Channel Robust Speaker Verification via Feature Mapping[•]

Douglas A. Reynolds
dar@ll.mit.edu

MIT Lincoln Laboratory, Lexington, MA USA

ABSTRACT

In speaker recognition applications, channel variability is a major cause of errors. Techniques in the feature, model and score domains have been applied to mitigate channel effects. In this paper we present a new feature mapping technique that maps feature vectors into a channel independent space. The feature mapping learns mapping parameters from a set of channel-dependent models derived from a channel-independent model via MAP adaptation. The technique is developed primarily for speaker verification, but can be applied for feature normalization in speech recognition applications. Results are presented on NIST landline and cellular telephone speech corpora where it is shown that feature mapping provides significant performance improvements over baseline systems and similar performance to Hnorm and Speaker-Model-Synthesis (SMS).

1. INTRODUCTION

One of the largest challenges in speaker recognition applications is dealing with channel variability. Typically a speaker will enroll his/her voice using one microphone or handset and then wish to be verified using a different microphone or handset. Since different microphones impose different characteristics on the acoustic signal, the spectrum-based features, pervasive in automatic speaker recognition systems, extracted for enrollment and verification will be different and hence result in a low match score. In addition to differing microphones, channel effects encompass other factors such as the acoustic environment (e.g., office, auto, etc.) and the transmission means (e.g., landline, cellular, VoIP, etc.). Since the speaker and channel information are bound together in the spectrum, anything that modifies the spectrum may cause difficulties.

Compensation techniques for channel effects have been applied in three domains. On the input side, feature domain compensation is aimed at removing the channel effects from the feature vectors prior to model training or verification. These include well-known and widely used techniques such as cepstral mean subtraction, RASTA filtering and spectral subtraction. On the output side, score domain compensation attempts to remove model score scales and shifts caused by varying input channel conditions. Examples of score domain compensation techniques are Hnorm [1] and Tnorm [2]. In model domain compensation the aim is to modify verification models to minimize the effects of varying channels. An example is Speaker Model Synthesis (SMS) [3], which learns how model parameters change between different channels and applies this transformation to synthesize speaker models under unseen enrollment conditions. Compensation in the different domains is of course not exclusive (indeed each seeks to remove different aspects of channel effects and so can have additive benefits) nor are all compensation techniques cleanly categorized into one of these domains. Of the three domains, feature domain compensation is perhaps the more general and widely useful since it is not tied to any particular model or score configuration. In this paper we present a new technique called *feature mapping* that extends the mapping idea from SMS to develop a more general feature domain channel compensation technique. The new technique is shown to be as effective as SMS for speaker verification on landline and cellular NIST speaker recognition corpora while also demonstrating a structure better suited for adaptation and as compensation for speech recognition systems.

In the next section we briefly describe the speaker verification system used throughout this paper. We next review the SMS approach and then describe the new feature mapping technique. This is followed by a description of the experiment data, design and results.

2. SPEAKER VERIFICATION SYSTEM

The speaker verification system discussed in this paper is shown in Figure 1 and fully described in [1]. In the front-end processing features are extracted from the speech signal and feature domain compensation is applied. In this work, the feature vector (extracted every 10 ms) is of 38 dimensions consisting of 19 mel-warped cepstra, derived from the frequency band 300-3300 Hz, and their first order derivatives, estimated with a 5-frame window. To compensate for linear channel effects (possibly time-varying), standard RASTA filtering is applied to the cepstra elements. The speaker and background models,

[•] This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government

which are used to form the likelihood ratio test statistic during verification, are both 2048 order Gaussian Mixture Models. The background model is typically trained using a 1-2 hours of speech from a large number of speakers over a variety of microphone/channel types. The speaker model is then derived from the background model using the available enrollment speech and one pass MAP estimation. For verification, the log likelihood of the input speech utterance is computed against both the background and speaker models, the difference taken and compared to a threshold to decide whether to accept or reject the putative speaker claim.

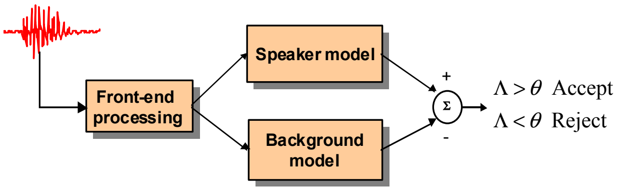

Figure 1 Structure of the speaker verification system in this paper.

