HANDSET-DEPENDENT BACKGROUND MODELS FOR ROBUST TEXT-INDEPENDENT SPEAKER RECOGNITION

Larry P. Heck and Mitchel Weintraub
Speech Technology and Research Laboratory
SRI International
Menlo Park, CA 94025

ABSTRACT

This paper studies the effects of handset distortion on telephone-based speaker recognition performance, resulting in the following observations: (1) the major factor in speaker recognition errors is whether the handset type (e.g., electret, carbon) is different across training and testing, not whether the telephone lines are mismatched, (2) the distribution of speaker recognition scores for true speakers is bimodal, with one mode dominated by matched handset tests and the other by mismatched handsets, (3) cohort-based normalization methods derive much of their performance gains from implicitly selecting cohorts trained with the same handset type as the claimant, and (4) utilizing a handset-dependent background model which is matched to the handset type of the claimant's training data sharpens and separates the true and false speaker score distributions. Results on the 1996 NIST Speaker Recognition Evaluation corpus show that using handset-matched background models reduces false acceptances (at a 10% miss rate) by more than 60% over previously reported (handset-independent) approaches.

1. INTRODUCTION

In telephone-based speaker recognition, it has been widely recognized that classification performance degrades because of corruptions of the signal in the transmission channel. Recently, however, research has been conducted suggesting that a significant part of the degradation might be attributed to a mismatch in handset types between training and testing (e.g., training on carbon button handsets only, but testing on electrets) [1]. Several well-established compensation techniques, including cepstral mean subtraction and delta coefficients as well as the newer RASTA filtering [2], have been applied to speaker recognition to compensate for channel and handset mismatches. While these methods can effectively compensate for linear channel distortions, they are generally less effective in treating the handset mismatch problem.

We present a new compensation method that significantly reduces the adverse effects of handset mismatch. The method compensates for handset effects by utilizing a new likelihood ratio scoring technique. Rather than comparing scores of a speaker whose identity is claimed (claimant) to scores of a set of "cohort" speakers, the method compares claimant scores with a speaker-independent "composite" model similar to the large random pooled model used in [3]. However, instead of using random speakers to build the composite model, we explicitly focus on the handset mismatch problem by training separate handset-dependent background models (speaker-independent, balanced gender). In this way, biases resulting from handset mismatches between the claimant and the normalizing impostor model are reduced. As compared to a state-of-the-art Gaussian mixture model (GMM) based text-independent speaker verification system using random pooled background speaker normalization without attention to handset types (baseline system), the new approach reduces the false alarm rate (at a 10% miss rate) by more than 60%.

2. BASELINE SYSTEM

The speaker recognition system utilizes EM-trained (expectation maximization) GMMs to represent the acoustic parameter distribution of each claimant speaker,

$$p(\vec{x_t} \mid \lambda_k) = \sum_{i=1}^{M} {p_i}^k {b_i}^k(\vec{x_t}), \tag{1}$$

where ${p_i}^k$ and ${b_i}^k$ are the mixture weight and the Gaussian density for the i-th mixture out of M for speaker k [1].

The acoustic parameters are 17th order mel-cepstra, with the zeroth-order term removed. The mel-cepstra are computed from a sliding 25 ms frame of speech, with a framerate of 10 ms. Cepstral mean subtraction is used in all experiments reported in this paper.

The average log-likelihood of a claimant speaker given an utterance $X = \{\vec{x_1} \ldots \vec{x_T}\}$ is computed as

$$\mathcal{L}(X \mid \lambda_k) = \frac{1}{T} \sum_{t=1}^{T} \log p(\vec{x_t} \mid \lambda_k). \tag{2}$$

For the baseline (open-set) system, a likelihood ratio detector is used that *normalizes* the score of the claimant speaker by the score of a single handset-independent composite model of all other impostor speakers. This method was described as "random pooled" in [3]. In the log-domain, the ratio can be expressed as a difference of terms,

$$\Lambda(X \mid k) = \mathcal{L}(X \mid \lambda_k) - \mathcal{L}(X \mid \overline{\lambda_k}) \tag{3}$$

where $\overline{\lambda_k}$ denotes the composite impostor model. The term composite refers to the fact that aspects of many persons' voices (90 speakers were used in the research reported

here) are combined into one model. This contrasts with cohort methods, where separate speaker-dependent models are used. The composite models are simple to implement, and have been shown to perform comparably with cohort-based methods [3]. For the experiments reported in this paper, we used 1280 Gaussian mixtures for the composite model.

