ASR DEPENDENT TECHNIQUES FOR SPEAKER IDENTIFICATION

Alex Park and Timothy J. Hazen

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139 USA
{malex,hazen}@sls.lcs.mit.edu

ABSTRACT

Traditional text independent speaker recognition systems are based on Gaussian Mixture Models (GMMs) trained globally over all speech from a given speaker. In this paper, we describe alternative methods for performing speaker identification that utilize domain dependent automatic speech recognition (ASR) to provide a phonetic segmentation of the test utterance. When evaluated on YOHO, several of these approaches were able outperform previously published results on the speaker ID task. On a more difficult conversational speech task, we were able to use a combination of classifiers to reduce identification error rates on single test utterances. Over multiple utterances, the ASR dependent approaches performed significantly better than the ASR independent methods. Using an approach we call speaker adaptive modeling for speaker identification, we were able to reduce speaker identification error rates by 39% over a baseline GMM approach when observing five test utterances from a speaker.

1. INTRODUCTION

The most common approach to speaker recognition today is the use of global Gaussian mixture models (GMM) [1]. The primary benefit of the GMM approach is that speaker identification can be performed in a completely text independent fashion (i.e., no knowledge of the words spoken by the speaker is required). However, because this approach ignores knowledge of the underlying phonetic content of the speech, it does not take advantage of all available information.

In this paper we strive to improve upon the GMM approach by combining it with other classification techniques which utilize information about the phonetic content of the speech. One of the disadvantages of the GMM's global model is that the acoustic variability of phonetic events in the test utterance is not taken into account when comparing different speakers. Although it has been shown that some phonetic classes have higher speaker distinguishing capabilities than others [2], much of this information is lost when all enrollment data is mapped to a single acoustic model. In our work we utilize a speech recognition engine to hypothesize the phonetic content of the test utterance. We then use this knowledge during speaker identification by applying refined phone dependent models in place of a global GMM. We believe that this approach is feasible in domain dependent applications where a reliable speech recognition engine is available.

This research was supported by an industrial consortium supporting the MIT Oxygen Alliance.

In addition to exploring these speech recognition based scoring techniques, we introduce a two-stage scoring framework which reduces computational demands presented by more refined speaker models. This framework also allows us to easily combine the output of several different classifiers.

Finally, we investigate the effect of performing speaker identification over multiple utterances. Traditionally speaker identification systems have focused on the goal of maximizing identification rates over individual, short utterances (1-3 seconds). While this is a reasonable metric for password driven verification tasks, recent research has also focused on data tasks where speaker recognition is performed over a collective set of utterances from a target speaker [3]. Forensic speaker identification, rich transcription of conversational data, and verification in transactional applications are all scenarios where a system would have access to multiple utterances prior to making a decision.

The rest of the paper is organized as follows. First we discuss the implementation of two baseline approaches that are based on the well-known GMM approach introduced by Reynolds [1]. Next, we detail two newer approaches which make use of speech recognition on the test utterance. Following that, we give a description of the corpora and conditions for our experiments. Finally, we discuss our results and give future directions for our work.

2. IMPLEMENTATION

We distinguish between traditional text independent approaches which we classify as ASR independent, and ASR dependent approaches which make use of speech recognition during speaker identification. For each of the different approaches, we used the same set of front-end acoustic features.

2.1. ASR Independent

2.1.1. Gaussian Mixture Models

Our baseline system was closely-based upon Reynolds' GMM approach [1]. For each input waveform, 14-dimension mean normalized MFCC vectors were computed at a frame rate of 10ms. From the MFCCs, 112-dimension input feature vectors were created by concatenating averages of MFCCs from eight different segments surrounding the current frame. Principal components analysis was then used to reduce the dimensionality of these feature vectors to 50 dimensions [4]. Global GMM models were then trained for each speaker using all non-silence frames in their enrollment data.

