FURTHER INVESTIGATION OF PROBABILISTIC METHODS FOR TEXT-INDEPENDENT SPEAKER IDENTIFICATION

J. Wolf, M. Krasner,
K. Karnofsky, R. Schwartz, S. Roucos

Bolt Beranek and Newman Inc.
Cambridge, MA 02238

ABSTRACT

In this paper, we present the preliminary performance of four methods for text-independent speaker identification using speech transmitted over radio channels. In a previous paper [1], we showed that for both laboratory-quality and simulated noisy-channel data in a single-session paradigm, new probabilistic classifiers yielded performance superior to that of a minimum distance classifier. We have recently compiled a speech database consisting of speech transmissions over a radio-channel. The lower quality and higher variability of this database differ markedly from the laboratory-quality databases often used in speech processing research. We present preliminary results with the same four methods of text-independent speaker identification using the radio-channel database with several experimental paradigms including multi-session paradigms. These results show that the probabilistic methods perform significantly better than a minimum-distance classifier for the multi-session paradigm.

INTRODUCTION

In our research on text-independent speaker identification, we have implemented four methods for modeling of training data from known classes (i.e., known speakers) and classification of test data from speakers with unknown identities. The training and test data consist of vectors that are points in a multidimensional space. The feature vectors currently used in the system are parametric representations of the short-time spectral envelope of the speech signal, although other information could be included, e.g., prosodic features. The four methods are Mahalanobis distance, Gaussian PDF estimation, Gaussian PDF estimation with "score clipping", and Non-Parametric PDF estimation [1].

The Mahalanobis distance method is a minimum distance classification technique. Each speaker is modeled by the mean vector and covariance matrix of the training data. Then, a minimum distance classifier based on the covariance-weighted distance from the mean of the test data to the mean of the training data of each class is used to classify test segments.

The probabilistic classification methods model each speaker by estimating a probability density function of the speaker's speech process from the training speech of the speaker. For each modeled speaker, the classifier computes the conditional joint probability of observing the feature vectors in the unknown test data under the assumption that the test data was spoken by that modeled speaker. Equal a priori probability of each possible speaker and independence of the test feature vectors are assumed. The classification is performed with the maximum a posteriori (MAP) probability (Bayes) classifier.

The Gaussian PDF estimation method models the distribution of each class as a multidimensional Gaussian PDF as estimated by the mean and covariance of the training data. A Gaussian distribution, however, is a poor model of a non-Gaussian process because the tails of the Gaussian distribution fall off very rapidly. The Gaussian PDF estimation with "score clipping" is a modification of the Gaussian PDF method designed to improve the robustness of the speech models. In this method, the tails of the Gaussian distribution are modified with a "soft-clipping" function. The modification is intended to reduce errors due to either (1) the underlying speech process being non-Gaussian or (2) the presence in the test data of feature vectors not generated by the modeled process, e.g., background noise during a pause in the speech. In addition, the modification mitigates the effect of underestimates of the process variance terms due to a small amount of training data. The Gaussian PDF methods are parametric modeling techniques: the shape of the distribution is fixed except for the mean and covariance parameters.

The last method tested, Non-Parametric PDF estimation, does not make assumptions about the overall shape of the distribution of each class of data. The Non-Parametric PDF method is based on a variation of the k-nearest neighbor (kNN) technique. The PDF for each known speaker is estimated at each test data point as a function of the distances from the test point to the k nearest training data points.

RADIO-CHANNEL DATABASE

We have recently compiled a new database using speech transmitted over radio channels. This database is distinctly different from our laboratory-quality database, containing speech that is often very noisy and distorted and, consequently, barely intelligible. Moreover, the noise and distortion levels vary widely over the database, with large differences between speech samples collected from a single speaker at times only a few minutes apart.

