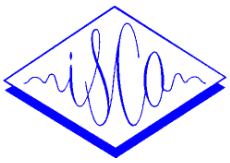

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A MODEL-BASED TRANSFORMATIONAL APPROACH TO ROBUST SPEAKER RECOGNITION

Remco Teunen, Ben Shahshahani, Larry Heck

{remco, ben, heck}@nuance.com
Nuance Communications, 1380 Willow Rd, Menlo Park, CA 94025, USA

ABSTRACT

A novel statistical modeling and compensation method for robust speaker recognition is presented. The method specifically addresses the degradation in speaker verification performance due to the mismatch in channels (e.g., telephone handsets) between enrollment and testing sessions. In mismatched conditions, the new approach uses speaker-independent channel transformations to synthesize a speaker model that corresponds to the channel of the testing session. Effectively verification is always performed in matched channel conditions. Results on the 1998 NIST Speaker Recognition Evaluation corpus show that the new approach yields performance that matches the best reported results. Specifically, our approach yields similar improvements (19.9% reduction in EER compared to CMN alone) as the HNORM score-based compensation method, but with a fraction of the training time.

1. INTRODUCTION

A major source of classification errors in speaker recognition systems is the distortion of the signal caused by the microphone. Mismatched microphones during the enrollment and test phases have been shown to cause up to an order of magnitude increase in the recognition error rate, even after standard channel compensation techniques are applied [1], [2]. This issue causes a significant barrier to successful deployment of the speaker recognition technology.

Channel compensation techniques can be classified into three broad categories: feature-based methods, model-based methods and score-based methods. Combinations of these techniques have also been studied in the literature.

Feature-based techniques attempt to reduce the signal variations due to channel differences at the signal processing and feature extraction stages. Cepstral mean normalization (CMN) [3], and RASTA [4] are among the most well-known channel normalization techniques used in speech and speaker recognition fields. More recent feature-based normalization techniques include the nonlinear power spectrum normalization technique [5], in which the distortion of the signal is approximated by a nonlinear function whose parameters are estimated by minimizing the mean squared spectral magnitude error. In [6], a neural network in used to obtain features that are discriminantly trained to achieve maximum classification accuracy. A survey of other feature-based techniques can be found in [7].

Model-based techniques attempt to reduce the effect of channel-variations by enhancing the speaker and background distribution models. Examples of model-based techniques include [8], in which affine transformations for variances of Gaussian models are trained off-line using stereo recordings. In [2] handset specific background models are used for score normalization and a handset detector is trained to identify the handset type of the incoming call.

Score-based normalization techniques such as ZNORM and HNORM [9] have been very successful in the recent NIST evaluations. These methods typically apply a speaker- and handset-dependent transformation to the final likelihood ratio scores. By doing this, the distribution of the impostor scores in all the channels are normalized to have zero mean and unit variance.

In this paper we present a new model-based channel compensation method. The method uses speaker-independent channel transformations to synthesize speaker models for channels for which no speaker data were available during the enrollment phase. The channel transformation parameters are estimated off-line without requiring stereo data. Upon receiving a test utterance, first a handset detector is used to identify the caller's handset. If the detected handset is different from the claimant's enrollment handset, then the appropriate transformation is applied to synthesize a claimant model for the detected handset type. Likelihood scoring and normalization is performed using the new synthesized model and the corresponding background model. In Section 2 we describe our algorithm in detail. Base systems used for comparison are described in Section 3. In Section 4 experiments on the NIST 1998 data are presented and the results are discussed in Section 4. Section 5 contains our concluding remarks.

2. SPEAKER MODEL SYNTHESIS

Our algorithm is best suited for speaker recognition systems that use a likelihood ratio detector approach. The verification score of an utterance is obtained by computing the average log-likelihood ratio as follows:

$$\Lambda(X|s) = \frac{1}{T}\sum_{t=1}^{T}\left(\log p(\underline{x}_t|\lambda_s) - \log p(\underline{x}_t|\lambda)\right), \qquad (1)$$

where $X = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_T\}$ denotes the set of feature vectors extracted from the utterance by the feature extraction front-end, λ_s is the speaker model (corresponding to the speaker that the caller claims to be), and λ is the background model used for normalizing the likelihood scores. Probability density functions of

both speaker and background models are modeled as Gaussian Mixture Models (GMMs) as follows:

$$p(\underline{x}_t|\lambda) = \sum_{i=1}^{L} w_i p(\underline{x}_t|b_i), \quad (2)$$

where b_i are multi-dimensional Gaussian densities, and w_i the corresponding weights. Each Gaussian is represented by a mean $\underline{\mu}$, and a variance $\underline{\sigma}^2$.