In *Proceedings of the 7th International Conference on Spoken Language Processing*, Sep. 16-20, 2002, Denver, Colorado, pp. 1337-1340.

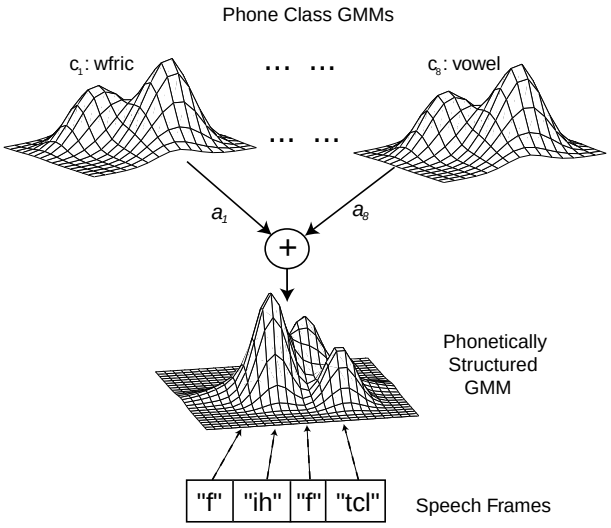

Fig. 1. Phonetically Structured GMM scoring framework

2.1.2. Phonetically Structured GMMs

A recent variant of the traditional GMM approach is the so-called "phonetically-structured" GMM method which has been proposed by Faltlhauser *et al.* [5]. This method trains smaller "granular" GMMs on separate phonetic classes for each speaker, then combines them into a larger single model which is used for identification. By combining the various phonetic models using a globally determined weighting, this method is less sensitive to phonetic biases present in the enrollment data of individual speakers. Examples of the phonetic classes that were used are: vowels, strong fricatives, liquids, etc. In total, eight phonetic classes were used for training. During identification, all speech frames from the test utterance are scored against the combined model, as illustrated in Figure 1.

2.2. ASR Dependent

The following two approaches require a speech recognition engine in order to generate a hypothesized phonetic segmentation of the test utterance. The generation of this hypothesis is described in Section 2.3.

2.2.1. Phonetic Classing

The use of separate phonetic manner classes for speaker modeling was studied previously by Sarma [6]. This technique is similar to the use of phonetically structured GMMs in that training is identical. Phonetic class GMMs were trained for each speaker, but instead of being combined into a single speaker model, the individual classes were retained. During identification, each test vector was assigned to a phonetic class using the phonetic segmentation hypothesis provided by the speech recognizer. The appropriate phone class model was then used to score the vector. This scoring procedure is illustrated in Figure 2.

Since test vectors were scored directly against the granular GMMs, this approach is similar to the "multigrained" method proposed by Chaudhari *et al.* [7]. However, by using the phone class

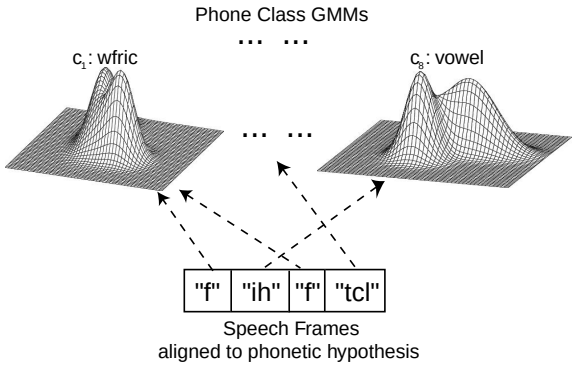

Fig. 2. Phonetic Class scoring framework

assignment provided by the speech recognizer, this approach eliminates the need to score against every model in the speaker's library, as is required by the multigrained method.

