

Speaker Verification using Text-Constrained Gaussian Mixture Models*

D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F. Quatieri

MIT Lincoln Laboratory, Lexington, MA USA

{sturim,dar,rbd,tfq}@sst.ll.mit.edu

ABSTRACT

In this paper we present an approach to close the gap between text-dependent and text-independent speaker verification performance. Text-constrained GMM-UBM systems are created using word segmentations produced by a LVCSR system on conversational speech allowing the system to focus on speaker differences over a constrained set of acoustic units. Results on the 2001 NIST extended data task show this approach can be used to produce an equal error rate of $< 1\%$.

1. INTRODUCTION

It is well known that using text-dependent speech, such as a fixed password, or text-constrained speech, such as the digits, for a speaker verification task produces much better performance than using text-independent speech. The advantage comes from constraining the speech so that comparisons can be made between how speakers produce certain specific sounds rather than needing to represent and compare speakers over large textual variations. It has been an aim over many years to close the gap between text-dependent and text-independent verification performance by effectively using speech recognition technology. The general idea proposed by several researchers (e.g., [1], [2]) is to use a speech recognition front-end to segment speech into specific units (words, phonemes, etc) and then train and verify speaker models conditioned on these units, thus effectively creating a text-dependent task.

To date, however this approach has not been as successful as simply using a speaker model trained on all the speech, such as with a Gaussian Mixture Model (GMM), for text-independent applications [3]. Some possible reasons for this are: (1) poor consistency of the speech recognition segmentation on unconstrained speech eliminating the specificity of the unit conditioned models, (2) limited data for model training and verification testing when speech is parsed into too many specific units yet too broad models when few units are used. Additionally, unlike in text-dependent applications where training and verification phrases can be designed to provide desired acoustic coverage, in truly text-independent applications the acoustic coverage is uncontrolled. To scientifically study approaches to effectively exploit the combination of speech recognition and speaker recognition for text-independent applications requires the use of large amounts of training and verification speech and high quality transcripts of the speech. Once approaches have shown promising results under these

liberal conditions, researchers can focus on engineering approaches to move to less data and more errorful transcriptions.

To aid in this area of research, the 2001 NIST speaker recognition evaluation introduced a new task called the *extended data* task. The new task allows the use of much more data for training and verifying a speaker than usually used for evaluations with the aim of encouraging the use of techniques and approaches precluded by limited training data. In addition, both manual transcripts and automatic word aligned transcripts were available for use. In this paper we present results from this extended data task where we have used speech recognition segmentation and GMMs to build text-constrained GMMs (TC-GMM) that are capable of producing equal error rates (EER) $< 1\%$ for a large scale text-independent verification task.

The rest of this paper is organized as follows. Section 2 describes the text-constrained GMM system. Section 3 describes the extended data corpus and the experiment paradigm. Section 4 presents experiment results and analysis using the TC-GMM - UBM system trained with various combinations of word sets. In Section 5, we discuss the results and suggest possible future directions.

2. TEXT-CONSTRAINED GMM-UBM SYSTEM

The basic building block of the TC-GMM-UBM system is the Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification system [4]. The GMM-UBM system is a likelihood ratio detector in which we compute the likelihood ratio for an unknown test utterance between a speaker-independent acoustic distribution (UBM) and a speaker-dependent acoustic distribution. To allow maximum flexibility and performance for text-independent applications, both acoustic distributions are modeled by GMMs. The speaker-dependent model is derived from the UBM using MAP adaptation with the speaker's training data. This allows not only a tight coupling between a well trained UBM and a speaker model trained with limited data, but also fast scoring techniques (see [4] for details). Typically 2048 mixtures are used for the UBM.

Standard mel-cepstral coefficients are used as features. After speech activity detection, mel-warped cepstral and delta-cepstral feature vectors are extracted every 10 ms using 20 ms windows. For telephone speech we use 19 mel-filters from the 300-3300 Hz range to compute cepstra and further process the features with RASTA filtering to remove linear channel effects. In the basic

* 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

GMM-UBM system, typically 1-2 hours of speech from 50+ speakers is used to train the UBM and, in previous NIST evaluations, 2 minutes of unconstrained conversational speech is used to train the speaker model.