The background models are channel- and gender-dependent. Previous work has shown that this gives improved performance over channel-independent background models for channel mismatched conditions [2]. Each background model is derived from a channel- and gender-independent root model (denoted by λ_r) using Bayesian adaptation. The adapted parameters $(w_i, \underline{\mu}_i, \underline{\sigma}_i^2)$ for mixture component *i* are computed as follows:

$$w_i = \alpha\frac{n_i}{N} + (1-\alpha)w_{r,i}, \quad (3)$$

$$\underline{\mu}_i = \alpha\left(\frac{1}{n_i}\sum_{j=1}^{N} Pr(i|\underline{x}_j, \lambda_r)\underline{x}_j\right) + (1-\alpha)\underline{\mu}_{r,i} \quad (4)$$

and

$$\underline{\sigma}_i^2 = \alpha\left(\frac{1}{n_i}\sum_{j=1}^{N} Pr(i|\underline{x}_j, \lambda_r)\underline{x}_j^2\right) + (1-\alpha)(\underline{\sigma}_{r,i}^2 + \underline{\mu}_{r,i}^2) - \underline{\mu}_i^2 \quad (5)$$

where

$$n_i = \sum_{j=1}^{N} Pr(i|\underline{x}_j, \lambda_r) \quad (6)$$

and $\underline{x}_j$, $j = 1, \ldots, N$ are training samples from the new channel/gender, and $(w_{r,i}, \underline{\mu}_{r,i}, \underline{\sigma}_{r,i}^2)$ are the parameters of the root model λ_r. The parameter α is the smoothing factor.

Channel-specific speaker models are also obtained by Bayesian adaptation. During the enrollment phase, the gender and channel of the speaker are first detected and then the corresponding channel- and gender-dependent background model is adapted using the speaker's enrollment data.

Our technique works best if during the adaptation process the root model is used for estimating the posteriors while the channel-dependent background model is used for smoothing. By doing this, the correspondence between individual Gaussians of the speaker model and the background models are better preserved.

We denote by $T_{ab}(.)$ a transformation from channel *a* to channel *b*. Using the channel-dependent background models we obtain a transformation for each pair of channels as follows:

$$T_{ab}(w_i) = w_i\left(\frac{w_{b,i}}{w_{a,i}}\right), \quad (7)$$

$$T_{ab}(\underline{\mu}_i) = \underline{\mu}_i + (\underline{\mu}_{b,i} - \underline{\mu}_{a,i}) \quad (8)$$

and

$$T_{ab}(\underline{\sigma}_i^2) = \sigma_i^2\left(\frac{\underline{\sigma}_{b,i}^2}{\underline{\sigma}_{a,i}^2}\right). \quad (9)$$

where $(w_{a,i}, \underline{\mu}_{a,i}, \underline{\sigma}_{a,i}^2)$ and $(w_{b,i}, \underline{\mu}_{b,i}, \underline{\sigma}_{b,i}^2)$ are the parameters of the *i*th mixture components in the background models corresponding to channels *a* and *b*, respectively. Notice that the correspondence between the Gaussians is valid, because models for both channels *a* and *b* were adapted from the same root model. Also notice that the above transformation ensures that $T_{ab}(\lambda_a) = \lambda_b$.

During the testing phase, the channel-dependent background models are first used to detect the handset type of the incoming call. If the speaker model that was trained during enrollment matches the detected handset type, then the speaker model and the corresponding channel-dependent background model are used in Equation (1).

However, if the detected handset does not correspond with the enrolled speaker model, then a new speaker model is synthesized for the new channel by applying the appropriate transformation between the enrollment channel and the detected channel. Log-likelihood ratio scoring is then applied using the synthesized speaker model and the corresponding channel-dependent background model. Notice that the Bayesian adaptation process used during the enrollment phase ensures that the correspondence between the Gaussians in the speaker model and the background models are valid.

Figure 1 shows an overview of the speaker model synthesis method.

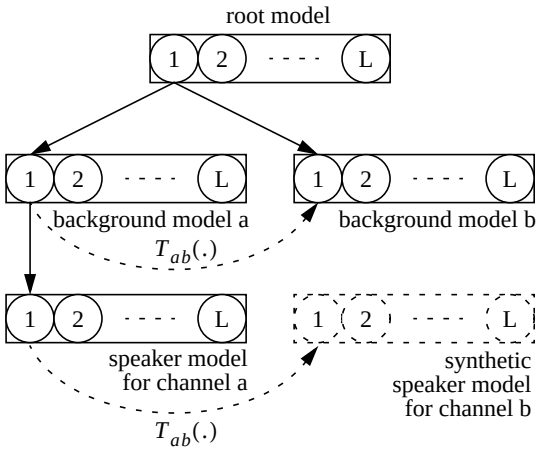

Figure 1: Overview of the speaker model synthesis method. The solid arrows denote Bayesian adaptation, and the dashed line model synthesis. The speaker-independent channel transformations are based on the background models and applied to the channel-dependent speaker models to synthesize speaker models for other channels.

