SPEAKER VERIFICATION WITH LIMITED ENROLLMENT DATA

Owen Kimball *Michael Schmidt* *Herbert Gish* *Jason Waterman*

BBN Systems and Technologies
70 Fawcett St., Cambridge, MA 02138 USA
okimball@bbn.com

ABSTRACT

New methods for speaker verification that address the problems of limited training data and unknown telephone channel are presented. We describe a system for studying the feasibility of telephone based voice signatures for electronic documents that uses speaker verification with a fixed test phrase but very limited data for training speaker models. We examine three methods for speaker verification that address these characteristics in different ways, including text-independent mixture models, a broad phonetic category model that has some of the properties of both text-dependent and text-independent approaches, and a text-dependent approach based on speaker adaptation. The speaker-adaptive approach is shown to have significantly better performance when the training and test channel conditions are mismatched, resulting in better overall performance across all conditions.

1. Introduction

In this paper we present new methods for speaker verification to deal with the problems of limited training data and unknown telephone channels. We first describe the application we are investigating, telephone based voice signatures for electronic documents, and the implications of the system's requirements for speaker verification. The limited training data available in this task leads us to examine alternative methods that can robustly estimate speaker characteristics, including text-independent mixture models, a model based on broad phonetic categories that has some of the desirable properties of both text-dependent and text-independent approaches and a text-dependent approach based on speaker-adaptive modeling. We find that the speaker-adaptive approach is able to estimate speaker models well with very limited data and that the models show a significant improvement over other methods when the channel conditions on training and test are mismatched.

2. Voice Signature Research System

The methods presented in this paper were developed as part of a system being developed to evaluate the feasibility of signing electronic documents by speaking over the telephone. There are a number of applications in commerce and government for which electronic documents represent a large potential savings in processing time as well as a convenience to customers, but such documents often require an alternative form of signature. Telephone-based voice signatures are one alternative to requiring a handwritten signature on a separate piece of paper. In this project we have developed a system to investigate some of the technical and practical issues of voice signatures.

The two key components of a voice signature are the words of the statement and the voice characteristics of the speech. The wording of a signature for an electronic document might be similar to the statement above the signature on a corresponding paper form, i.e. essentially affirming that the document is true to the best of the signer's knowledge. The voice characteristics present in the speech play a role analogous to a person's distinctive handwriting in a written signature. One of the questions we hope to address with this research is whether a voice signature can function as an adequate replacement for a written signature in determining that a signer is who they claim to be, and in particular how well speaker verification can perform in automatically screening for possible fraudulent signatures.

In an actual deployment of a voice signature system for a large customer base, it would be both inconvenient and impractical to request that every user of the system enroll in the system before signing a document by contributing training speech prior to calling in their first voice signature, particularly if documents are signed infrequently. In this project we are considering the alternative of collecting a single voice signature as training from each system user and building a speaker model based on it. While this approach avoids unduly inconveniencing the user, it poses a challenge to the verification system of providing a single phone call and utterance for training.

The system we are using to research these issues has two major roles, compliance checking and speaker authentication. In compliance checking, the system determines if a caller is making a good faith attempt at speaking the signature statement correctly. The determination is made using speech recognition at the time of the call in order to give the caller feedback and reprompt them if there are any problems. The recognizer uses speaker-independent (SI) acoustic models and has a grammar of the expected signature phrase that also allows a number of the words to be omitted. A simple mechanism that essentially counts the number of content words in the recognized text and compares the sum to a threshold has proven to be effective at allowing moderate disfluencies and occasional missed words while eliminating sentences that are far from the target phrase. If the system has trouble recognizing a caller's speech, the caller is given three tries before the system asks them to call back and try again later.

The second function of the system is to verify that a caller's voice matches their claimed identity using speaker verification based on the model built from the target speaker's training signature. Verification is run after the caller has hung up, since the process of identifying suspect signatures will not be perfectly reliable and may in some cases require human involvement. Note that the verifier's task is simplified by the system's compliance check, since both training and test utterances must be reasonable approximations to the target phrase.