3. SPEAKER MODEL SYNTHESIS

In [4] it was shown that better speaker verification performance can be obtained when the speaker and background model are channel matched, that is trained using speech from the same channel/microphone type. However, in many applications, it is unlikely to have user enrollment speech from all channel types that the user will use for later verification. Thus the motivation behind SMS is to synthesize a speaker model from "unseen" channels so that channel matched background scoring can be applied.

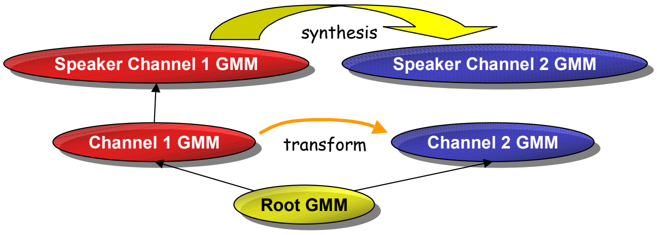

Figure 2 Speaker Model Synthesis (SMS).

This is accomplished as follows (see Figure 2). First a channel independent root GMM is trained using an aggregation of data from many different channels. Next, channel dependent GMMs are trained by using channel dependent data to adapt the channel independent root GMM. Since all models are derived from a common root, there is a correspondence between Gaussian components in the models. Transformations between the different channel dependent model parameters are then created by simply computing the mean shift, variance scale and weight scale to transform one channel dependent model into another The transformation of parameters from component i, $(\omega_i, \mu_i, \sigma_i)$ [1], from channel dependent model 1 to channel dependent model 2 would be

$$T_i^{CD1 \to CD2}(\omega_i) = \omega_i(\omega_i^{CD2} / \omega_i^{CD1})$$

$$T_i^{CD1 \to CD2}(\mu_i) = \mu_i + (\mu_i^{CD2} - \mu_i^{CD1})$$

$$T_i^{CD1 \to CD2}(\sigma_i) = \sigma_i(\sigma_i^{CD2} / \sigma_i^{CD1})$$

When a speaker enrolls, the most likely channel dependent background model is detected and adapted via MAP adaptation. Synthetic speaker models for the other channel dependent types are also generated[2]. During verification, again the most likely channel dependent background model is detected and the likelihood ratio with the corresponding channel dependent speaker model is computed and reported for the accept/reject decision. Note that since all models are derived via MAP adaptation from a single root model, a fast scoring technique [1] is available making this operation computationally inexpensive.

4. FEATURE MAPPING

While SMS focuses on synthesizing speaker models for unseen channels, the feature mapping approach, described in this section, focuses on mapping features from different channels into a common channel independent feature space. The two approaches are related in that they both learn transformations or mappings by examining how model parameters shift and scale after MAP adaptation. This new approach is motivated by several factors. First, a feature domain approach potentially has wider use since it is not tied to any particular recognition structure or model. Second, mapping features into a single space allows aggregation of information potentially obtained from several different channel types. For example, speech for enrollment or adaptation from several different channel types can be aggregated into a single features space before model building or updating. With SMS, separate speaker models must be maintained and an arbitrary selection of a common channel dependent model is made to combine model parameters.

Figure 3 shows the structure for the feature mapping system. As in SMS, a channel independent root GMM is trained using an aggregation of data from many different channels and channel dependent GMMs are trained by adapting the root GMM using channel dependent data. The model parameter changes between the channel independent and a channel dependent model indicate how the feature space distributions between the two spaces are

[1] mixture weight, mean and standard deviation

[2] Practically, synthetic models are generated on the fly during verification.

related and thus are used to create feature-mapping functions

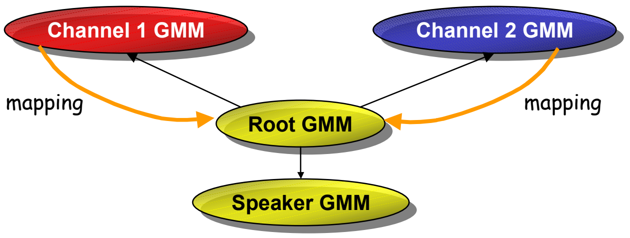

Figure 3 Feature Mapping

For simplicity, assume we are working with diagonal covariance GMMs; the full covariance case can be similarly derived. Let x be a feature from the space modeled by channel dependent 1 GMM (CD1) and $i = \arg\max_{1 \le j \le M} \omega_j^{CD1} p_j^{CD1}(x)$, where $p_j^{CD1}(x) = N(\mu_j^{CD1}, \sigma_j^{CD1})$ is the jth mixture component of the CD1 GMM. The mapping of x to a channel independent feature, y, is then given by

$$y = M_i^{CD1 \to CI}(x) = (x - \mu_i^{CD1}) \frac{\sigma_i^{CI}}{\sigma_i^{CD1}} + \mu_i^{CI}$$

The effect of this mapping is to transform $x \sim N(\mu_i^{CD1}, \sigma_i^{CD1})$ into $y \sim N(\mu_i^{CI}, \sigma_i^{CI})$. For a feature vector, the above mapping is applied to each vector element separately to create a mapped feature vector.