3. Baseline systems

The speaker model synthesis method is compared against three different baseline systems:

1. **Universal background model (UBM)**. A single, speaker-independent background model is used for score normalization [9].

2. **Universal background model with HNORM channel compensation (UBM + HNORM)**. The same system as system 1, but the scores are normalized using HNORM [9]. For each speaker 5.6 hours of speech from the NIST 1996 corpus are used to train the HNORM parameters.

3. **Channel- and gender-dependent background models (HD + GD)**. Test utterances are compared against the enrolled speaker model and the corresponding background model. For channel mismatched conditions, the channel of the background model matches the channel of the speaker model, not the channel of the test utterance [2].

All baseline systems are GMM-based. All GMMs use 1024 Gaussians per mixture. Background models are trained using at least 6 hours of speech from 86 different speakers. The front-end and the training procedure are the same for all systems. The feature vectors consist of 14 mel-cepstrum coefficients, including energy, and their first and second order derivatives. The mel-cepstrum coefficients are computed from a sliding 25 ms frame of speech, with a frame shift of 10 ms, and they are normalized with respect to the mean of the utterance (CMN).

4. Experiments

The speaker model synthesis (SMS) method was evaluated on the 1998 NIST Speaker Recognition Evaluation corpus [10]. This corpus was chosen because of its high number of channel mismatched tests. The evaluation is focused on speaker detection, where the task is to determine whether a specified target speaker is speaking during a given speech segment. This task is posed in the context of conversational telephone speech with limited training data.

The test corpus has 500 speakers, balanced for gender. All speakers serve as both target speakers and as impostor speakers. The training data for each speaker consist of two minutes of speech taken from a single conversation ("1-session training" condition).

Performance was computed on test segments that have a duration of 10 seconds. The performance was computed separately for the following four test conditions:

1. Same phone number. The test segments are from the same phone number as the training data.

2. Different phone number. The test segments are from different phone numbers as the training data. This test contains both matched and mismatched channel conditions.

3. Training on carbon-button, testing on electret.

4. Training on electret, testing on carbon-button.

The total number of target speaker trials is 4,950, and the total number of impostor trials is 44,530.

5. Results

Figures 2 through 5 show the results of the four systems for the four test conditions. Figure 2 shows the results for the "same phone number" test condition; the telephone handset used for training and testing is the same. The performance of the four systems is very similar. For high false alarm rates the data is not very reliable due to the limited amount of data.

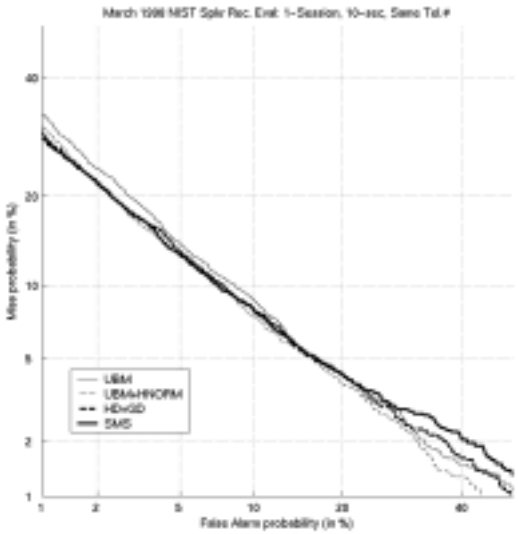

Figure 2: Performance of the four speaker recognition systems for the same phone test conditions. The handset is exactly the same between training and testing.

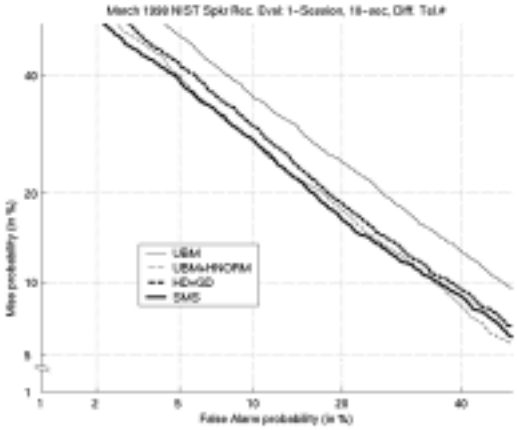

Figure 3: Performance curves for test condition 2, different testing and training phone numbers. The EER is best for the SMS system.

Figure 3 shows the results for test condition 2, the "different phone number" condition. This curve contains both matched and mismatched channel conditions. However, given that the performances for the matched conditions are very similar for the four systems (Figure 2), the differences are mainly due to the mis-

matched channel conditions. The performance of the SMS system is comparable to the performance of the UBM + HNORM system. Both these systems perform significantly better than the UBM system (19.9% at the equal error rate, EER). The HD + GD system also performs significantly better than the UBM system (15.6% at the EER), and only sightly worse than the UBM + HNORM system at the EER (1.9%).

