MODEL TRANSFORMATION FOR ROBUST SPEAKER RECOGNITION FROM TELEPHONE DATA

Françoise Beaufays and Mitch Weintraub

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA.
e-mail: francois,mw@speech.sri.com

ABSTRACT

In the context of automatic speaker recognition, we propose a model transformation technique that renders speaker models more robust to acoustic mismatches and to data scarcity by appropriately increasing their variances. We use a stereo database containing speech recorded simultaneously under different acoustic conditions to derive a *synthetic variance distribution*. This distribution is then used to modify the variances of other speaker models from other telephone databases.

The technique is illustrated with experiments conducted on a locally collected database and on the NIST'95 and '96 subsets of the Switchboard Corpus.

1. INTRODUCTION

Many applications of speaker identification systems (speaker-ID for short) assume that the users access the system remotely. Typically, the channel involved in the communication is that of the telephone. Because the handset and the line can vary from call to call, there is often an acoustic mismatch between the data collected to train the speaker models and the speech produced by the speakers at run-time or during testing. Such mismatches are known to severely affect the performance of the speaker-ID system [1]. In addition, the typically limited amount of training data further accentuates the problem.

The issue of acoustic mismatches can be tackeled at different levels: speech features can be extracted that are less sensitive to channel effects than the traditional cepstrum (see *e.g.*, [2, 13]) the effect of mismatches can be reduced via cepstral mean/bias removal (see *e.g.*, [3-6]), the speaker models can be transformed to compensate for the mismatches, rescoring techniques can be used to normalize the speaker scores and reduce the channel effects (see *e.g.*, [7]), etc. This paper concentrates on the model transformation approach.

The method we propose is a channel *compensation* method, as opposed to a channel *adaptation* one. It aims at making the speaker models more robust to channel mismatches rather than adapting them to the test environment. Adaptation algorithms used in speech recognition (*e.g.*, [8-11]) are not well-suited to speaker recognition: if the speaker models are adapted with the test data, they all converge towards the same model and the speaker discriminability is lost.

In this work, we don't assume any a priori knowledge about the communication channel. If such knowledge is available (for example if we know that the handset has an electret or a carbon-button), it can be used to map the speaker models from one environment to another (*e.g.*, using POF filters [12]) or to refine the variance transformation described here by making it telephone-dependent; but we show that even without such information a significant performance gain can be achieved.

The model transformation we describe uses an auxiliary database containing stereo recordings to compute what we refer to as a *synthetic variance distribution*, that is, a distribution of variances constructed artificially by comparing clusters of data points recorded simultaneously in different acoustic environments, and with a large amount of training data. This distribution is then used as a "target" to which the variances of other speaker models can be compared, and based on which they can be modified.

2. BASELINE SYSTEM

The front-end used in our baseline system extracts, from each frame of speech, a 17-dimensional filterbank-based cepstrum. The cepstra of the training data are used to build a set of Gaussian mixture models (GMM) trained with the EM algorithm. During testing, unknown speakers are recognized by a classifier that determines which trained model maximizes the log-likelihood of the speaker's test utterances. Cepstral mean substraction is applied to both training and testing utterances.

3. DEVELOPMENT DATABASE

Deriving a synthetic variance distribution requires a stereo database. Ideally, this should contain telephone speech recorded simultaneously through several telephones. Because this type of data is hard to collect, we used instead the Stero-ATIS database. This database, collected at SRI, consists of 10 series of 30 4-second long sentences that are recorded simultaneously with a close-talking Sennheiser microphone, and at the other end of one of 10 telephone units (by "telephone unit" we mean the combination of a handset and a line). The database contains the voices of 13 male speakers reading flight-related sentences.

4. COMPENSATION FOR ACOUSTIC MISMATCHES

The most severe channel mismatch situations occur when the training data is collected from only a few telephone units. If, instead, many telephone units are represented in the training data, chances are that the classifier will find a subset of the speaker model that fits well the test data. The variance transformation that we propose increases the acoustic coverage of the speaker models to make them more robust to unseen data.