3. EXPERIMENTAL DATABASE

The database we used in our experiments is the March 1996 NIST Speaker Recognition Evaluation. The database is a subset of Switchboard, a conversational-style corpus of long distance telephone calls. The subset consists of 40 claimant speakers (21 male and 19 female) and approximately 400 impostor speakers (200 male, 200 female). There are three training conditions for each claimant speaker: "one-session" (all training data from one phone call, i.e., one handset), "one-handset" (training data from two phone calls, but with one handset), and "two-handset" (training data from two different handsets). Each training condition uses 2 minutes of training speech from the claimant speaker. There are two testing conditions: "matched" and "mismatched" telephone numbers, referring to whether or not the telephone used during testing was the same as that used in training. In addition, there are three test utterance durations: 30, 10, and 3 seconds. The results of this paper are focused on the most difficult portion of the database: the "one-session" condition with males only for both the claimant and impostor speakers. Both test conditions over all three test durations are examined.

4. HANDSET-DEPENDENT COMPOSITES

To motivate the use of our new handset compensation technique, we begin with an analysis of the baseline system. Figure 2 shows histograms of scores for true and false speaker verification trials using the evaluation database described above. However, the test data were separated into two handset types: carbon and electret. In this way, we generated three plots: the top plot shows the score histograms for all of the data (all handsets), the middle figure plots the portion of scores from matched handset types (e.g., electret in both training and testing), and the bottom figure plots the scores from mismatched handset types (e.g., electret in training, carbon in testing). As can be seen, the true and false speakers are much more difficult to distinguish in the mismatched case. In addition, the plots show that the true speaker scores are largely bimodal, with the rightmost mode dominated by the matched handset type scores, and the leftmost mode dominated the mismatched scores. As a result, the overall true speaker distribution (top plot) is spread out, overlapping the false speaker distribution which results in a high number of false rejects and false accepts.

To reduce the number of errors in the verification system, we have developed an alternative method for score normalization. Instead of normalizing the claimant speaker score with a generic composite model trained over speakers using various handsets, we normalize with models that are specific to the handset type. That is, if a claimant speaker model was trained on a carbon button handset, then a carbon handset composite model is used to normalize the scores, whereas if the claimant model was trained on an electret, then an electret composite model is used. The use of handset-dependent models is motivated by the fact that models built from carbon handset data have shifted and scaled score distributions as compared to distributions from models built using electret handset data [1]. Our goal is to normalize variability due to handset effects by computing differences between scores from claimant models and composite models trained on the same handset type.

Implementing this approach, however, requires a handset detector to determine what type of handset was used in training (unless the training data is already marked with handset labels). We implemented a GMM for the two handset classes: carbon and electret. The GMM-based handset detector was trained and tested with an SRI speech corpus called Stereo ATIS. Stereo ATIS consists of approximately 10 hours of read sentences from the Air Travel Information System (ATIS) task. Each sentence was recorded by 13 male speakers in stereo over a telephone handset and a Sennheiser noise-canceling microphone. Ten different handsets were used, including seven electrets and three carbon button transducers. The database was split into two portions for training and testing, with all 10 handsets represented equally in both. Approximately 3.6 hours and 1.5 hours of speech data from electret and carbon handsets, respectively, were used to train the handset classifier. Approximately 5 hours of the speech data consisting of sentences of about 10 seconds in length were used to test the classifiers.