In the text-constrained GMM system, a speech segmenter is used to segment the incoming speech into pre-defined acoustic units, such as words or phones (see Figure 1). The speech corresponding to each acoustic unit or group of units is then used to train unit-constrained GMM-UBM systems. The UBM for each unit group is trained using speech from a large number of speakers, but only consisting of speech detected as coming from that specific group of units. Similarly, only speech from a specific group of units is used to adapt the speaker model. During verification, only speech from the same group of units as was used to train the GMM-UBM models is used in the likelihood ratio calculation. By constraining the GMM-UBM system to use only specific set of acoustic units we hope to focus in on speaker differences for these particular units (perhaps due to pronunciations) and improve accuracy.

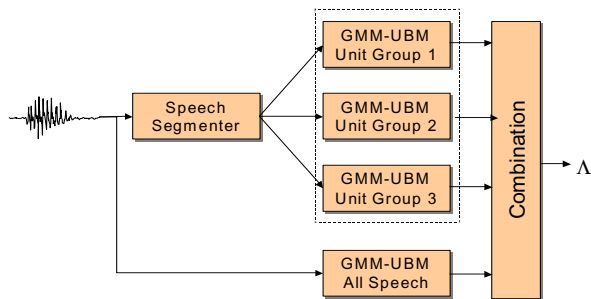


Figure 1 The text-constrained GMM-UBM system. Speech is segmented into acoustic units (e.g., word or phones) and GMM-UBM verifiers are trained and tested using only speech from constrained groups of units.

The speech segmenters often used are speech recognizers, producing word units, or phone recognizers, producing phone units. In this current work we focus on using word units from a speech recognizer¹. In general, however, the segmenter need not segment the speech into pre-defined linguistic units such as words or phones. Since the aim is to compare the same acoustic unit as spoken by different speakers, consistency of segmentation is most important. However, obtaining consistent segmentation by using too-broadly defined acoustic units may undo the effects we wish to achieve by constraining the data.

Recent work [5] has also shown that particular word usage can convey information about speaker identity. By using speech from speaker dependent word or phone n-gram sequences to train GMM-UBMs, we may be able to couple idiolect and acoustic constrains directly. This would allow generating a score of not only whether word usage matches a person but also how well the acoustics of high occurrence words match the speaker. In [6] experiments are presented which demonstrate that simple fusion of idiolect and phonotactic scores with text-constrained GMM scores improves performance.

¹On-going work is focused on using phone recognition segmentations

During verification, different unit-constrained GMM-UBM systems can be used in parallel with their scores combined to produce a final result. The combination could use prior knowledge of the discriminating ability of units (e.g. for phonemes [7]) or be based on trained parameters. In the experiments we examine the use of individual and collections of TC-GMM-UBM systems combined by simple linear combination.

3. THE EXTENDED DATA TASK

For the 2001 NIST speaker recognition evaluation (NIST SRE <http://www.itl.nist.gov/iad/894.01/tests/spk/2001/doc/index.htm>) the extended data task was added with the aim of encouraging the use of techniques and approaches precluded by limited training data. The Switchboard I corpus was used for the task since manual and automatic transcripts for the entire corpus are available². To supply a large number of target and non-target trials and speaker models trained with up to 16 conversations of training speech (~40 minutes), the evaluation consisted of a cross-validation processing of the entire corpus. The corpus was divided into 6 partitions of ~80 speakers each. All trials within a partition involved models and test segments from within that partition only; data from the other 5 partitions were available for background model building or normalization. Within a partition, multiple models per speaker were defined using 1, 2, 4, 8 and 16 conversation sides and all possible permutations of a speaker's conversation sides are used so that all conversations sides for a speaker are used for a target trial. A conversation side consists of nominally 1-2.5 minutes of speech. The trials consist of matched and mismatched handset trials and some cross-sex trails. For the evaluation, NIST supplied speaker model training lists and index files indicating which models were to be scored against which conversation sides for each partition.

4. EXPERIMENTS

Experiments were conducted using the extended data setup. Except as noted, results reported below are for partition 1 with UBMs trained using data from partitions 2-6. The experiments examine three areas: performance using individual words, selection and performance of groups of words, and combination of individual word, word group and all-speech systems.

4.1 Individual Words

To examine the performance of individual words we did the following. We first found the words that occurred in > 70% of the conversation sides, so as to focus on words that are likely to occur for all speakers. There are 50 words that met this criterion (**and, I, that, yeah, you, just, like, uh, to, think, the, have, so, know, in, but, they, really, it, well, is, not, because, my, that's, on, its, about, do, for, was, don't, one, get, all, with, oh, a, we, be, there, of, this, I'm, what, out, or, if, are, at**). These 50 words represent only 0.2% of the entire lexicon, but 54% of all word occurrences. We next built and tested TC-GMM-UBM systems for these individual words for the 8 conversation training

² The manual transcripts were supplied by the Institute for Signal and Information Processing (ISIP). The automatic transcripts were supplied by Dragon Systems and have a word error rate of 20.8%.

condition. In Figure 2 we show the sorted EER for these words (words are in the order listed above).