This is illustrated conceptually in Fig. 1 for a two-dimensional feature space. If G1 is a cluster of features

collected from one telephone unit, the same speech frames transmitted by another unit might look like G2 or G3 or G4. Since our speaker-ID system uses GMMs, each cluster in Fig. 1 can be thought of as one Gaussian of one speaker model. The exact mean and variance changes from G1 to G2, G3, or G4 is generally unknown at the time of testing. Instead of trying to estimate these changes and use them in an adaptation-like algorithm (*i.e.*, replacing G1 with an estimate of G2, G3, or G4), we compute the variance of a cluster G that "covers" the possible regions where we may expect the data to lie when transmitted by different telephone units. The cluster G has the same mean as G1, but its variance is larger because it has to compensate for the mean shifts occurred between G1 and G2, G3, G4. The variances of the G clusters of all the speaker models form what we refer to as the *synthetic variance distribution*.

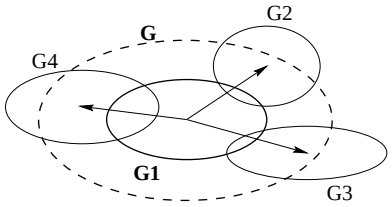

Figure 1. Clusters of data points in a two-dimensional feature space.

5. EFFECTS OF TRAINING DATA SCARCITY

In mismatched as in matched conditions, speaker-ID performance strongly depends on the amount of training data. More training data, even if it is mismatched with the test data, gives better performance and allows larger models to be built. This is illustrated in Fig. 2 where four matched and mismatched systems trained with increasing amounts of data are compared. The experiment was conducted on the Stereo-ATIS database described earlier. The speaker models were built with Sennheiser-recorded data. The test data consisted either of Sennheiser sentences (lower four curves), or of their telephone stereo recordings (upper four curves). The GMM sizes varied from 64 to 256 Gaussians.

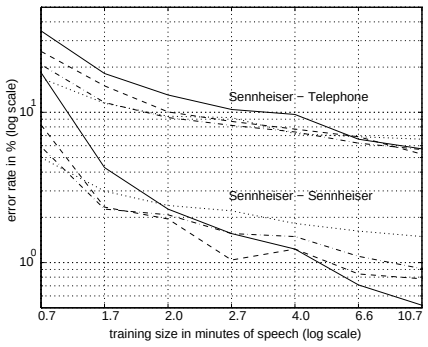

Figure 2. Speaker-ID error-rate as a function of the amount of training data. '...' = 64 G, '-.-.' = 128 G, '- - ' = 256 G, '—' = 512 G

The amount of data used to build a GMM and its number of Gaussians are directly reflected by the variance distribution of the Gaussians. For illustrative purposes, we plotted

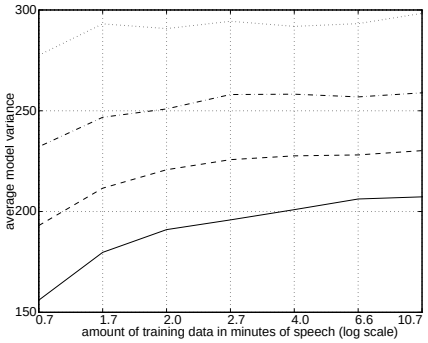

Figure 3. Average variance along c_2 vs. the amount of training data, for four GMMs. '...' = 64 G, '-.-.' = 128 G, '- - ' = 256 G, '—' = 512 G

in Fig. 3 the average variance along cepstral coefficient c_2 of the Gaussians built in the previous experiment. The figure shows that (1) for a given amount of data, the Gaussians of large GMMs have lower variances (each Gaussian models fewer data points); (2) for a given model size, the average variance increases with the amount of training data. This can be attributed to the fact that the EM algorithm tends to minimize the Gaussian variances, and that it can achieve this better when there are fewer data points per Gaussian.

These last observations suggest that increasing the variances of a GMM can also be useful in compensating for the lack of training data, and in allowing larger models to be built. We take this factor into account by constructing the synthetic variance distribution with a large amount of data.