Table 1 shows the performance of the handset detector used to detect carbon handsets for 256 and 512 Gaussian mixtures. The false alarm rate indicates the percentage of times that a handset was incorrectly classified as carbon, while the miss rate indicates the percentage of times that a carbon handset was classified as another handset type (in these experiments, there are only two classes: carbon and electret). The performance leveled off after 512 mixtures, which is the number of mixtures used for handset detection in the experiments.

Number of Mixtures	False Alarm Rate (%)	Miss Rate (%)
256	0.4	1.8
512	0.2	1.1

Table 1. Performance of carbon microphone handset detector.

Using the handset detector, we labeled all the training data from the March 1996 NIST evaluation database (used for both the claimant and imposter modeling). In this initial implementation, we made hard class decisions, separating the data into carbon and electret classes. Finally, we encoded associations between claimant speaker models and their matched handset composite models. Table 2 shows the results of applying the handset-dependent composite (H-Composite) modeling technique to the "one-Session" training condition of the March 1996 NIST evaluation. The table compares the new method with the baseline system for the 30-,10-, and 3-second tests. Results for two operating points are shown: the equal error rate (EER) and the false alarm rate when the miss rate is 10%. As expected, the improvement for the mismatched telephone case (mismatch

30 Second Test Length, 2 Minutes Training				
Scoring Method	Mismatched Tel.#		Matched Tel.#	
	EER	Pfa($P_{m=10\%}$)	EER	Pfa($P_{m=10\%}$)
Baseline	20.0	41.6	6.0	2.3
H-Comp.	12.7	15.8	4.8	2.1
10 Second Test Length, 2 Minutes Training				
Baseline	21.6	39.8	7.2	4.3
H-Comp.	14.2	18.5	5.8	3.0
3 Second Test Length, 2 Minutes Training				
Baseline	24.6	51.8	9.9	10.5
H-Comp.	18.9	33.5	10.6	11.0

Table 2. Performance of composite (Baseline) and handset-dependent composite in EER, and the false alarm rate at a 10% miss rate. Results shown separately for matched tel. number (for both training and testing) and mismatched tel.

between training and testing) is much greater. The technique gives significant improvements except in the matched 3-second test case.

Table 3 compares the H-Composite method to cohort-based methods. Cohort normalization uses one or more speaker-dependent impostor models as the background model. The methods include "Cohort" (one cohort), 10 closest cohorts, and 10 maximally spread closest (10-msc) cohorts [1, 3]. The total number of Gaussians used to normalize each target speaker for the H-Composite, Composite and the 10 closest and maximally spread cohorts is the same (1280 Gaussians)[1]. All of the methods yield significant improvement over the unnormalized scores (no background model). For the matched telephone number, the H-Composite performs the best at EER, while the 10 msc performs the best at a 10% miss rate. For the mismatched telephone, the H-Composite method outperformed the cohort methods, giving a 19.0% reduction in false alarm rate (at a 10% miss). An interesting feature of the cohort methods was observed during the testing. For each target speaker, the closest cohort had the matched handset type as the target speaker. This suggests that the handset type has at least as significant of an impact on cohort selection as the speaker characteristics.

Scoring Method	Mismatched Tel.#		Matched Tel.#	
	EER	Pfa($P_{m=10\%}$)	EER	Pfa($P_{m=10\%}$)
Unnorm.	32.1	73.3	21.2	32.6
Cohort	22.1	48.9	11.2	12.2
Baseline	20.0	41.6	6.0	2.3
10-closest	14.8	19.4	5.6	1.8
10-msc	13.6	19.5	5.0	1.6
H-Comp.	12.7	15.8	4.8	2.1

Table 3. Comparisons for 30-second test of normalization methods of comparable complexity, including cohort (closest speaker), 10-closest speakers, and 10 maximally spread closest speakers.

To further study the effects of handset type, we used our handset labels on Switchboard to divide the "mismatched tel.#" category into two classes: matched handset type (but mismatched tel.#'s), and a mismatched handset type (e.g., trained on carbon, tested on electret). Table 4 shows verification performance for this additional split of Switchboard. When comparing training and testing data from the "matched" telephone numbers and from "mismatched" telephone numbers, the major factor in predicting performance is the type of the telephone handset (whether they were the same in training and testing or different), not the difference across lines nor within a handset type class (e.g., electret).