2.2.2. Speaker Adaptive Scoring

The previous two approaches attempt to improve upon the global GMM approach by using broad phonetic class models which are more refined than the global GMM. At a further level of granularity, models can be built for specific phonetic events. Unfortunately, the enrollment data sets for each speaker in typical speaker ID tasks are not large enough to build robust speaker dependent phonetic-level models. To compensate for this problem, we can draw upon techniques used in the speaker adaptation field. This allows us to build models that learn the characteristics of a phone for a given speaker when sufficient training data is available, and rely more on general speaker independent models in instances of sparse training data.

In this approach, speaker dependent segment-based speech recognizers were trained for each speaker. During identification, the hypothesized phonetic segmentation produced by the speaker independent speech recognizer was used to generate the best path speaker dependent score, which was then interpolated with the recognizer's speaker independent score. This method approximates the MAP strategy for speaker adaptation [8], and is similar to the LVCSR-based speaker verification system developed by Dragon Systems and described by Weber *et al.* in [9]. Mathematically, if the word recognition hypothesis assigns each test vector x to a phone j, then the score for speaker i is given by

$$S_i = \sum_x \log[\lambda_{i,j} p(x|M_{i,j}) + (1 - \lambda_{i,j}) p(x|M_j)]$$

where $M_{i,j}$, M_j are the speaker dependent and speaker independent models for phone j, and $\lambda_{i,j}$ is an interpolation factor given by

$$\lambda_{i,j} = \frac{n_{i,j}}{n_{i,j} + \tau}$$

In this equation, $n_{i,j}$ is the number of training tokens of phone j for speaker i, and τ is an empirically determined tuning parameter that was the same across all speakers and phones.

2.3. Two Stage Scoring

In our system, we utilized a two-stage method to calculate speaker scores. This framework is illustrated in Figure 3. In the first stage,

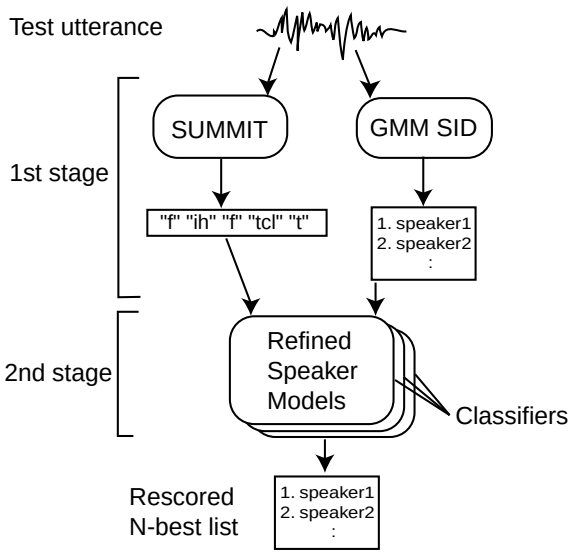

Fig. 3. Two stage scoring framework indicating parallel ASR and GMM speaker ID computation

the test utterance is passed in parallel through a speech recognition module and a GMM speaker ID module, which is implemented using the baseline approach. The speech recognition module produces a time-aligned phonetic hypothesis, while the GMM speaker ID module produces an N-best list of hypothesized speakers. These results are then passed to the next stage, where a second classifier rescores each speaker in the N-best list using one of the refined techniques described above.

This two-stage scoring method is useful in a number of ways. First, by using the GMM speaker ID module for fast-match, we reduce post-recognition latency by limiting the search space of speakers presented to the second stage. Identification performance is not significantly affected since the probability of N-best exclusion of the target speaker by the GMM module can be made arbitrarily low by increasing N. Furthermore, there is little increase in pre-identification latency for the ASR dependent approaches since the GMM scoring proceeds in parallel with word recognition. Another advantage of this framework is that scores from multiple classifiers can be used and combined in the second stage.

3. EXPERIMENTS

3.1. Corpora

For evaluation, we used two significantly different corpora. The first corpus, YOHO, consisted of 138 speakers reading six digit combination lock phrases, and was recorded in a low noise office environment [10]. Although recording was done on a high-quality telephone handset, the speech was not passed through a telephone channel. Training data is made up of 96 enrollment phrases which are identical over all speakers. On average, each speaker has approximately 180 seconds of speech training data.