6. SYNTHETIC VARIANCE DISTRIBUTION

The synthetic variance distribution is computed as illustrated in Fig. 1, using a stereo database. For sake of clarity, we will assume that this database is Stereo-ATIS. The Sennheiser utterances of Stereo-ATIS will be used to build the G1 clusters, and their telephone stereo recordings will be used to estimate the variances of the G clusters. The algorithm is summarized below.

1. Use a few Sennheiser sentences from each speaker to build a set of GMMs that will serve as frame classifiers.
2. For each speaker in the database:
   (a) Label each frame of the speaker's remaining Sennheiser data with the index of the Gaussian that maximizes its log-likelihood, that is, classify the Sennheiser frames using the speaker's GMM.
   (b) For each Gaussian in the GMM (for each cluster):
      i. Compute the mean, μ_S, and the variance, σ_S^2, of the Sennheiser frames clustered by this Gaussian.
      ii. Compute the variance, σ_T^2, of the stereo recordings of these frames. These stereo recordings comprise frames recorded on all 10 telephone units. To compensate for the shift in the means, the variance σ_T^2 is computed wrt. the mean μ_S of the Sennheiser frames instead of being computed wrt. its own mean, μ_T.

   The variances, σ_T^2, form the desired synthetic variance distribution.

We built a synthetic variance distribution for our baseline system. We kept 30 sentences per speaker to build a set of 64-Gaussian GMM classifiers, and used the other 270 sentences per speaker to derive the synthetic variance distribution. Figure 4 displays pairs of variances (σ_S^2,σ_T^2) computed along two different cepstral coefficients (the data points in each plot were normalized to have zero-mean and unit-variance). The figure shows that (1) as we expected, most of the synthetic telephone variances are larger than the corresponding Sennheiser variances (*i.e.*, most data points are above the diagonal), and (2) the variances along c_{17} show more dispersion than the variances along c_1. This is due to the fact that higher-order cepstral coefficients are more sensitive to channel effects.

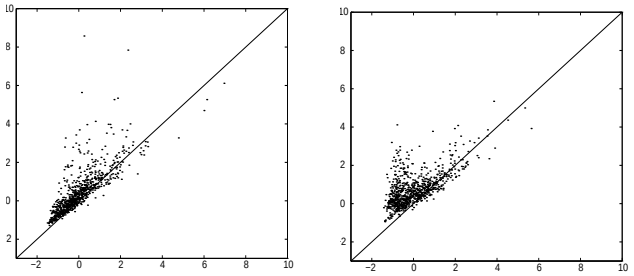

Figure 4. Pairs of normalized variances, σ_T^2 vs. σ_S^2, along c_1 (left) and c_{17} (right).

In first approximation, the points in each plot could be fitted with a straight line, thereby defining an *affine transformation* from the Sennheiser to the synthetic telephone variance distributions. Assuming that the acoustic coverage of Sennheiser data is similar to that of data collected from a telephone unit under similar conditions (same amount of data), we could apply this transformation to speaker models trained for other telephone databases. This approach is further described in [13]. It gave good results but was outperformed by the variance transformation described in the next section.

7. SPEAKER MODEL TRANSFORMATION

This transformation can be seen as an extension of the affine transformation mentioned previously. It translates the variances of the speaker models so as to make the mean of their distribution equal to the mean of the synthetic variance distribution. Mathematically, the transformation can be described as:

$$\begin{aligned} \sigma^2_{\text{tfmed},p,j}(i) &\triangleq \sigma^2_{p,j}(i) + t_i, \\ t_i &= < \sigma^2_{T,q,l}(i) > - < \sigma^2_{p,j}(i) >, \end{aligned}$$

where $\sigma^2_{p,j}(i)$ denotes the variance of the j^{th} Gaussian of the p^{th} speaker along the i^{th} cepstral coefficient, $< \sigma^2_{p,j}(i) >$ denotes the average of the variances along c_i of all the Gaussians of the speaker models to transform, and $< \sigma^2_{T,q,l}(i) >$ represents the average along c_i of the variances of the telephone synthetic distribution.

