INVESTIGATION OF TEXT-INDEPENDENT SPEAKER IDENTIFICATION OVER TELEPHONE CHANNELS

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

Bolt Beranek and Newman, Inc.
10 Moulton Street
Cambridge, MA 02238

1 Abstract

In this paper, we examine several methods for text-independent speaker identification of telephone speech with limited duration data. The issue addressed is the assessment of channel characteristics, especially linear aspects, and methods for improving speaker identification performance when the speaker to be identified is on a different telephone channel than that data used for training.

We show experimental evidence illustrating the cross-channel problem and also show that the direct approach, of using simple channel-invariant features, can discard much speaker dependent information. The methods we have found to be most effective rely on the training process to incorporate channel variability.

2 Introduction

In this paper, we consider the effect of channel variability on speaker recognition over telephone channels and methods for accounting for that variability. In this scenario, modeling of a speaker set is done with data from a variety of telephone channels and the identification is performed for an unknown speaker on an unknown channel.

We use experimental results to illustrate this problem and then illustrate the performance of a straightforward method of approaching the problem, i.e., using channel invariant features. This approach achieves channel invariance at the expense of poor performance.

The methods that we find most useful incorporate channel variation in the training process. One method, probabilistic channel modeling, models the channel as a Gaussian random vector and incorporates this information into the likelihood functions. The other method is a simple modification of the Gaussian method that calculates the covariance matrix in a manner that reflects the data variability. Both approaches make different assumptions, make different demands on the data and will be useful in different situations.

We will, for speaker identification, rely primarily on the Gaussian pdf (GPDF) classifier. Comparison of this and other classifiers such as Robust Gaussian pdf estimation, K-NN pdf estimation, and the Mahalonobis distance can be found in our previous papers [1, 2, 3].

In this research, we consider only the cepstral coefficient feature vectors. This is a natural choice when the channel is assumed to be a linear, time-invariant filter, since it results in the channel effect being one of adding a vector to the speaker feature vector.

As in our previous papers [2, 3] we have restricted ourselves to working with real signals and channels because we have found that simulations of aspects of the problem, while useful for illustrating certain points, invariably are inadequate for understanding the total problem.

3 Telephone Database

For this research, we have compiled a database composed of speech recorded over long-distance telephone channels. Telephone calls between Florida and South Carolina were placed to ten male speakers over a period of about two weeks. Each of the speakers is represented by four sessions, where each session corresponds to the data recorded from a single telephone call.

The speech data was carried over public telephone lines and is of typical long-distance toll quality, i.e., a bandlimited signal with low-level broadband noise. For digitization, the signals were filtered to a bandwidth of 300 Hz to 3300 Hz. We have observed the presence of a faint tone at 2600 Hz at a level 10 dB above the noise floor that appears to be an artifact of telephone signaling. On several of the calls, there is a low-level of environmental (background) noise present, i.e., the conversations of other people who happen to be in the same room as the speaker (telephone caller). On some of the calls, there is "pop and click" channel noise present. An examination of signal to noise ratios for this telephone data indicates that it is substantially cleaner than the radio channel data used as the basis of our earlier speaker identification work with real-world data [2, 3]. The average SNR of the telephone data is about 27 dB with a range of 19 to 33 dB. The most significant attribute of the telephone data, in terms of inter-session variability, is the presence of a significant linear channel effect. This variability can be seen graphically by comparison of the average spectra for a single speaker for each of several sessions. In addition, this variability is demonstrated by the substantial performance degradation from single-session to multi-session experiments described below.

4 Single-Channel/Multi-Channel Experimental Paradigm

The purpose of this experiment is to show the magnitude of the effect of the channel on speaker identification performance. For this experiment, the data for ten speakers is used. The single-channel experiment uses training data from each of the speakers from a single session; test material for a speaker is from a disjoint interval of the same session as used in training. The associated multi-channel experiment also uses training data from a single session; test material,

CH2118-8/85/0000-0379 $1.00 © 1985 IEEE

however, for each speaker is from a different session than the training and, therefore, from a different channel. Thus, speaker identification in this experiment is hindered by the inter-session variability of the telephone channel, as well as the inter-session variability of the speaker.

In addition to the use of cepstral features, the same experiment was performed using cepstral features that had been processed by the removal of the feature mean over each session. The removal of the feature mean removes an additive channel term, which, for the cepstral features, renders them invariant to linear channel effects. Unfortunately, the removal of the feature mean removes speaker mean information in addition to the channel term. This has a significant effect on the performance that can be achieved with these channel-invariant features.

The results, shown in Fig. 1, compare the single-channel and multi-channel experiments for the cepstral features and the channel-invariant features. Single-channel performance for cepstral features is 83%. Performance for the multi-channel experiment is 44%, an increase of the error by a factor of 3.

Performance for the channel-invariant cepstral features degraded by a much smaller amount from the single-channel to the multi-channel experiment: 56% down to 41%, an increase in the error by a factor of 1.35. Although this degradation is much smaller than the cepstral features, the performance on the single-channel experiment was only 56%, showing that the loss of the speaker mean information is significant to the speaker identification.