To simulate a more difficult variable condition task, we created a second corpus out of speaker-labeled data taken from the MIT MERCURY airline travel information system [11]. The MERCURY data set consisted of variable length spontaneous speech utterances gathered from 38 speakers across a variety of telephone channels and handsets. Training data is limited to approximately 50 variable length utterances per speaker. The total amount of training data per speaker ranges from 30 seconds to 90 seconds of actual speech.

	Error Rate (%)	
Method	YOHO	MERCURY
Baseline GMM	0.83	22.4
Phonetically Structured GMM (PS)	0.31	21.3
Phone Classing (PC)	0.40	21.6
Speaker Adaptive (SA)	0.31	27.8
Multiple Classifiers (GMM+SA)	0.53	19.0
Multiple Classifiers (PS+SA)	0.25	18.3

Table 1. Comparison of identification error rates for each approach on YOHO and MERCURY data sets

3.2. Experimental Conditions

For both corpora we used domain dependent implementations of the MIT SUMMIT speech recognizer [4]. On the YOHO data set, the vocabulary was limited to allow only the set of possible numerical combination lock phrases. On the MERCURY data set, the recognizer was limited to a 2200 word vocabulary for conversational queries regarding airline travel. Empirically determined parameters such as classifier combination weights and interpolation parameters were found by tuning on an independent set of MERCURY development data.

4. RESULTS

For this project, we chose to confine our experiments to the task of closed-set identification rather than speaker verification. The motivation for doing so was to compare the relative speaker distinguishing capability of each system without having to consider the effect of different background model normalization schemes required for verification tasks.

We first computed results for the closed set identification task on individual utterances. These results are shown in Table 1. When comparing the performance of the different classifiers, we observed that error rates on the YOHO corpus were uniformly low. In particular, we noted that our best results on the YOHO corpus were better than the 0.36% identification error rate obtained by a system developed at Rutgers [10]. This is the best result that we are aware of that has been reported for this task. With the exception of systems involving the GMM baseline, each of the classifiers produced between 14 and 22 total errors out of 5520 test utterances, making the differences between these approaches statistically insignificant.

On the MERCURY data set, the comparative performance of each system is more evident. Both the phonetically structured GMM system and the phonetic classing system give slight improvements over the baseline, while the speaker adaptive system has a higher error rate than any of the other approaches. Across all systems, we observed that error rates were significantly higher on the MERCURY task than on YOHO, clearly illustrating the increased difficulties associated with spontaneous speech, noise, and variable channel conditions. These factors also led to a higher word error rate on the MERCURY data, which partially explains why the recognition aided systems did not yield improvements

	Error Rate (%)		
Method	1 Utt	3 Utt	5 Utt
Baseline (GMM)	22.4	14.3	13.1
PS	21.3	15.6	14.3
PC	21.6	14.9	13.8
SA	27.8	10.3	7.4
GMM+SA	19.0	9.7	7.5
PS+SA	18.3	11.2	8.0

Table 2. ID error rates over 1, 3, and 5 utterances on MERCURY

over the baseline GMM method as observed with YOHO. However, we saw that by combining the outputs of multiple classifiers, lower overall error rates were achieved on both corpora.

In order to test the performance of these methods on multiple utterances, we performed additional experiments on the MERCURY corpus. Identification error rates over 1, 3, and 5 utterances are shown in Table 2. For all methods, scoring over multiple utterances resulted in significant reductions in error rates. We observed that the speaker adaptive approach attained the lowest error rates among the individual classifiers as the number of test utterances was increased (Figure 4). Moreover, as the number of utterances was increased past 3, the performance of the combined classifiers exhibited no significant gains over the speaker adaptive approach. When compared to the next best individual classifier, the speaker adaptive approach yielded relative error rate reductions of 28%, 39%, and 53% on 3, 5, and 10 utterances respectively.

