than any previous test. The final results of this study are shown in Table 3.

Sessions		Speaker Identification
Ref.	Test	P(CI)
1-5	6-10	98.65%
6-10	1-5	97.45%
Average		98.05%

Table 3. Performance with fundamental frequency and reflection coefficient mean an^d standard deviation long-term features, L_V=1000 (average real-time interval = 38 seconds).

Summary

These results are extremely promising for future studies in many areas of speaker identification. One important extention would be to reprocess the "clean-text" audio tapes through various channel disturbances such as the telephone system to determine the robustness of the approach in less ideal environmental conditions. In conclusion, this substantially large testing effort has shown that realistic and acceptable speaker identification can be achieved with text-independent linguistically unconstrained speech. A future article will discuss in detail procedures and results for speaker identification, speaker verification and feature selection from this data base.

References

1) B.S. Atal, Effectiveness of linear prediction characteristics of the speech waves for automatic speaker identification and verification, The Journal of the Acoustical Society of America, vol. 55, pp. 1304-1312, 1974.

2) K.P. Li and G.W. Walker, Talker differences as they appear in correlation matrices of continuous speech spectra, The Journal of the Acoustical Society of America, vol. 55, pp. 833-837, 1974.

3) K.O. Mead, Identification of speakers from fundamental frequency contours in conversational speech, Joint Speech Research Unit, Report 1002, 1974.

4) M.R. Sambur, Speaker recognition using orthogonal linear prediction, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, pp. 283-289, 1976.

5) R.D. Arnott and J.D. Markel, Fortran control of real-time signal processing, submitted to IEEE Transactions on Acoustics, Speech, and Signal Processing, 1977.

6) J.D. Markel, B.T. Oshika and A.H. Gray, Jr., Long-term feature averaging for speaker recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-25, pp. 330-337, 1977.

7) L. Kanal, Patterns in Pattern Recognition: 1968-1974, IEEE Transactions on Information Theory, vol. IT-20, pp. 697-722, 1974.

8) A.H. Gray, Jr. and J.D. Markel, A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-22, pp. 207-217, 1974.

9) H. Wakita, Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms, IEEE Transactions on Audio and Electroacoustics, vol. AU-21, pp. 417-427, 1973.

10) P.D. Bricker, R. Gnanadesikan, M.V. Mathews, S. Pruzansky, P.A. Tukey, K.W. Wachter and J.L. Warner, Statistical Techniques for talker identification, Bell Systems Technical Journal, vol. 50, pp. 1427-1454, 1971.

11) A.E. Rosenberg, Automatic speaker verification: a review, Proceedings of the IEEE, vol. 64. pp. 475-487, 1976.

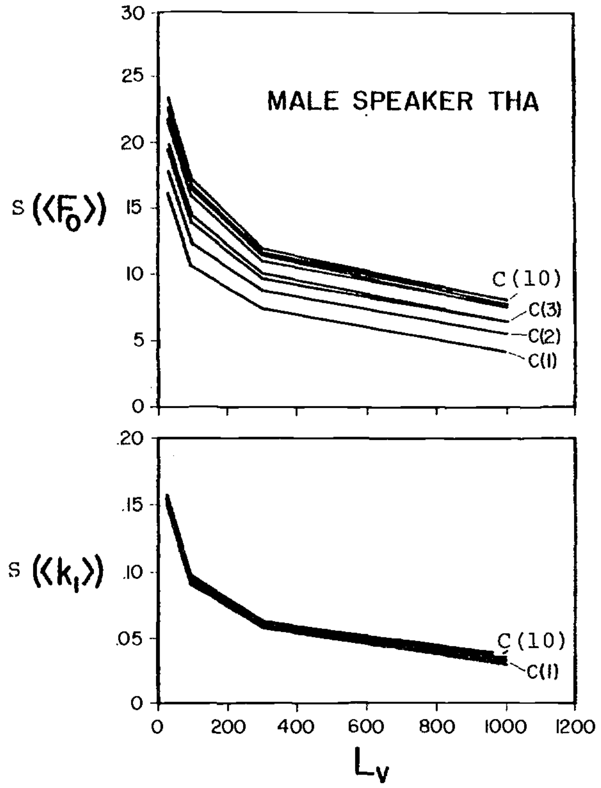

Fig. 1 Standard deviation of long-term features as a function of L_V.

