COMPARISON OF BACKGROUND NORMALIZATION METHODS FOR TEXT-INDEPENDENT SPEAKER VERIFICATION *

Douglas A. Reynolds
Speech Systems Technology Group
MIT Lincoln Laboratory
E-mail: dar@sst.ll.mit.edu

ABSTRACT

This paper compares two approaches to background model representation for a text-independent speaker verification task using Gaussian mixture models. We compare speaker-dependent background speaker sets to the use of a universal, speaker-independent background model (UBM). For the UBM, we describe how Bayesian adaptation can be used to derive claimant speaker models, providing a structure leading to significant computational savings during recognition. Experiments are conducted on the 1996 NIST Speaker Recognition Evaluation corpus and it is clearly shown that a system using a UBM and Bayesian adaptation of claimant models produces superior performance compared to speaker-dependent background sets or the UBM with independent claimant models. In addition, the creation and use of a telephone handset-type detector and a procedure called *hnorm* is also described which shows further, large improvements in verification performance, especially under the difficult mismatched handset conditions. This is believed to be the first use of applying a handset-type detector and explicit handset-type normalization for the speaker verification task.

1. INTRODUCTION

For the task of speaker verification, it has been shown by several researchers that performance can be greatly improved by normalizing raw speaker model likelihood scores by the likelihood score from a background speaker model [1, 2, 3, 4]. That is, for an utterance U, instead of making an accept/reject decision by comparing the claimant model score to a threshold, $p(U|C) > \theta$, a ratio between the claimant model score and a background model score is used, $p(U|C)/p(U|B) > \hat{\theta}$. This ratio is an approximation to the likelihood ratio test for the hypothesis that the utterance was spoken by the claimant speaker. Typically, the background speaker score is approximated by using a collection of alternative speaker models which are "close" to the claimant model in some sense [1] or by using a speaker-independent model trained on a large number of speakers [4].

In this paper we examine techniques for defining background speaker models and the effects different background models have on performance, threshold stability, and computation for a text-independent, telephone-speech, speaker verification task using Gaussian mixture speaker models. Specifically, we compare the use of speaker-dependent background speaker sets to the use of a universal, speaker-independent background model, exploring issues of the size and composition of the background models. A previous study for text-dependent verification was presented in [5]. In addition we also show how claimant models built from the universal background model using Bayesian adaption can greatly improve recognition performance. Experiments are performed on the 1996 NIST Speaker Recognition evaluation corpus, which encompasses the entire Switchboard corpus.

The remainder of this paper is organized as follows. The next section describes the two background modeling procedures. This is followed in section 3 by a description of the 1996 NIST evaluation corpus and results for the different systems on this corpus.

*THIS WORK WAS SPONSORED BY THE DEPARTMENT OF THE AIR FORCE. OPINIONS, INTERPRETATIONS, CONCLUSIONS AND RECOMMENDATIONS ARE THOSE OF THE AUTHOR AND ARE NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

2. RECOGNITION SYSTEMS

The basic model used in all experiments is the Gaussian Mixture Speaker Model [3]. In this model, the distribution of acoustic observation is represented by a Gaussian mixture model (GMM),

$$p(x_t|\lambda) = \sum_{i=1}^{M} p_i \; b_i(x), \tag{1}$$

with mixture weights p_i and Gaussian densities $b_i(x)$. Maximum likelihood parameters are estimated using the EM algorithm. Concatenated mel-cepstra and delta-cepstra features are used as acoustic observations in the experiments. cepstral mean subtraction and RASTA filtering are used for channel equalization in all experiments.

The average log-likelihood of a model given an utterance (parameterized into the a sequence of acoustic observations vectors) $X = \{x_1 \ldots, x_T\}$ is computed as

$$\mathcal{L}(X|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(x_t|\lambda) \tag{2}$$

2.1. Speaker-Dependent Background Sets

For speaker-dependent background sets, a likelihood ratio score between the claimed speaker's model, λ_s, likelihood score and the average likelihood score of a set of background models, $\{\lambda_b\}$, is used [1].

$$\Lambda(X|s) = \mathcal{L}(X|\lambda_s) - \log \sum_{b} \exp\{\mathcal{L}(X|\lambda_b)\}. \tag{3}$$