Figure 1 shows the corresponding 30-second performance curves of the handset-independent Composite (baseline) and new H-Composite modeling methods for matched and mismatched handset types. The plots show significant improvement for the mismatched case, and even improvement for the matched case at low and high miss rates. Figure 3 shows histograms of scores for true and false speaker verification trials for the new H-Composite method. As compared to the baseline histogram plots in Figure 2, the mismatched case is greatly improved by shifting the mean (Mt) of the target score distribution to the right, resulting in a sharpening of the overall distribution. In addition, the means (Mi) and standard deviations (Si) were changed, albeit less than the target distributions, illustrating the nonlinearly relation between the H-Composite normalization and the scores.

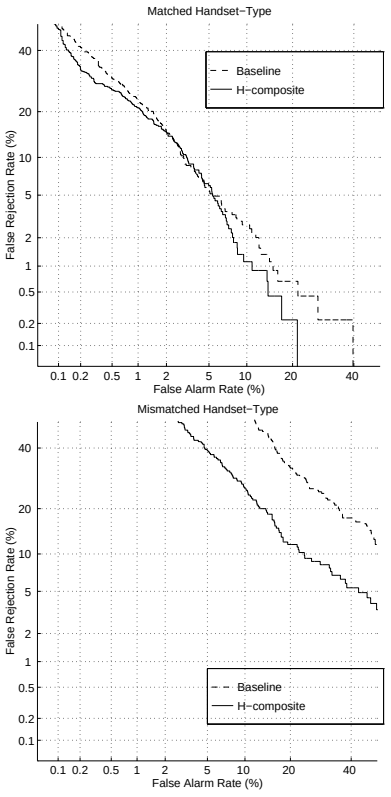

Figure 1. Comparison of baseline and handset-dependent composite modeling methods for 30-second test length with matched and mismatched handsets

[1] The cohort methods potentially require a unique set of cohorts for each claimant speaker, increasing the enrollment time and storage needs over the composite methods.

Condition	EER	Pfa(P_m=10%)
Matched tel.#	6.0	2.3
Mismatched tel.#, matched handset	5.4	4.9
Mismatched tel.#, mismatched handset	25.9	47.0

Table 4. Baseline performance for the matched telephone in both training and testing, matched handset type (but mismatched telephones), and a mismatched handset type (e.g., trained on carbon, tested on electret).

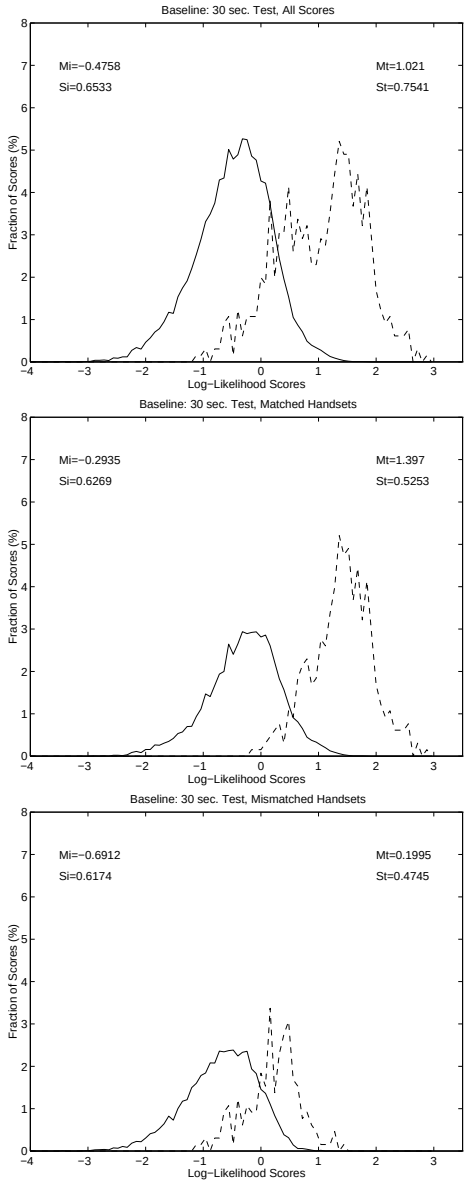

Figure 2. Handset-indep. composite (baseline) models: histograms of scores for true (right) and false (left) speaker trials with for all (top), matched (middle), and mismatched (bottom) handset types.