The feature mapper operates as follows. Given an input utterance, the most likely channel dependent model is first detected and then each feature vector in the utterance is mapped to the channel independent space based on its top-1 decoded Gaussian in the channel dependent GMM. The top-1 Gaussian decoding comes as a no cost by-product from the fast scoring technique used in computing the most likely channel dependent model. For multi-speaker speech verification cases, where the channel type may be changing within an utterance, the system can detect the channel type and map features over a short-term window of 1-2 seconds rather than the whole input utterance.

Although the mapping is independent of the follow-on recognition system, it is possible to couple the verification structure with the mapping models for greater efficiency. This is done during training by using the mapped features from enrollment speech for MAP adaptation of the channel independent root GMM. The system then uses the root GMM as a universal background model. During verification, the mapped features from the input speech are scored against the speaker and root GMM and the likelihood ratio score reported.

Note that both SMS and feature mapping are related in spirit to work on Stochastic Matching [5].

5. EXPERIMENTS

In this section we report on speaker detection experiments conducted on landline and cellular data from the NIST speaker recognition evaluations (SRE). The landline data is conversational telephone speech derived from the Switchboard-II phase-1 and phase-2 corpora[3]. The evaluation includes 457 male and 546 female speakers. For each speaker, approximately 2 minutes of speech extracted from a single telephone call is used for enrollment. Verification utterances are nominally 30 seconds in duration but vary between 0 and 60 seconds and come from phone numbers different than those used for enrollment. There are 3026 male and 3026 female verification utterances. Each verification utterance is scored against 11 putative speaker models with no cross-sex trials[4]. Additionally some results on the 2002 cellular corpus derived from Switchboard-II phase 4 are also presented. This corpus consists of 139 male and 191 female speakers with 2 minutes of training speech from a single telephone call and 1140 male verification utterances and 2119 female verification utterances of nominally 30 seconds duration.

Results are presented using Detection Error Tradeoff (DET) plots, which show the system tradeoff of misses versus false acceptances. Performance is computed by pooling all scores from male and female trials. Along with equal error rate (EER), the minimum decision cost function (DCF), defined as DCF = 0.1*Pr(miss) + 0.99*Pr(false_alarm), is also used as an overall performance measure.

In Figure 4 we show a DET comparing four systems run on the landline corpus. The baseline system uses a single 2048 background GMM trained using data from Switchboard-II phase 3 that is balanced for sex and handset type (carbon-button and electret). The handset labels are derived from an automatic system [6]. The baseline+Hnorm system is the baseline system with the standard Hnorm [1] score normalization applied. The SMS and feature mapping systems use the above background model as a channel independent root model and have four channel dependent models (male-electret, female-electret, male-carbon, female-carbon) trained using subsets of the complete root training data. It is clear from these results that (a) feature mapping provides significant improvement over the baseline system, (b) feature mapping provides the same performance improvement as Hnorm and (c) the feature mapping and SMS systems provide similar performance. The second point is significant since Hnorm

[3] The 2000 NIST SRE corpora is available from the LDC; see http://www.ldc.upenn.edu/Catalog/LDC2001S97.html

[4] The 2000 NIST SRE evaluation plan can be found at http://www.nist.gov/speech/tests/spk/2000/doc/spk-2000-plan-v1.0.html

requires computationally expensive normalization parameter estimation after enrollment.