We employ an algorithm described in [3] which uses an inter-speaker distance measure to select background speakers. Given an utterance X_A from speaker model λ_A and X_B from speaker model λ_B, the distance is defined as

$$d(\lambda_A, \lambda_B) = \log \left\{ \frac{p(X_A|\lambda_A)}{p(X_A|\lambda_B)} \; \frac{p(X_B|\lambda_B)}{p(X_B|\lambda_A)} \right\}. \tag{4}$$

Using this distance measure, the selection algorithm finds "close" speakers, to represent the most competitive voices of imposters,

which are also maximally spread from each other, to reduce redundancy in the background speaker set. A dual set of maximally-spread "far" speakers are also included in the background set to provide representation of dissimilar imposters in the likelihood ratio. In the experiments, we used 15 maximally-spread close (msc) and 15 maximally-spread far (msf) background speaker models per claimant model. All GMMs used 128 mixtures.

2.2. Universal Background Model

One of the difficulties in using speaker-dependent background sets is that it requires an extra step in training to select background sets and more computation and storage during scoring. The goal with a universal background model (UBM) is that one background model can be trained once and used for all claimant speakers. The UBM is a large GMM (2048 mixtures in the experiments) trained on a large number of speakers (in the experiments 6 hours of speech from 45 males and 45 females) to create a speaker-independent model. The number of speakers used to train the UBM should be large enough to cover the general acoustic space of expected imposters and not be overly tuned to any particular speakers.

We examined two way to employ such a UBM for the speaker verification task: independent claimant and UBM models and claimant models adapted from the UBM. In the case of independent models, we simply train a speaker-dependent GMM for each claimant and compute the following likelihood ratio for a test message

$$\Lambda(X|s) = \mathcal{L}(X|\lambda_s) - \mathcal{L}(X|\lambda_b). \tag{5}$$

In the experiments, independent claimant models were 64 mixture GMMs.

For the speaker adapted claimant models, we use Bayesian adaptation to train claimant models from the UBM [6]. The Bayesian adaptation is performed using a fixed relevance factor r to adapt mixture weights, means and variances in the following way. Given a UBM (which acts as the prior distribution) and training observations from the claimant speaker, $X = \{x_1 \ldots, x_T\}$, we determine the probabilistic alignment of the training data into the prior mixture components. For mixture component i we have a probabilistic count of

$$n = \sum_t \Pr(i|x_t) \tag{6}$$

training observations mapping into the prior parameters (w_p, μ_p, σ_p^2) for mixture component i. (Since all adaptation equations refer to a single mixture component, the i notation on parameters is avoided for clarity) The adaptation coefficient for this mixture component is

$$\alpha = \frac{n}{n+r} \tag{7}$$

where r is a fixed relevance factor (typically 16 for 2048 mixtures in our experiments, but dependent on the number of mixtures). The adapted mixture weight is

$$w_a = [\alpha w_s + (1-\alpha) w_p]\, \gamma \tag{8}$$

where

$$w_s = n/T$$

is the mixture weight for the new observations and T the total number of training observations. The scale γ is computed over all adapted mixture weights to ensure they sum to unity.

The adapted mixture mean is

$$\mu_a = \alpha \mu_s + (1-\alpha)\mu_p \tag{9}$$

where

$$\mu_s = \frac{1}{n} \sum_t \Pr(i|x_t) x_t$$

is the mixture mean of the new observations.

The adapted mixture variance is

$$\sigma_a^2 = \alpha E\{x_t^2\} + (1-\alpha)(\sigma_p^2 + \mu_p^2) - \mu_a^2 \tag{10}$$

where

$$E\{x_t^2\} = \frac{1}{n} \sum_t \Pr(i|x_t) x_t^2$$

is the expected squared value of the new observations in the mixture component. Note that the sufficient statistic is adapted not the variance parameter itself.

This approach allows mixture dependent adaptation of parameters. If a mixture component has a low probabilistic count, n, of new data, then $\alpha \to 0$ causing the de-emphasis of the new (potentially under-trained) parameters and the emphasis of the prior (better trained) parameters. For mixture components with high probabilistic counts, $\alpha \to 1$, causing the use of the new speaker-dependent parameters. The relevance factor is used as a way of controlling how much new data should be observed in a mixture before the new parameters begin replacing the prior parameters. This approach should thus be robust to limited training data.