For a given front-end and for a given feature vector, the synthetic variance distribution is *fixed*. We therefore refer to this transformation as the *translation to a fixed target* of the model variances. This scheme allows speaker models that have small variances because they were trained with little data and/or with little acoustic variety are compensated more (t_i larger) than models that already have a good acoustic coverage.

For completeness, it should be noted that because the synthetic distribution is derived to compensate for the worst case mismatch (training data from a *single* acoustic environment), and because it is constructed with a lot of training data, it almost always has larger means than the distributions of variances computed under less favorable conditions. However, as a measure of security, we check that t_i is positive before applying it to the model variances. If t_i is negative, we set it to zero. We saw this happening only once in our experiments, for two higher-order cepstral coefficients.

8. EXPERIMENTS

8.1. Experiments on the SRI-digits Database

A first series of experiments was conducted on SRI-digits, a database of 10 male speakers reading series of digits over different telephone units. Six sets of GMMs were built with speech collected from six different telephone units. Each GMM had 64 Gaussians and was trained with one minute of speech. The GMMs were tested with 4-second long speech segments collected from 10 other telephone units. Three scenarios were considered: "no variance transformation", "affine transformation", and "translation to a fixed target". The speaker-ID error-rates were measured for the six systems and averaged. The experiment was repeated with 128- and 256-Gaussian GMMs. Table 1 summarizes the results.

	64 G	128 G	256 G
No transformation	**41.70**	42.73	44.17
Affine transformation	37.62	**37.49**	37.47
Translation to a fixed target	36.36	**35.79**	36.74

Table 1. Speaker-ID error-rate on SRI-digits, 1-line 1-minute training, 4-second testing.

The table shows that (1) the variance transformations made the models less sensitive to the number of Gaussians, but they favored larger models, and (2) the affine transformation gave a 10% relative improvement while the translation to a fixed target gave a 14% improvement.

The experiments were repeated with GMMs trained with data collected from two telephone units. Results are summarized in Table 2. Because of the larger amount of training data and because of the reduced mismatch, the error-rates are lower than in the previous experiment. The overall conclusions regarding the variance transformation remain the same. The improvements in error-rate are roughly 14% for both transformations.

	64 G	128 G	256 G
No transformation	**28.63**	28.73	35.72
Affine transformation	26.33	25.43	**24.74**
Translation to a fixed target	25.52	24.93	**24.66**

Table 2. Speaker-ID error-rate on SRI-digits, 2-line 2-minute training, 4-second testing.

8.2. Experiments on the Switchboard Corpus

So far, we considered only cepstrum-based systems. In [13], we propose a new speaker-ID feature that measures the slope of the filterbank used to derive the cepstrum. Cepstrum-based and filterbank slope-based GMMs can be combined by averaging the log-likelihoods of the test utterances wrt. the two GMMs (see [13] for more details).

Two sets of 64-Gaussian GMMs were built for the 30-second training, 5-second testing close-set task of the NIST'95 Evaluation (26-speaker subset of Switchboard).

One GMM was based on a 17-dimensional cepstrum, the other used a 28-dimensional filterbank slope feature [13]. Synthetic variance distributions were computed for each front-end, and fixed-target translations were applied to the corresponding GMMs. Table 3 summarizes the error-rates of the different systems. The variance transformation brought a 7% relative error-rate improvement to each classifier and improved the combination of the two classifiers by 11%.

cepstrum	var. transf.	ftbk slope	var. transf.	error-rate in %
√				24.89
		√		24.15
√	√			23.08
		√	√	22.44
√		√		23.40
√	√	√		22.33
√		√	√	22.54
√	√	√	√	20.83

Table 3. Close-set speaker-ID error-rates on the NIST'95 Evaluations subset of Switchboard (30-second training, 5-second testing)

A similar experiment was performed on hub "s1b" of the NIST'96 Evaluations. The training data consisted of one two-minute call. The test data consisted of one 10-second call. Roughly 50% of the handsets used for testing were identical to those used for training. This was an open-set speaker recognition task, with 21 male target speakers, and 400 imposter speakers. The figure of interest was the probability of false alarm at 10% miss (see [7] for more details about the open-set baseline system). Table 4 summarizes our results with and without variance transformation.