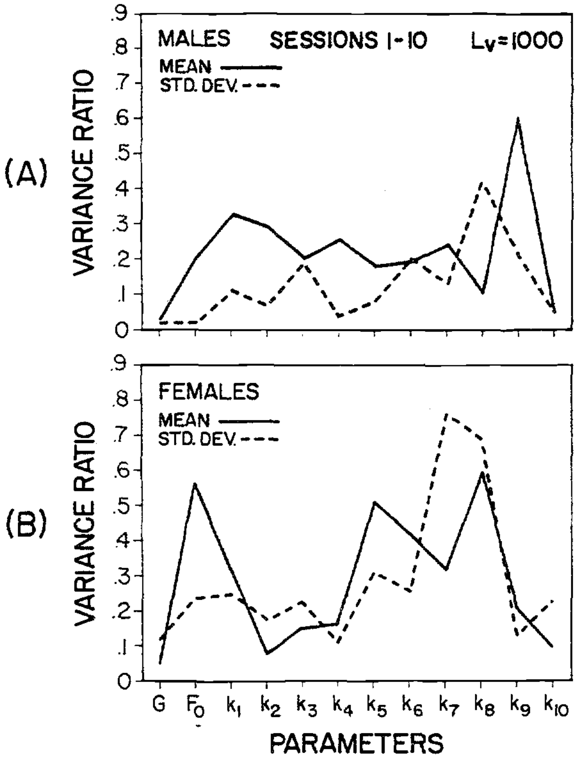

Fig. 2 Variance ratios from all 10 sessions - L_V = 1000.

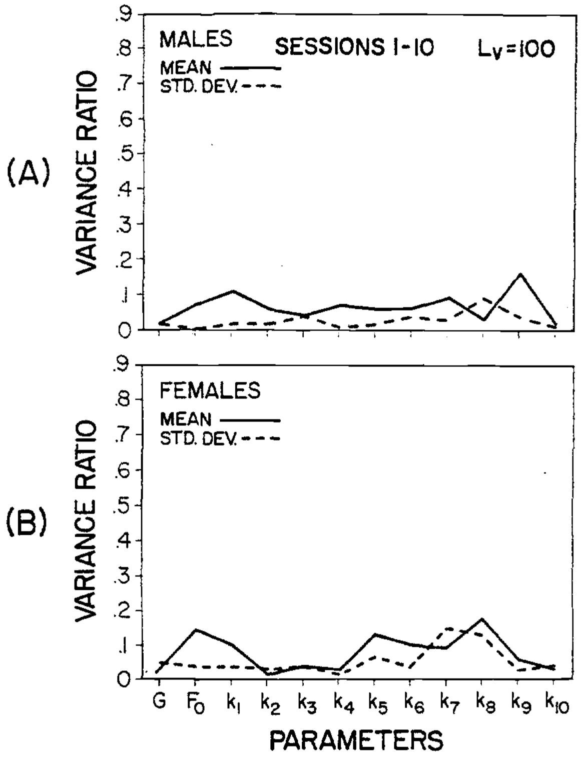

Fig. 3 Same as Fig. 2 except L_V=100.

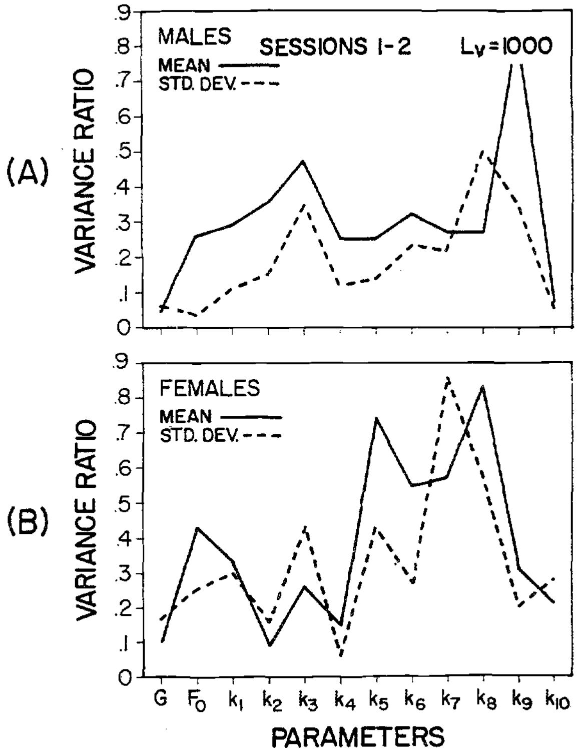

Fig. 4 Same as Fig. 2 except sessions 1-2.