5 Channel Modeling

In the above experiment, we have limited the problem such that the substantial performance degradation is attributable to two factors: the variability in the channel from training to test and the inter-session variability of the speaker. In this section, we begin to develop appropriate compensation for the effects of the channel variability. Unfortunately, the channel variability is a composite of many effects including a linear filtering component, non-linear distortions, and additive and other noise effects. The channel modeling that we introduce is aimed at modeling only the linear filtering component of the inter-channel variability.

5.1 Probabilistic Channel Modeling

The idea behind probabilistic channel modeling is that since we observe the speaker to be identified only over an unknown telephone channel, speaker modeling should be done with observations from an ensemble of channels.

Below we will start with a representation for GPDF modeling and show that when there is a channel term that can be modeled by a Gaussian feature vector, we obtain a new representation of the likelihood function.

To this end, we consider cepstral feature vectors of size p, to have a Gaussian distribution with mean μ_j and covariance Σ_j for speaker j. In GPDF modeling it is μ_j and Σ_j that are estimated from the training data. If we observe N cepstral samples from unknown speaker m we can write these samples as:

$$x_{m,n} = y_{m,n} + \mu_m \qquad n=1,\ldots,N \tag{1}$$

where $y_{m,n}$ is a zero-mean Gaussian vector with covariance matrix Σ_m and μ_m is the mean component for speaker m.

Assuming that the observations are statistically independent the likelihood function for the N observations is given by:

$$L_j = \prod_{n=1}^{N} p(x_{m,n} ; \mu_j, \Sigma_j) \tag{2}$$

where p is a multivariate Gaussian density with mean μ_j and covariance Σ_j.

We can write log L_j, the log likelihood function for speaker j, in the form

$$\log L_j = \log h(\Sigma_j, S) + \log p(\bar{x} ; \mu_j, \Sigma_j/N) \tag{3}$$

where p is again a multivariate Gaussian density; $\bar{x}$ is the mean of the observations and h is a function that depends on the data only through the sample covariances. More explicitly,

$$\log h = -(N-1)/2 \log|\Sigma_j| - (N/2)\mathrm{tr}(\Sigma_j^{-1}S) + K \tag{4}$$

where $|\Sigma_j|$ is the determinant of Σ_j, tr denotes trace, and K is a constant that is the same for all L_j.

Also,

$$\log p = -\frac{1}{2}\log|\Sigma_j| - \frac{N}{2}(\bar{x}-\mu_j)' \Sigma_j^{-1} (\bar{x}-\mu_j) + K, \tag{5}$$

L_j, given above, is the GPDF classifier, and the modeling above assumed that there was no difference in channel between training and test.

When there is a significant time-invariant, channel difference between the training and test observations, it can be incorporated in the GPDF model by replacing μ_j by $\mu_j + c$, where c is the channel term. We now have for the likelihood

$$L_j(c) = h(\Sigma_j,S)\ p(\bar{x} ; \mu_j+c, \Sigma_j) \tag{6}$$

If we now assume that the channel has the multivariate Gaussian density $p_{ch}(c; \mu_c, \Sigma_c)$ with μ_c being the mean of all channels and Σ_c the channel covariance function. The classifier that minimizes the probability of misclassification is given by:

$$L_j^* = \int L_j(c) p_{ch}(c ; \mu_c, \Sigma_c)\, dc \tag{7}$$

It turns out that the above integral is a simple convolution and

$$L_j^* = h(\Sigma_j, S)p(\bar{x}; \mu_j+\mu_c, \Sigma_j/N+\Sigma_c) \tag{8}$$

where now p is a multivariate Gaussian density with mean $\mu_j+\mu_c$ and covariance

$$\Sigma_j/N + \Sigma_c \quad (9)$$

5.2 Estimation of Channel Parameters

We can readily see that application of probabilistic channel modeling requires data for estimating the channel mean and covariance. In our application, we have observations for use in training for each of the speakers on three different channels, i.e. three training sessions. We used the difference mean of the feature vector occurring on a particular channel for the particular speaker, and the average of these means for the particular speaker to generate a channel sample (of the probabilistic channel distribution), since these differences were attributable to the effects of the channel. The entire collection of such channel samples was then used in estimating the channel mean and covariance. This resulted in our obtaining three channel samples for each of ten speakers, giving 30 channel samples, only 20 of which were linearly independent.

Although we constrained ourselves to use only channel information obtained from the training data set, it is certainly possible to obtain it by other methods, e.g., obtaining channel samples from speakers not in the experimental group.

5.3 Speaker Covariance by Pooling the Data

An alternative to probabilistic channel modeling is to pool the entire training dataset available for each speaker in calculating the speaker mean and covariance and to use the standard Gaussian classifier, relying on the pooled dataset to represent channel variation. In fact, we can show that the pooled covariance will equal the average of the covariance of a speaker feature vector obtained on each of the channels plus the estimate of channel covariance, Σ_c, using only those channel samples obtained from the data for the particular speaker.