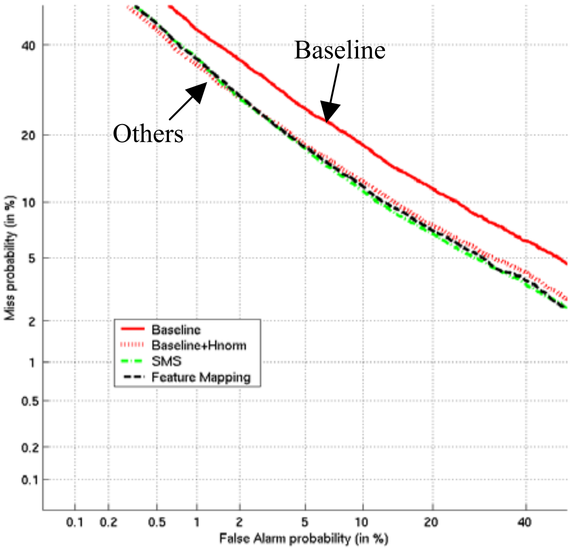

Figure 4 Landline Corpus: DET comparing baseline (red solid), baseline+Hnorm (red dots), SMS (green dash-dot) and feature mapping (black dashed).

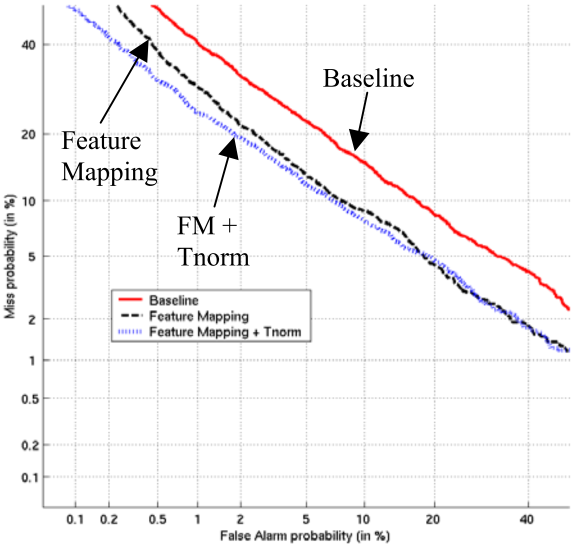

Figure 5 Cellular Corpus: DET comparing baseline (red solid), feature mapping (black dashed) and feature mapping+Tnorm (blue dots)

The feature mapping system was also applied to the cellular corpus by using cellular data to update the channel independent root model and adding six cellular channel dependent models (male and female GSM, analog and digital)[5]. The results are shown in Figure 5. Again we see the feature mapping produces significant gains over the baseline system using the same channel independent background model. The third DET on this plot shows the feature mapping system with Tnorm [2] applied, demonstrating that different domain normalizations can be effectively combined. A set of 100 male and 100 female speakers models from the 1997 SRE landline corpus was used for Tnorm. With Tnorm, the EER slightly reduces from 9.1% to 8.7%, while the minimum DCF value drops from 39x10^{-3} to 34x10^{-3}. Similar reductions were obtained when applying Tnorm to the landline corpus. Additionally, a form of Hnorm, called *Cnorm* when used with feature mapping, where we estimate and remove score bias and scales for each channel type, has been used, but was found to produce less of a performance gain than Tnorm and is more expensive to apply.

It is significant to note that for the 2002 NIST SRE, a single system using feature mapping was successfully applied to the landline, cellular and multimodal corpora tasks without background model retraining.

6. CONCLUSIONS

This paper has presented a new feature mapping approach for minimizing channel variability in the feature domain and demonstrated the performance advantage for both a landline and cellular telephone speaker detection task. The approach is a general compensation technique suitable for application to other speech recognition tasks. Future work will focus on applying it to a telephone speech recognition task and to other cross-channel speaker detection tasks. Additional development will examine an iterative approach to refining the mappings (as in [5]) and ways of adding new channels without updating the root model.

REFERENCES

[1] D. Reynolds, T. Quatieri and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” Digital Signal Processing, vol. 10, pp. 19-41, January 2000.
[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score Normalization for Text-Independent Speaker Verification Systems,” Digital Signal Processing, vol. 10, pp. 42-54, January 2000.
[3] R. Teunen, B. Shahshahani, and L. Heck, ``A Model-based Transformational Approach to Robust Speaker Recognition,'' ICSLP October 2000
[4] L. Heck and M. Weintraub, “Handset-dependent Background Models for Robust Text-Independent Speaker Recognition,” ICASSP April 1997.
[5] A. Sankar, and C-H. Lee, “A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition", IEEE Trans. on SAP, 1995
[6] D. Reynolds, “HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects,” ICASSP April 1997.

[5] The GSM data is from the 2001 NIST cellular corpus and the analog and digital data is from the OGI National Cellular Database (http://cslu.cse.ogi.edu/corpora/natcell/).