In Figure 4, the performance for a specific mismatched condition is shown: training on carbon-button, testing on electret. The performance of the UBM + HNORM system is the best for this condition. The EER is 26% better than the EER of the SMS system. It should be noted that a very limited amount of data is available for this test (only 212 true speaker trials).

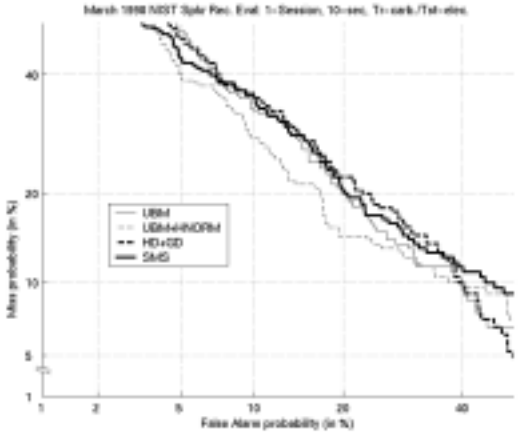

Figure 4: Performance curves for a specific mismatched condition: training on carbon-button, testing on electret. The curves are somewhat noisy due to the small number of tests.

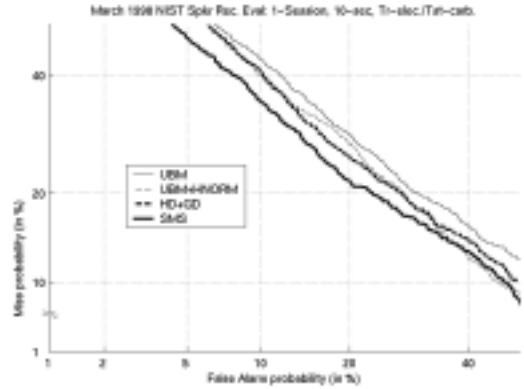

Figure 5: Performance for a specific mismatched condition: training on electret, testing on carbon-button.

The performance for the reverse test condition is shown in Figure 5: training on electret and testing on carbon-button handsets. The SMS system outperforms all other systems. The EER is 6.9% better than the EER of the HD + GD and UBM + HNORM systems, and 25% better compared to the UMB system.

For commercial systems this particular mismatched test condition is most relevant, since speakers typically enroll in the system using an electret handset type (e.g., office phone).

6. Conclusion

A novel statistical modeling and compensation method for robust speaker recognition was presented. In mismatched conditions, the new approach uses speaker-independent channel transformations to synthesize a speaker model that corresponds to the channel of the testing session. Effectively verification is always performed in matched channel conditions. Results on the 1998 NIST evaluation corpus show that the new approach yields performance that matches the best reported results. Specifically, our approach yields similar improvements as the HNORM score-based compensation method, but with a fraction of the training time.

The approach developed in this paper can be readily extended to perform on-line, unsupervised adaptation. This is presented in [11].

7. References

1. D.A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models”, *Speech Communications*, vol. 17, pp. 91-108, 1995.

2. L. Heck et. al, “Handset-Dependent background models for robust text-Independent speaker recognition”, *ICASSP*, 1997.

3. S. Furui, “Cepstral analysis technique for automatic speaker verification”, *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-29, pp. 254-272, 1981.

4. H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, “Compensation for the effects of the communication channel in auditory-like analysis of speech (RASTA-PLP)”, *EUROSPEECH*, pp. 1367-1370, 1991.

5. T.F. Quatieri, D.A. Reynolds, and G.C. O’Leary, “Magnitude-only estimation of handset nonlinearity with application to speaker recognition”, *ICASSP*, vol. 2, pp 1027-1030, 1994.

6. L. Heck et. al, “Robustness to telephone handset distortion in speaker recognition by discriminative feature design”, *Speech Communications*, 2000.

7. R.J. Mammone, X. Shang, and R.P. Ramachandran, “Robust speaker recognition”, *IEEE Signal Proc. Magazine*, vol. 13, pp. 58-71, 1996.

8. H.A. Murthy, F. Beaufays, L. Heck, and M. Weintraub, “Robust text-independent speaker verification”, *IEEE Trans. on Speech and Audio Proc.*, vol. 7, pp. 554-568, 1999.

9. D. Reynolds, “Comparison of background normalization methods for text-independent speaker verification”, *EUROSPEECH*, 1997.

10. NIST, “Speaker recognition workshop”, in *NIST Workshop Notebook*, linthicum Heights, Maryland, 1998.

11. L. Heck, and N. Mirghafori, “On-line unsupervised adaptation in speaker verification”, *ICSLP*, 2000.