The source data consists of short radio transmissions during task-oriented conversations between speakers. Individual transmissions range in duration from about 0.5 seconds to 5 seconds, with an average duration of approximately 2 seconds. The speech data appears to have a slow roll-off below 500 Hz and a 10 to 15 dB notch at 2600 Hz. The data is additionally band-limited to a frequency range of 300 to 3200 Hz. (This bandwidth is roughly equivalent to the bandwidth of the laboratory-quality database.)

The principal degradations of the speech are time-varying environmental and channel additive noise and time-varying distortion that has some

ICASSP 83, BOSTON CH1841-6/83/0000-0551 $1.00 © 1983 IEEE

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on October 22, 2008 at 02:06 from IEEE Xplore. Restrictions apply.

correlation with speech loudness. We have calculated a dynamic range figure for each transmission. This figure is calculated as the ratio of the energy level that is greater than 95% of the frames within a transmission to the energy level that is greater than 5% of the frames within the transmission. This dynamic range figure varies from 4 to 40 dB over the database with an average of approximately 19 dB. For most transmissions, this dynamic range can be interpreted as the ratio of voiced speech energy to the energy of noise during speech pauses, silences before plosive bursts, etc. It should be noted, however, that the shorter transmissions may not contain any frames that do not contain speech. The 5% energy level, in that case, would reflect the weakest speech in the transmission, e.g., weak fricatives, nasals, etc.

An additional variation in the data is that of speaker state. The speakers vary from being calm and talking "normally" to being very excited and yelling. This variation often occurs in a set of transmissions collected from a single speaker within a period of approximately 30 minutes. To a human listener, a speaker may sound quite different over this set of transmissions; without the context from the source conversation, it is not obvious that different transmissions are indeed by the same speaker. This information, in the form of subjective observations, is noted and stored with the data during the compilation of the database.

The source speech data was collected for each speaker in several sessions, with the sessions separated in time by 1 to 20 days. Although sessions were designed with participants involved in conversations for similar tasks, the channel conditions were not constant. Changes in channel conditions include differences in radio equipment and variation in environmental acoustic noise and channel noise. Each session spans approximately 30 minutes and consists of many short transmissions. For compilation into an appropriate database, individual transmissions for each speaker were manually segmented and stored by speaker. The resultant database contains, at present, data for 19 speakers, each speaker represented by a minimum of 30 seconds of data on each of 4 separate sessions. Our multi-session experimental paradigms, described below, incorporate this information with experiments that train with data from one or more sessions and test with data from sessions not used in training.

Speaker identification is a much different and harder problem when using the low-quality radio-channel data. Relative to our laboratory-quality database, there are many differences worth summarizing. The greatest difference is that the noise and distortion levels are much higher and vary greatly even within a set of transmissions from a single speaker collected in a single session. In addition, the speaker's emotional state varies dramatically, causing a wide range in such objective features as pitch, speech rate, etc. The data from a single speaker on a single session spans a 30 minute period; in the laboratory-quality database, 30 seconds of data spanned less than one minute real time. Finally, in this database, we have available data collected in several sessions over a span of several weeks.

This database differs markedly from those often employed in speech processing research and, in particular, from the one used in our earlier research [1]. This data is of much lower signal quality and higher variability along several dimensions. The variability in channel conditions, in particular, suggests that a processing algorithm that estimates and provides compensation for the channel may result in superior performance. In the results reported below, however, we have not yet implemented such a strategy.

RESULTS

For all of the following experiments, the speech data was analyzed with a 20 ms Hamming window with an overlap of 10 ms between adjacent windows. The experiments were performed using LPC log-area-ratios (LARs) as the recognition features. In general, data for training or test consisted of speech from several transmissions since individual transmissions are of short duration as described above. A sub-population filter was used to automatically eliminate low-energy frames of data that are not generated by the speaker, e.g., frames that correspond to silences, etc. These eliminated frames, however, were counted in the duration of a data set.