The likelihood ratio for a test message is then computed using equation 5. The fact that the claimant model was adapted from the UBM, however, allows a faster scoring method than merely evaluating the GMM for the claimant and UBM. The approach is based on two observed effects. The first is that when a large GMM is evaluated for an observation, only a few of the mixtures contribute significantly to the likelihood value. This is simply because the GMM represents a distribution over a large observation space, but a single observation will be near only a few components of the GMM. Thus likelihood values can be approximated very well using only the top C best scoring mixture components.

The second observed effect, is that the components of the adapted claimant GMM retain a correspondence with the components of the UBM, so that observations close to a particular component in the UBM will also be close to the corresponding component in the claimant model. Using these two effects, we see that for each incoming observation, it is possible to first determine the top C scoring components in the UBM, which are also used to evaluate the likelihood of the UBM for the observation, and then score only the corresponding C components in the claimant model to evaluate the likelihood of the claimant model for the observation. For a UBM with M mixtures, this requires only $M + C$ component computation per observation compared to $2M$ component computation for normal likelihood ratio evaluations. For multiple claimant models per test message, the savings becomes even greater. In the experiments, we use $C = 5$.

3. EXPERIMENTS

3.1. 1996 NIST Speaker Recognition Evaluation Corpus

Speaker verification using the above background modeling approaches was evaluated on the 1996 NIST Speaker Recognition Evaluation corpus [7]. This corpus is derived from the Switchboard corpus, using all available speakers. The available telephone numbers per conversation was exploited to create matched and mismatched telephone number (handset) test conditions. This is believed to be the largest text-independent speaker recognition task run to date.

The corpus consists of 21 male claimants, 19 female claimants, 204 male imposters and 172 female imposters. A set of 90 speakers, separate from the claimants or imposters, was used for background modeling. There are three training conditions consisting of

- two minutes of training speech extracted from a single session (referred to as 1s1h, 1-session 1-handset),
- two minutes of training speech extracted from two sessions (one minute from each session) originating from the same telephone number (ostensibly the same handset) (referred to as 2s1h), and
- two minutes of training speech extracted from two sessions (one minute from each session) originating from different telephone numbers (ostensibly different handsets) (referred to as 2s2h).

Three test utterance durations of nominally 3 seconds, 10 seconds and 30 seconds were used. For each test utterance duration there are a total of 653 male and 680 female claimant tests (each equally split between tests from one of the telephone numbers used in training and tests from a telephone number not used in training) and 1183 male and 1197 female imposter tests. Verification performance is computed using a speaker-independent threshold on pooled scores from true claimant scores and same-sex imposter scores.

In these experiments, we focused on the male claimants using the 2s2h training condition and the 30 second test condition.

3.2. Baseline Results

Figure 1 shows a plot of the false rejection versus false acceptance errors on an inverse probability warped scale (referred to as Detection Error Tradeoff or DET curves) for the two background approaches for the male claimants on the 2s2h train, 30 seconds test condition. In the top plot of Figure 1 we show results for the matched telephone claimant tests; in the bottom plot Figure 1 we show results for the mismatched telephone claimant tests.

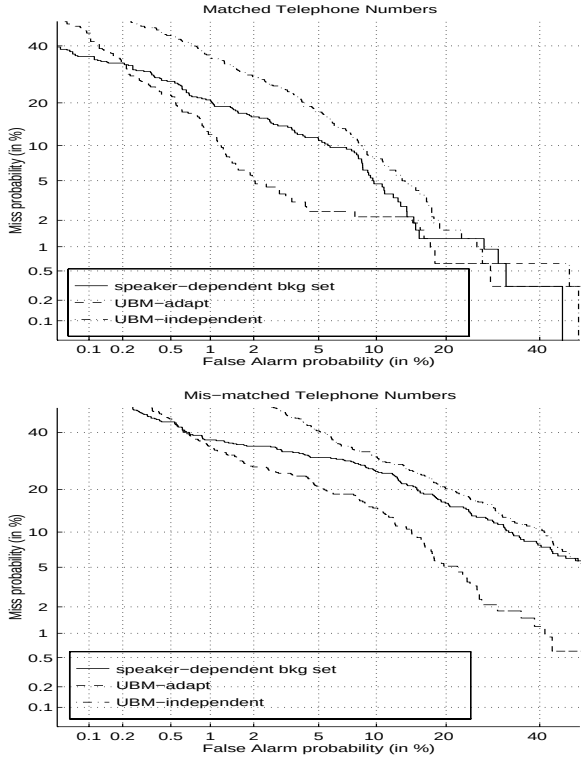

Figure 1. DET curves comparing speaker-dependent background sets and the universal background model (both independent and adapted claimant models) score normalization techniques for matched (upper) and mismatched (lower) telephone number claimant tests.