The limited training data available in the scenario outlined leads us to consider verification methods that permit robust model estimation with very limited training data. With text-independent models such as the Gaussian mixture model (GMM) [1], the number of parameters can be readily configured to allow training with small amounts of data. We explore using such a model in the experiments. However, given our knowledge of the sentences spoken, we would prefer to take advantage of the more detailed modeling possible with text-dependent methods. One possibility is to use phonetic-class-dependent models: with a sufficiently small number of classes we may get some of the benefits of text-dependent models while keeping the number of free parameters manageable. Another approach is to take advantage of recent advances in speaker adaptation and attempt to train speaker-dependent models of the detailed phonetic contexts that appear in the text, using a technique such as Maximum Likelihood Linear Regression (MLLR) [2]. We describe these alternatives in more detail in the following sections.

3. Broad Phonetic Category Model

The Broad Phonetic Category (BPC) model is based on identifying the phonetic categories in an utterance and modeling each separately. Separating categories in this way should make the model more discriminative than typical text-independent methods that use a single class for all speech. The system relies on an automatic labeling of the categories by a speech recognition system. Although this recognizer can make mistakes, the broad nature of the categories may help reduce the problems due to labeling errors. To score an utterance against a speaker model, we model each of the phonetic classes following the approach described in [3] and take a simple combination of the scores across the classes.

In our implementation, we considered 5 phonetic categories consisting of plosive, fricative, semi-vowel, and two vowel classes. We investigated two methods for labeling the speech. In the first we used a speaker independent recognition system to perform word recognition and phonetic alignment within words; the phoneme labels were then mapped to their broad phonetic categories. In the second method we directly label the speech using a recognizer that has acoustic models of the phonetic classes and a "language model" consisting of a statistical bigram that gives the probability of phonetic class given preceding class. Our initial experiments showed that the phonetic labeling produced in the former approach was sensitive to small errors in the word recognition as well as phenomena such as mis-starts and other disfluencies. For this reason, we focused exclusively on the second method using a phoneme recognizer in the experiments reported later in this paper.

Given a phonetic labeling of the training speech, we estimate models of each of the separate phonetic classes, using the approach described in [3]. Specifically, each BPC is modeled by the mean and covariance of the cepstral features and the covariance of the delta cepstra estimated from all training speech in that category for a speaker. In testing, we estimate the corresponding sample statistics for the same phonetic classes in an utterance. We then obtain the likelihood of the test statistics given the training models, where the cepstral mean is modeled as a normal distribution and the two covariance matrices are modeled as Wishart distributions. We take a simple sum of the likelihoods across the means and covariances of all phonetic classes to determine the score for an utterance given a speaker's model. The score for an utterance is normalized by subtracting the best score for the utterance across a set of cohort speakers. Experiments using this model are described below.

4. Speaker-adaptive model estimation

Another method considered uses supervised speaker adaptation to address the problem of estimating the parameters of a text-dependent, HMM-based system with limited enrollment data. Starting from a SI model based on the BYBLOS system's continuous density HMM [4], we apply maximum likelihood linear regression (MLLR) adaptation [2] using each speaker's enrollment speech to adapt the SI model. Specifically, using the known enrollment text, we adapt the means of the mixture model, keeping the mixture weights and covariance estimates fixed, and we estimate a single transformation matrix per speaker. With a single utterance for adaptation data, we find that a single EM pass is sufficient in estimating counts in the MLLR process. The estimated transformation is then applied to the SI model to generate the speaker model that is used at verification time.

There are a number of variants of the SI model that can be adapted to speakers. We have investigated phonetically-tied mixture (PTM) models, in which all triphones with the same center phoneme share a common set of Gaussians with different mixture weights, as well as state-clustered tied-mixture (SCTM) models, where the tying of codebooks as well as states that share mixture weights is determined by clustering the states of the HMM. The number of Gaussians for each of these model types can also be varied, and we have investigated PTM models with 64 and 256 Gaussians per mixture codebook and SCTM with 8 and 32 Gaussians per codebook. In our experiments to date we have found only minor differences between these different models, and we discuss later results based on the PTM model with 64 terms (PTM-64).

Verification is accomplished by scoring putative target speakers against their corresponding model and normalizing the result with an appropriate cohort model score [5]. Scoring consists of using constrained recognition to find the log likelihood of the verification speech given the HMMs corresponding to the known text.