The effect of the dimensionality (number of features in the feature vector) on the performance of classification rules is an important aspect of the design of pattern recognition systems. Given a fixed training set size, there is a trade-off between the information added by one more dimension and the loss in accuracy of the estimate of the joint density of the features [2]; that is, adding a feature can decrease performance. As the training set size increases, the joint density estimates become more accurate, increasing performance and allowing for the use of more features. Due to the lower quality of our radio-channel database, less information is conveyed by each feature, which should result in both lower performance and a decrease in the optimum number of features. The results presented below are consistent with these relationships.

The methods reported on and the symbols used in the figures are as follows:

- o Mahalanobis Distance (MD)
- □ Gaussian PDF estimation (GPDF)
- ⊞ Gaussian PDF estimation with "score clipping" (GPDF+C)
- △ Non-Parametric PDF estimation (NPDF)

Single-Session Paradigm: The first set of experimental results are for a single-session experimental paradigm. In this paradigm, each of 19 speakers is modeled using training data from a single session. Test data for a speaker is taken from the same session as the training data, but no data is used for both training and test. Performance for each method is plotted in Fig. 1 as a function of the number of features used in the recognition. Each speaker model was trained with 10 s of data; five 2 s segments of data were used for testing. For comparison, recognition performance on the laboratory-quality database with the same training and test durations [1], is shown in Fig. 2.

For the laboratory-quality data, the probabilistic classifiers perform significantly better than the MD minimum-distance classifier. In terms of absolute performance, all classification methods had poorer performance on the radio-channel data, reflecting the poorer quality of the data. For the radio-channel data, however, the MD classifier performs as well as the GPDF and GPDF+C parametric methods, whereas NPDF method does not perform as well.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on October 22, 2008 at 02:06 from IEEE Xplore. Restrictions apply.

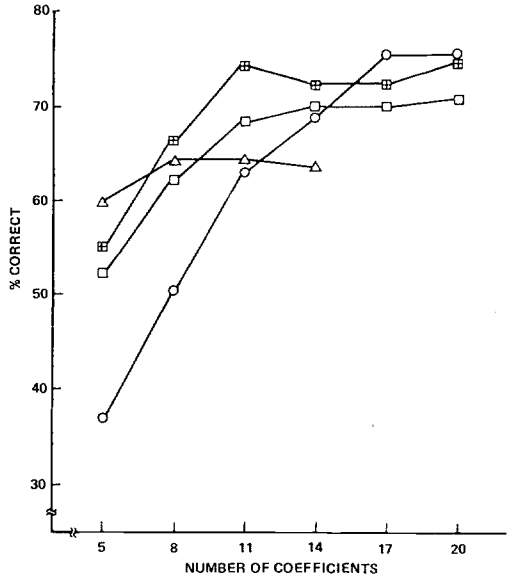

Figure 1: Performance on single-session paradigm for the radio-channel database.

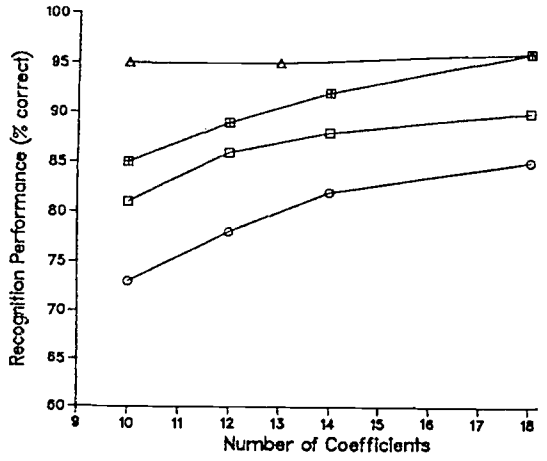

Figure 2: Performance on single-session paradigm for the laboratory-quality database (from [1]).

As was discussed above, the data even within a single session varied greatly with respect to noise and distortion levels. We previously reported [1] that a simulation of different channel conditions for test and training can cause significant degradation in performance, especially with the NPDF method. Our new results are consistent with this previously observed effect.