5. CONCLUSIONS

In this paper, we evaluated speaker modeling techniques which make use of speech recognition. By focusing on domain dependent tasks and using a two-stage scoring system, we were able to implement these techniques in a computationally feasible manner. On the YOHO corpus, we were able to use classifier combination to attain extremely low identification error rates. For the more difficult MERCURY task, we also observed significant improvement on single utterances by using classifier combination. Over multiple utterances, however, we found that speaker adaptive scoring yielded the greatest gains when compared to the other approaches.

6. FUTURE WORK

We plan to further investigate the use of speaker adaptive scoring in extended speaker verification tasks by implementing a background model scoring scheme. One useful application of this technique would be in the MERCURY domain, where users identify themselves at the beginning of the session, but usually go through several non-critical queries before attempting to perform a secure transaction, such as ticket purchase. In this type of scenario, the system would have access to several utterances from the target speaker prior to making a verification decision.

In addition to the speaker modeling approaches discussed in this paper, we plan to incorporate the use of noise robust measurements, such as formant locations, fundamental frequency, and duration into the feature set used for speaker identification.

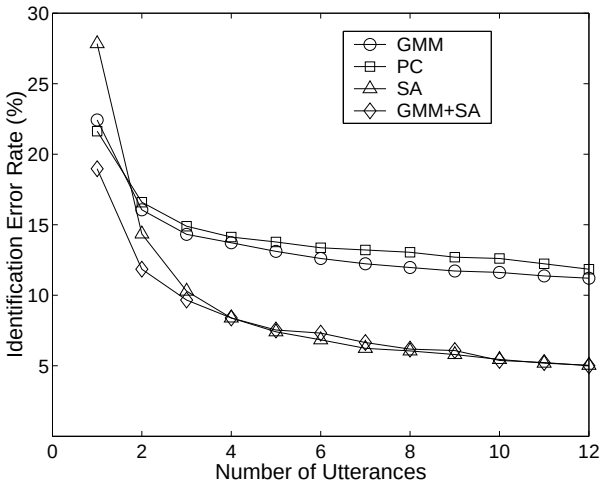

Fig. 4. ID error rates over multiple utterances on MERCURY

7. REFERENCES

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

[2] J. Eatock and J. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Proc. ICASSP*, Adelaide, Apr. 1994, vol. 1, pp. 133–136.

[3] W. Andrews, M. A. Kohler, and Joseph P. Campbell, "Phonetic speaker recognition," in *Proc. Eurospeech*, Aalborg, Sept. 2001, pp. 2517–2520.

[4] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP*, Philadelphia, Oct. 1996, pp. 2277–2280.

[5] R. Faltlhauser and G. Ruske, "Improving speaker recognition performance using phonetically structured gaussian mixture models," *Proc. Eurospeech*, Aalborg, Sept. 2001, pp. 751–4.

[6] S. Sarma and V. Zue, "Segment-based speaker verification system using SUMMIT," in *Proc. Eurospeech*, Rhodes, Sept. 1997, pp. 843–846.

[7] U. Chaudhari, J. Navratil, and S. Maes, "Transformation enhanced multi-grained modeling for text independent speaker recognition," in *Proc. ICSLP*, Beijing, Oct. 2000, vol. 2, pp. 298–301.

[8] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, 291-298, April 1994.

[9] F. Weber, B. Peskin, M. Newman, A.C-. Emmanuel, and L. Gillick, "Speaker recognition on single- and multispeaker data," *Digital Signal Processing*, vol. 10, 75-92, Jan. 2000.

[10] J. Campbell, "Testing with the YOHO CD-ROM voice verification corpus," in *Proc. ICASSP*, Detroit, May 1995, pp. 341–344.

[11] S. Seneff and J. Polifroni, "Dialogue management in the MERCURY flight reservation system," in *Satellite Dialogue Workshop, ANLP-NAACL*, Seattle, April, 2000.