Both the adaptation and verification require a transcription of the speech, which can be obtained either by using the prompted text (and assuming no significant errors in reading) or using the transcript produced by the compliance recognizer. While neither method is completely accurate, since speakers do deviate from the prompt and since the recognizer misses effects like stuttering and occasionally deletes or inserts words, we find

that using the recognizer's output gives slightly better performance than using the prompt.

5. Database

To evaluate these methods, we collected a database of speakers to simulate telephone voice signatures for electronic documents. There were 135 speakers who called from different telephones over long distance lines to a collection system. Each speaker made two phone calls in which they spoke a dummy signature statement, with the first call reserved for training and the second for test. The dummy signature statement contained 34 words and the average speaking time for the signature across all speakers was 11 seconds.

For 82 of the calls, a record of the originating telephone number used for the call is available. Of these cases, 45 of the callers used different numbers for training and test calls and 37 used a single number for both. In the experiments below, we look at the effect of channel on performance. Our assumption is that the largest component of the channel effect is the telephone handset used. We also assume calls made from the same telephone number use the same telephone handset, and therefore categorize such calls as "matched" channel, whereas calls from different telephone numbers are assumed to have "mismatched" channel.

6. Experimental Results

In the experiments described below, the speech was analyzed every 10 msec with a 20 msec Blackman window and band-pass filtered between 300 and 3300 Hz. LPC cepstra were computed and then RASTA filtered to remove a causal estimate of the long-term average cepstra [6]. In experiments with the BPC and speaker-adaptive models, the cepstral order was 14, while for the GMM model, previous experiments indicated better performance with $19th$ order cepstra. We did not test to see if 19 cepstra also improved the BPC and adaptive models.

In all the methods evaluated, we normalize the raw likelihoods obtained from a model by cohort normalization. The basic normalization is as follows: models are estimated for every speaker in the cohort set, and in testing, an utterance is scored against a target model and all cohort speaker models. The normalization score for an utterance is the maximum of the log likelihoods across all cohort speakers. In order to maximize the number of test speakers while keeping cohort and target speakers separate, we split the test data into two groups and jackknife the roles of target and cohort between them, with each of the groups serving as the cohort set for the other. Within a group of targets we further jackknife so that each speaker's test utterance serves as a true test for their own model and an impostor test for all other speakers' models. With a voice signature system, we assume that each speaker's accept/reject threshold would be tuned separately and therefore find the ROC curve for all speakers as the average of the per speaker ROC's. Averaging speaker ROC's in this way rather than forming a single speaker-independent ROC gives a small but consistent improvement across all the methods described.

The first model tested on the voice signature task was a text-independent GMM model. One potential advantage of the GMM for our problem is that its text-independence allows training with a relatively small amount of data simply by scaling the number of terms in the mixture. We evaluated the GMM using a 16-mixture and 32-mixture model. In each case, diagonal covariance Gaussian mixtures for each speaker were initialized with k-means clustering and trained with 5 passes of EM. We found that the 32-mixture model performed slightly better for this task and results for that model are reported below.

We compared the performance of the 32-mixture GMM, the BPC model and the speaker adaptive model. For the speaker adaptive model, the adaptation was from a PTM-64 SI model and the transcriptions used in both adaptation and verification were from the recognition done in compliance checking. Figure 1 shows the ROC curves for the three different systems. The horizontal lines to the y-axis indicate the equal error rate (EER) points, which are 8.1%, 15.4% and 16.2% for the MLLR-adapted, BPC and GMM model respectively.

To better understand the systems' behavior we compared their performance when the test and training channel were either matched or mismatched. Figure 2 compares the ROC's for the BPC and adapted systems for same and different channel conditions (the GMM model's performance characteristics were essentially the same as the BPC's). For these experiments, we looked at the performance of the 82 callers from the complete set of 135 for whom we had records of their originating test and training telephone number. The EERs for the four conditions are given in Table 1. The BPC model obtains very good performance on same-channel tests, but it is more than a factor of 3 worse on the mismatched condition. In contrast the speaker-adapted model, although performing slightly worse than the BPC model on the same-channel condition, seems to be unaffected by channel mismatch. In fact, the adapted model performs slightly better on mismatched channel than on matched, although this difference may well be insignificant.