Multi-Session Paradigm: The other experimental results we report are for a multi-session paradigm. In this paradigm, each of 19 speakers is modeled using training data taken from three sessions. The test data for a speaker is then taken from a fourth session, a session that was not used for the training of that speaker. The use of multiple sessions, each session having data collected on a separate occasion, yields more variability for a speaker within the training data and from the training data to the test data than occurs with the single-session paradigm.

Performance for the multi-session paradigm is shown in Figs. 3 through 6. For all of the combinations of training duration, test duration, and number of features, the probabilistic methods performed significantly better than the MD minimum-distance classifier.

In Fig. 3, 10 s training with the radio-channel data is sufficient for use of 11 features by the three probabilistic methods. Maximum MD performance, however, is with 8 features. Figure 4 shows that with 20 s training, the number of effective features increases for each method. The improvement in performance for 20 s training compared to 10 s training is shown in Fig. 6. As discussed above, in the multi-session paradigm, performance decreases when too many features are used.

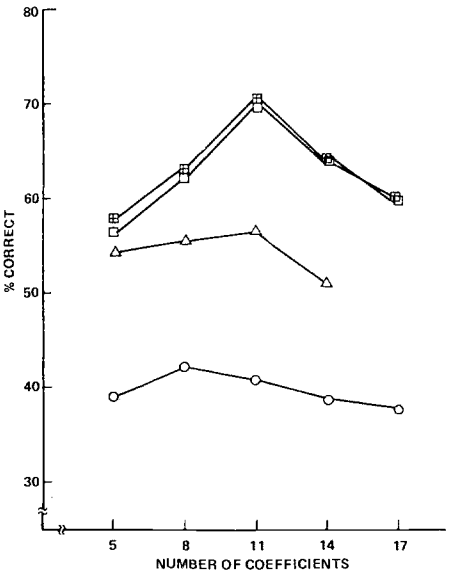

Figure 3: Performance for 10 s training, 2 s test on the multi-session paradigm for the radio-channel database.

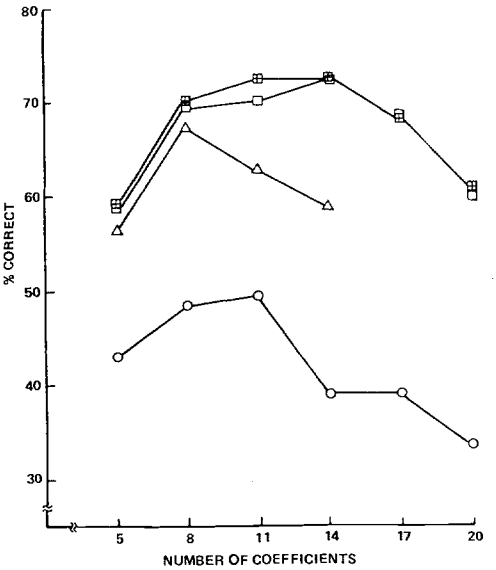

Figure 4: Performance for 20 s training, 2 s test on the multi-session paradigm for the radio-channel database.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on October 22, 2008 at 02:06 from IEEE Xplore. Restrictions apply.

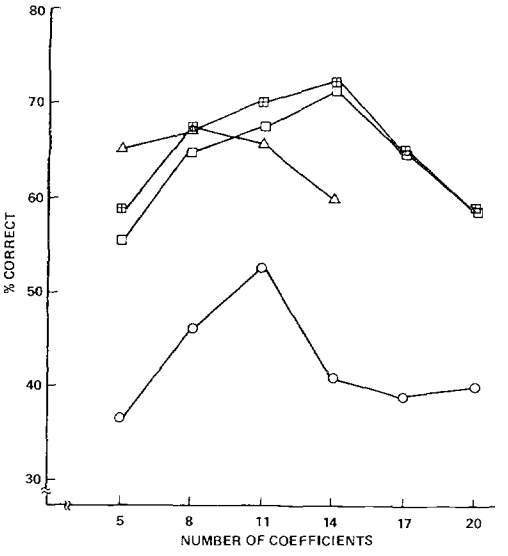

Figure 5: Performance for 20 s training, 4 s test on the multi-session paradigm for the radio-channel database.