cepstrum	var. transf.	ftbk slope	var. transf.	% False alarm @ 10% miss
√				21.7
		√		19.8
√	√			20.3
		√	√	20.1
√		√		17.6
√	√	√		19.2
√		√	√	10.7
√	√	√	√	12.7

Table 4. % false alarm at 10% miss on the NIST'96 Evaluations subset of Switchboard (2-minute training, 10-second testing)

Table 4 shows that the variance transformation improved the cepstrum-based GMM performance by 9%. The filterbank-slope GMM was practically unaffected by the variance transformation. The system formed by combining both GMMs improved by 28% when variance transformation was applied to both features, by 40% when only the cepstrum-based GMM was transformed, but its performance decreased when only the filterbank slope-based GMM was transformed. We are currently investigating the reasons of this last result.

9. CONCLUSION

We have discussed the issue of speaker recognition over the telephone. We have proposed a variance transformation technique that renders Gaussian mixture models more robust to acoustic mismatches and to training with limited amounts of data. We have shown through several examples that the method improved significantly the speaker recognition error-rate of cepstrum-based systems. We are currently investigating the behavior of the variance transformation on the filterbank slope-based GMMs.

Several extensions of this technique can be investigated: (1) combining the transformations with handset detectors to make the transformations telephone-dependent, (2) make the transformations speaker-specific, (3) normalize the variance translation to eliminate differences between Stereo-ATIS and the application database that are due to factors other than the training conditions and the amount of training data, *e.g.*, the type of speech (read vs. spontaneous, constrained vs. unconstrained, etc.).

REFERENCES

[1] D.A. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Proc. Letters*, vol. 2, no. 3, pp. 46-48, March 1995.

[2] C.R. Janowski Jr., T.F. Quatieri, D.A. Reynolds, "Measuring fine structure in speech: Application to Speaker Identification," in *Proc. ICASSP-95*, pp. 325-328.

[3] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. ASSP*, vol. ASSP-29, pp. 254-272, April 1981.

[4] R.M. Stern, F.-H. Liu, P.J. Moreno, A. Acero, "Signal processing for robust speech recognition", *Proc. ICSLP-94*, vol. 3, pp. 1027-1030, Sept. 1994, Yokohama, Japan.

[5] F.-H. Liu, R.M. Stern, A. Acero, P.J. Moreno, "Environment normalization for robust speech recognition using direct cepstral comparison," *Proc. ICASSP-94*, vol. 2, pp. II/61-64, Adelaide, Australia.

[6] A.E. Rosenberg, C.-H. Lee, F.K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification", *1994 Intl. Conf. on Spoken Language Proc.*, Yokohama, Japan.

[7] L.P. Heck, H. Murthy, F. Beaufays, M. Weintraub, "Speaker Recognition at SRI International", NIST Speaker Recognition Workshop, Maritime Institute of Technology, Baltimore, Md. March 1996

[8] J.L. Gauvain, C.-H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 2, pp 291-298, April 1994.

[9] C.J. Legetter, P.C. Woodland, "Flexible Speaker Adaptation using Maximum Likelihood Linear Regression", *Proc. of the Spoken Language Systems Techn. Workshop*, pp. 110-115, 1995.

[10] V. Digalakis, D. Rtischev, L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures", *IEEE Trans. Speech and Audio Proc.*, vol. 3, no. 5, pp. 357-366, 1995.

[11] A. Sankar, C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. Speech and Audio Proc.*, pp. 190-202, May 1996.

[12] L.Neumeyer and M.Weintraub, "Probabilistic optimum filtering for robust speech recognition," *Proc. ICASSP-94*, vol. 1, pp. 417-420, April 1994.

[13] H.A. Murthy, F. Beaufays, L.P. Heck, M. Weintraub, "Robust Text-Independent Speaker Identification over Telephone Channels", in preparation - to be submitted to *IEEE Trans. Speech and Audio Proc.*.