An advantage of this approach is its computational simplicity and its minimal assumption about channel statistics. However, the approach doesn't share the channel information from one speaker among the others.

6 Performance for Channel Modeling Method

For the experiments described in this section, ten speakers, each with data from four sessions, were used. A total of 20 seconds of training data for each speaker was taken from three sessions, an equal duration from each. Test data was taken from the fourth session that was not used in the training. Each test trial employed a 4 second test segment. The performance reported here represents 200 test trials. As described in the previous section, the Gaussian modeling and classification method is used with refinements for modeling of the telephone channels.

The Gaussian method without channel modeling serves as a baseline experiment for comparison. In this method, we observe that the data for each speaker for each session is, in general, from a separate telephone channel. In addition, the test segment is from a single (but different) telephone channel. To estimate the covariance of a speaker's single session data, first, the covariance for each training session is computed, and then for each speaker, the covariances for that speaker's three training sessions are combined into a single estimate. This covariance estimate is then used in the Gaussian model of the speaker. Performance in this baseline experiment was 68% correct speaker identification.

The Gaussian method with probabilistic channel modeling gave performance of 76% and represents an improvement of 8% over the baseline experiment performance.

For the Gaussian method using a pooled covariance, the covariance for each speaker was computed by pooling the data from a speaker's three training sessions. The performance for the pooled covariance method was 73%, an improvement of 5% over the baseline, and 3% less than Gaussian with probabilistic channel modeling.

Table 1 shows the performance obtained with each of the methods.

7 Discussion of Results

In compensating for the linear effect of inter-session channel variability, we refined the Gaussian modeling and classification method to reflect our knowledge of the channel. In the probabilistic channel modeling method, we make estimates of both the covariance of a single session's data for each speaker and the variability introduced by the channel. The effect of this method is reflected in both of the component terms that comprise the likelihood function in Equation 8. First, the component that is a function of the test data covariance employs a training covariance estimate of a single session. The component that is a function of the test data mean employs the training mean and accounts for the variability in that mean due to the channel. The performance improvement from these refinements is significant.

For the pooled-covariance method, we also note that the likelihood function component that is a function of the test mean tries to account for the variability due to the channel in a simple manner. The pooled covariance estimate, however, is not a good estimate of the test sample covariance. The performance of the pooled covariance method, as might be expected, is less than the probabilistic channel modeling case.

The gains that we have achieved are modest but significant, and we feel that additional channel data would further improve performance. Recall from Section 5.2 that only twenty independent channel samples were available for computation of the channel mean and covariance.

8 Summary

We have shown that the cross-channel problem for telephone speaker identification can result in serious performance degradations and have developed methods for improving performance. These methods have relied on incorporating the channel variability in the modeling and have achieved a degree of success in the limited set of experiments with limited data.

References

1. R. Schwartz, S. Roucos, and M. Berouti, "The Application of Probability Density Estimation to Text-Independent Speaker Identification", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, May 1982, pp. 1649-1652, Vol. 3.

2. J. Wolf, M. Krasner, K. Karnofsky, R. Schwartz, and S. Roucos, "Further Investigations of Probabilistic Methods for Text-Independent Speaker Identification", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, MA, April 1983, pp. 551-554, Vol. 2.

3. M. Krasner, J. Wolf, K. Karnofsky, R. Schwartz, S. Roucos, and H. Gish, "Investigation of Text-Independent Speaker Identification Techniques Under Conditions of Variable Data", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, March 1984, pp. 18B.5.1-18B.5.4, Vol. 2, Paper 18B.5.

Table 1 Performance of Baseline Experiment with Channel Modeling and Pooled Covariance Methods

	Percent Correct
Baseline	68
Pooled Covariance	73
Channel Modeling	76

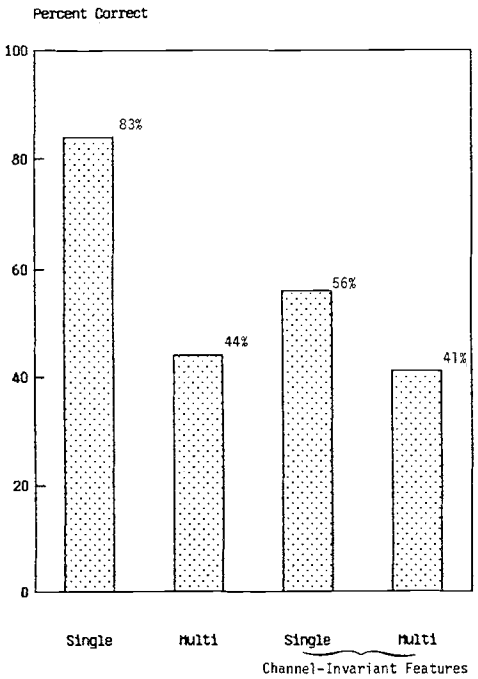

Fig. 1 Single-Channel and Multi-Channel Experiments