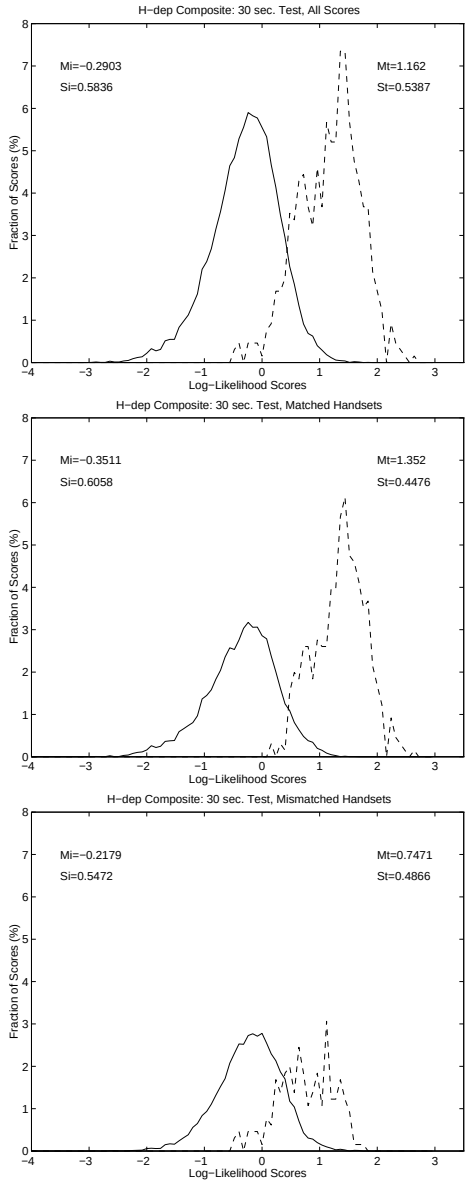

Figure 3. Handset-dep. composite models: histograms of scores for true (right) and false (left) speaker trials with for all (top), matched (middle), and mismatched (bottom) handset types.

REFERENCES

[1] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: experiments on the Switchboard Corpus", *ICASSP*, May 1996.

[2] H. Hermansky *et al.*, "RASTA-PLP speech analysis technique," *Proc. of ICASSP*, pp. I.121-I.124, March 1992.

[3] A.E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," *Proc. of ICASSP*, pp. 81-84, 1996.

[4] L.P. Heck and M. Weintraub, "Rescoring methods for robust telephone-based speaker recognition," *IEEE Trans. on Speech and Audio Proc.*, in preparation.

HANDSET-DEPENDENT BACKGROUND MODELS FOR ROBUST TEXT-INDEPENDENT SPEAKER RECOGNITION

Larry P. Heck and Mitchel Weintraub
SRI International, Menlo Park, CA

This paper studies the effects of handset distortion on telephone-based speaker recognition performance, resulting in the following observations: (1) the major factor in speaker recognition errors is whether the handset type (e.g., electret, carbon) is different across training and testing, not whether the telephone lines are mismatched, (2) the distribution of speaker recognition scores for true speakers is bimodal, with one mode dominated by matched handset tests and the other by mismatched handsets, (3) cohort-based normalization methods derive much of their performance gains from implicitly selecting cohorts trained with the same handset type as the claimant, and (4) utilizing a handset-dependent background model which is matched to the handset type of the claimant's training data sharpens and separates the true and false speaker score distributions. Results on the 1996 NIST Speaker Recognition Evaluation corpus show that using handset-matched background models reduces false acceptances (at a 10% miss rate) by more than 60% over previously reported (handset-independent) approaches.