As clear from these results and also found under the other train/test conditions, the UBM using adapted claimant models significantly outperforms the speaker-dependent background sets and the independent claimant model used with the UBM.

3.3. Handset Dependent Score Normalization - hnorm

In examining the scores produced by the different recognition systems, it became clear that speaker models were producing different distributions of scores for the same test utterances, most significantly for the mismatched telephone number tests. Since a pooled (speaker-independent) threshold is being used, this caused significantly higher false alarm rates for a given miss rate.

Based on earlier work [8, 9], we believed that handset differences associated with different telephone numbers was the root cause of the observed differences. Since handset information is not available, we created a handset detector to label the test utterances a being either from a carbon-button type handset (CARB) or an electret type handset (ELEC). The handset detector is a simple maximum likelihood classifier in which handset dependent GMMs were trained using the HTIMIT corpus[1] [10]. Using these labels, we did indeed observe that different claimant models responded differently to different handset types. This occurs because the claimant model not only represents the speaker but also the handset characteristics over which the training data was collected. Thus a claimant model trained on speech from a CARB handset would tend to score better to other utterances also collected over a CARB handset. There is a similar affinity for claimant models trained with ELEC speech to score well on ELEC test data. These observations and the utility of the handset labeler are supported by work reported in [11].

To normalize out these effects, we developed a handset score normalization technique called hnorm. In hnorm, we first determine the response of a claimant's model to speech with CARB and ELEC labels, The response to CARB speech is parameterized as the mean and variance of the likelihood ratios produced by the claimant model for development utterances labeled as CARB. Likewise for ELEC. Note that the speech used to determine the claimant's response is not from the claimant, but from non-claimant development speakers. Each claimant s then has two sets of parameters describing his/her model's response to CARB and ELEC type speech:

$$\{\mu_s(\text{CARB}), \sigma_s(\text{CARB}), \mu_s(\text{ELEC}), \sigma_s(\text{ELEC})\}$$

During testing, an input utterance is first labeled as CARB or ELEC and scored as normal to obtain the likelihood ratio score for a particular claimant model. To hnorm the likelihood ratio, $\Lambda(X|s)$, we apply the hnorm parameters as follows (assume X was labeled as CARB):

$$\Lambda_{\text{HNORM}}(X|s) = \frac{\Lambda(X|s) - \mu_s(\text{CARB})}{\sigma_s(\text{CARB})} \tag{11}$$

This has the effect of causing each claimant model to produce zero mean and unit standard deviation scores for non-claimant speech, independent of the handset characteristics of the test utterance or of those used in training the claimant model. In addition to helping normalize out handset-dependent biases for a particular claimant model, this normalization also makes a speaker-independent threshold more effect for all claimant speakers.

The hnorm procedure was applied to the evaluation corpus. A comparison of the baseline UBM using claimant model adaptation with and without applying hnorm is shown in Figure 2. It is evident that hnorm produces a significant reduction in errors for the mismatched condition. At 10% miss rate, the false alarm rate decreases from 14.5% to 2.4% – an 83% reduction in error.

[1] HTIMIT is a handset dependent corpus derived by playing a subset of TIMIT through known CARB and ELEC handsets.