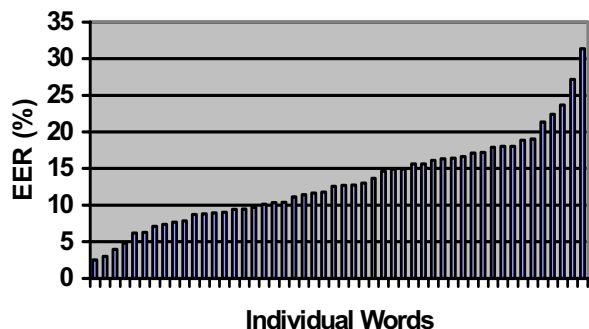


Figure 2 EER for individual word with >70% conversation coverage

It is particularly interesting to note that a single word, **and**, with an average total duration of 4.3 s per conversation side, produced an EER of 2.5%, compared to using all the speech to train and test the GMM-UBM which produces an EER of 1.3%. As might be expected, there is a relatively high negative correlation (-0.65) between EER and the occurrence of the word since high occurrences implies more speech is available for training and testing.

4.2 Word Groups

In the next set of experiments we examined jointly using groups of words. The first set of words, used in the NIST 2001 evaluation, was selected in an ad-hoc fashion looking at word frequency and desirable phonetic content (such as nasals or strong steady-state vowels). This 11 words set, referred to as the NIST set, consists of **uh** (7), **um** (7), **in** (9), **know** (9), **not** (11), **on** (13), **my** (12), **uh-huh** (12), **some** (13), **one** (16), **I'm** (19); where the EER for the individual words is shown in parentheses.

In contrast to the ad-hoc NIST set, we wanted to test if we could use the individual word EER as a way to select a good word set. Using the top 11 words based on EER given in section 4.1, our second word set consists of (**and**, **I**, **that**, **yeah**, **you**, **just**, **like**, **uh**, **to**, **think**, **the**). We will refer to this as the min EER set.

In Figure 3 we present a DET plot showing results for the two word sets and the system using all speech. The two word sets are performing comparably in the EER region, but the min EER set is performing much better in the low false-alarm region. We also see the min EER and all speech systems are performing about the same even though the min EER system is using only a small fraction of all the available speech.

4.3 Combination of Systems

In the last set of experiments we examined different ways of combining systems to produce better performance. The first combination was to fuse the scores from the individual word systems and compare the result to the system using the words jointly for the NIST word set. The fusion was a simple linear combination with equal weights. In Figure 4 we show a DET plot comparing these two systems for the 8 conversation training

condition. The jointly trained and tested system performs significantly better than the fused system. One factor in this could be the extra amount of speech available to the joint set for both training and verification decisions. The joint set had an average 26 seconds of speech per conversation, versus between 1.3-4.3 seconds for each of the individual words. Of course, it is possible that more sophisticated fusion of the individual scores could substantially improve the fusion results.

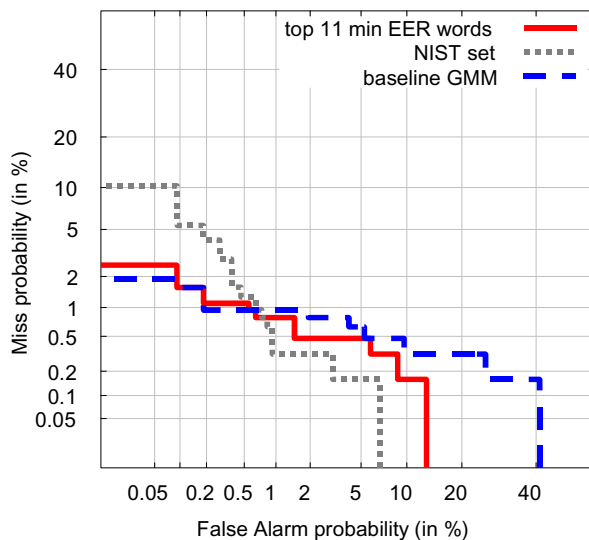


Figure 3 DET plot for the NIST and min EER word sets and the all-speech system on the 8 conversation training condition.