System	EER Same	EER Diff
Speaker Adapted	.081	.065
BPC	.066	.20

Table 1. Channel effects on EER.

The greatly improved performance on cross channel seems to account for the better overall performance of the adapted model approach. We surmise that the adapted model's improved cross-channel performance is due to adapting an SI model that is trained on calls from a number of channels: the adaptation moves the model toward a speaker, but carries with it a number of channel variants. This hypothesis is consistent with the slightly better same-channel performance for the BPC model: training only on a single speaker and channel is more effective when testing on that same channel than trying to adapt a somewhat broader, multi-channel SI model to a speaker.

We have also investigated alternative cohort normalization methods. We find that normalizing by the sum of the top-N cohort likelihoods, where N is chosen to be about 10% of the number of cohort speakers, rather than choosing the single highest cohort likelihood, gives a small but consistent improvement. In the speaker adaptive system, this change improves the equal error rate from 8.1% to 7.4%, but it can also be applied to any of the methods discussed.

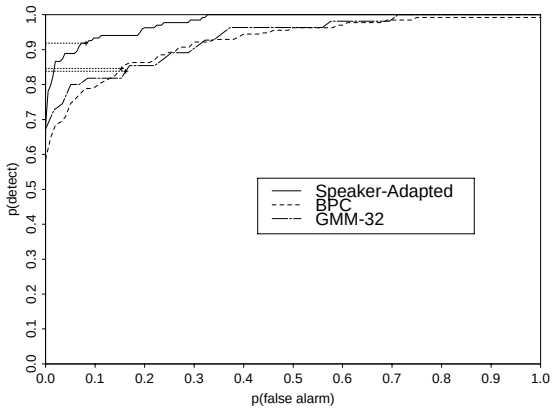

Figure 1. Comparison of GMM, BPC and Speaker-Adaptive systems on all tests.

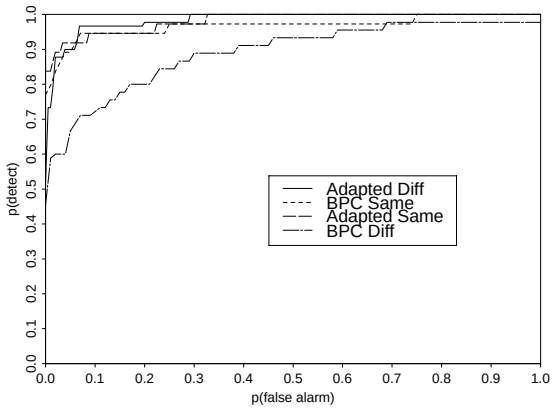

Figure 2. Comparison of BPC and Speaker-Adaptive systems for same and different channel tests.

7. Conclusions

We have investigated alternative methods of model estimation for speaker verification with very limited training data, including two new approaches, the BPC model, and a model using MLLR adaptation of a SI model. The adapted model shows significant improvement for cross-channel performance and better overall performance than the BPC and GMM models.

7. References

[1] D. A. Reynolds "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Communication,* Vol 17, August 1995.

[2] V.J. Leggetter and P.C. Woodland, "Speaker Adaptation Using Linear Regression," Technical Report CUED/F-INFENG/TR.181, Cambridge University, Engineering Department, June 1994.

[3] H. Gish, M. Schmidt, A. Mielke, "A Robust Segmental Method for Text-Independent Speaker Identification," Proc. ICASSP '94, April 1994, Adelaide, South Australia, pp. 145-148.

[4] L. Nguyen, T. Anastasakos, F. Kubala, C. Lapre, J. Makhoul, R. Schwartz, N. Yuan and G. Zavaliagkos, "The 1994 BBN/BYBLOS Speech Recognition System," Proc. Spoken Language Systems Technology Workshop, Morgan Kaufmann Publishers, pp. 77-81, 1995.

[5] A. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang, and F. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," Proc. ICSLP '92, Banff, Canada, vol 2, pp. 599-602.

[6] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP Speech Analysis Technique," Proc. ICASSP '92, March 1992, San Francisco, pp. 121-124.