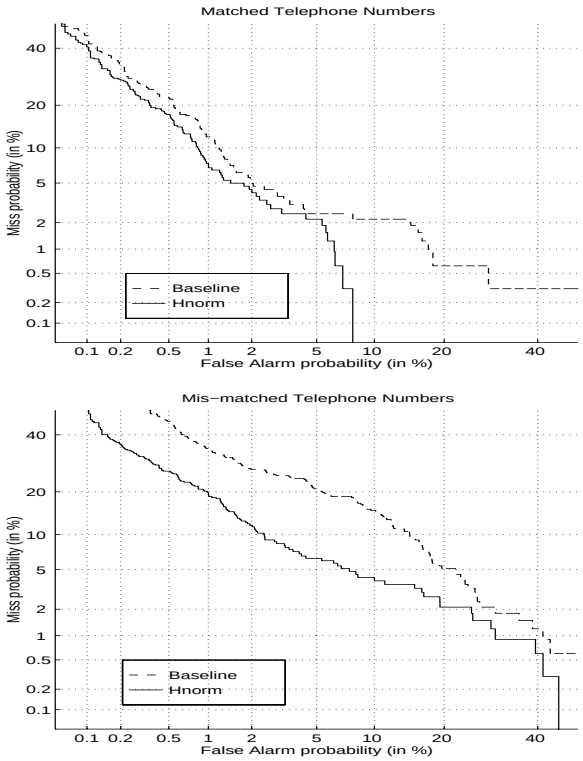

Figure 2. DET curves comparing the UBM with claimant model adaptation baseline with hnorm score normalization for matched (upper) and mismatched (lower) telephone number conditions.

In Figure 3 we show the distribution of log-likelihood ratio scores both with and without hnorm applied. Note that prior to hnorm (upper plots), the score distribution for mismatched claimant tests is clearly bi-modal, indicating differing responses for different utterances. After applying hnorm, we see that the distribution for mismatched claimant tests is tighter although still appearing bi-modal and not as tight as the distribution for matched claimant tests. Even for the matched claimant tests, hnorm appears to help with a few low scoring tests[2].

4. CONCLUSION

This paper has compared two approaches to background model representation for a text-independent speaker verification task using GMMs. Experiments were conducted on the 1996 NIST Speaker Recognition Evaluation corpus and it was clearly shown that a system using a universal background model (UBM) and Bayesian adaptation of claimant models produced superior performance compared to speaker-dependent background sets or the UBM with independent claimant models. In addition, a procedure called *hnorm* was described which was shown to further improve verification performance, especially under the difficult mismatched handset conditions. It is believed that this is the first use of applying a handset-type labeler and explicit handset-type normalization for the speaker verification task.

Future work is expected to focus on a different uses of the handset labels to normalize features or model parameters directly for improved handset variability compensation.

[2]These likely occur when calls from a single phone number use different telephone handsets.

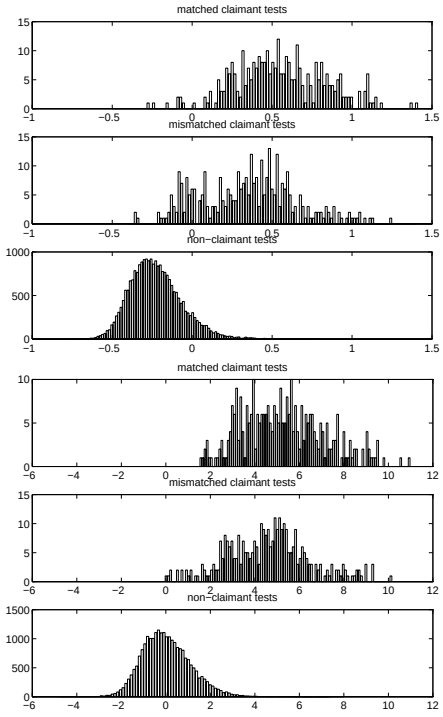

Figure 3. Distribution of log-likelihood ratio scores for matched claimant tests, mismatched claimant tests and non-claimant tests. All scores are from the UBM with claimant adaptation. The upper three plots are baseline scores. The bottom three plots are for scores after hnorm has been applied.

REFERENCES

[1] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, pp. 89–106, 1991.

[2] A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," ICSLP, pp. 599–602, November 1992.

[3] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, August 1995.

[4] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech Communication*, vol. 17, pp. 109–116, August 1995.

[5] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," ICASSP, pp. 81–84, May 1996.

[6] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Trans on Speech and Audio Proc*, vol. 2, pp. 291–298, April 1994.

[7] NIST, "March 1996 NIST speaker recognition workshop notebook." NIST administered speaker recognition evaluation on the Switchboard corpus, March 27–28 1996.

[8] D. Reynolds, M. Zissman, T. Quateri, G. O'Leary, and B. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," ICASSP, pp. 329–332, May 1995.

[9] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus," ICASSP, pp. 113–116, May 1996.

[10] D. A. Reynolds, "HTIMIT and LLHBD: Speech corpora for the study of handset transducer effects," ICASSP, April 1997.

[11] L. P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," ICASSP, April 1997.