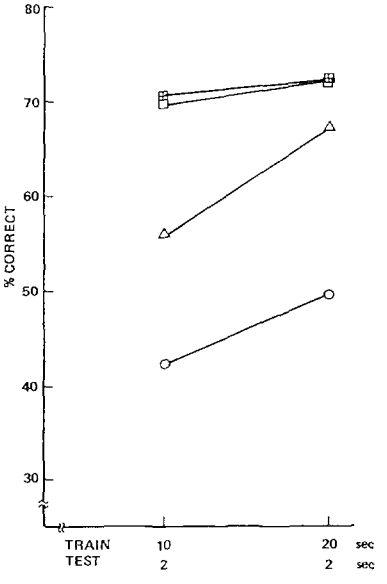

Figure 6: Comparison of 10 s and 20 s training on the multi-session paradigm for the radio-channel database (using best number of features).

Figure 5 shows the performance for 4 s test segments, using 20 s training. A comparison of Figs. 4 and 5 reveals that, averaged over the different test conditions and methods, the performance improved by only 1% when the test segment duration doubled. For these average performance levels, the expected improvement is about 7% (assuming the performance on successive 2 s test segments is independent). However, careful examination of the individual test results revealed that the performance on successive 2 s test segments is highly correlated, which leaves little room for improvement due to combining their scores into 4 s test segments.

The performance in the single-session and multi-session paradigms are compared in Fig. 7. It is important to note that while the performance of the MD minimum-distance method degrades significantly for the multi-session case, the performance of the probabilistic methods decreases only slightly.

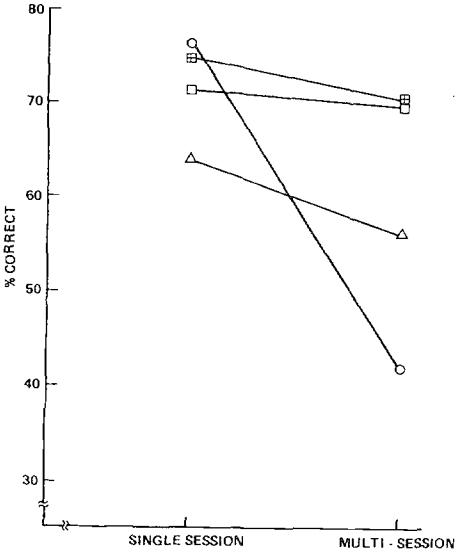

Figure 7: Comparison of performance on single-session and multi-session paradigms for the radio-channel database (using best number of features).

SUMMARY

In this paper, we have described our radio-channel database. The database, at present, contains data for 19 speakers, each speaker represented by a minimum of 30 s of data on each of 4 separate sessions. This database is distinctly different from a laboratory-quality database, containing speech that is often very noise and distorted and, consequently, barely intelligible. In addition, the noise, distortion, and speaker's emotional state vary dramatically even within the data collected in a single session.

We have presented some preliminary results for four methods of text-independent speaker identification using the radio-channel database for both single-session and multi-session paradigms. The results of the multi-session paradigm show that for all tested combinations of training duration, test duration, and number of features, the probabilistic methods performed significantly better than the minimum distance classifier.

REFERENCES

[1] R. Schwartz, S. Roucos, and M. Berouti, "The Application of Probability Density Estimation to Text-Independent Speaker Identification," Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Paris, France, May 1982, pp. 1649-1652.

[2] S. Roucos, "On Small Sample Performance of Pattern Recognition Machines," Ph.D. dissertation, University of Florida, 1980.

12.8

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on October 22, 2008 at 02:06 from IEEE Xplore. Restrictions apply.