The second combination experiment was to combine the joint NIST word set system with the system trained using all speech. Again, the combination was a simple equal-weight linear combination. One interpretation of this combination is as a back-off system wherein the all-speech system acts as a back-off score when there is too little information for a good decision with the text-constrained system. This can actually be significant in cases in which the particular words in the set do not occur in a test utterance.

The DET plot comparing the text-constrained, the all-speech and the combined system performance is shown in Figure 5. These results are for the 8 conversation training condition and come from the complete cross-validation set. From these results we see that the text-constrained system is performing better than the all-speech system and that the combination produces significant improvement³.

Lastly, in Figure 6 we present EER as a function of the number of training conversations from the complete cross-validation set. Error bars indicate the 95% confidence intervals. The results in this plot are for the GMM-UBM using all-speech, the text-constrained GMM-UBM using the NIST word set, and the linear combination of the two. The combination of the two systems produces statistically significant improvement for all but the 8 and 16 conversation training conditions⁴. As expected, adding

³ Because we are using all splits for this DET, these results are different than those shown in Figure 3.

⁴ The result for the 16 conversation condition is less reliable due to many fewer trials than the other conditions.

more training sessions improves performance, most notably in going from one to two training conversations. From two conversations and greater we also see that the text-constrained system is outperforming the all-speech system. For the 8 conversation training condition, the combined system gives an EER of 1%. When we apply score normalizations, such as H_{norm} and T_{norm} , prior to combining the systems, we can reduce this EER to 0.65%.

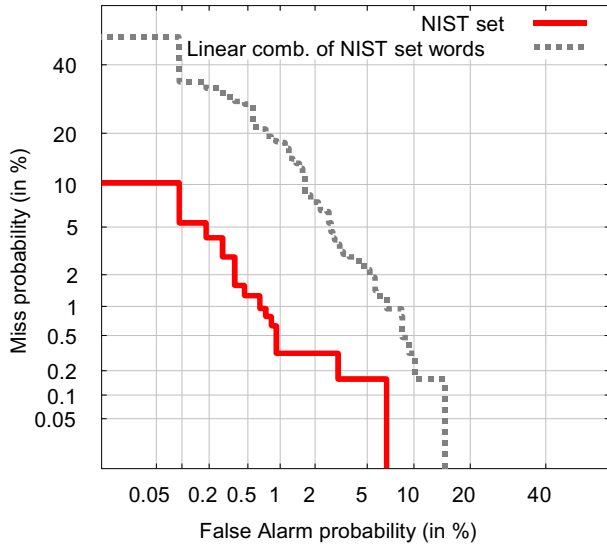


Figure 4 DET plot comparing the fused individual words and joint word systems for the NIST word set on the 8 training conversation condition.

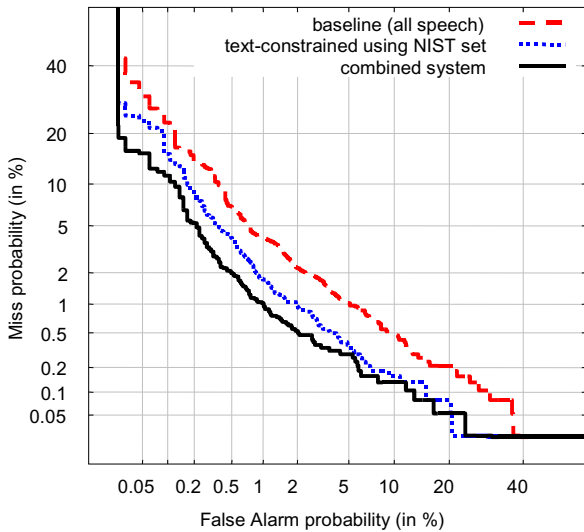


Figure 5 DET plot of all-speech, joint NIST word set and combination of the two systems for complete cross-validation on 8 conversation training condition.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented results of using a text-constrained GMM-UBM system to improve speaker verification performance for a text-independent task. The work demonstrated

on the NIST extended data task that using knowledge about the spoken text could produce low error rates by focusing only on limited acoustic units in the speech. Further, we found that a simple combination of the text-constrained and all-speech systems significantly improves performance

Future work will focus on better selection of word groups and using speaker-dependent word groups. Additionally we will examine the use of other acoustic units, such as phones or automatically derived units.

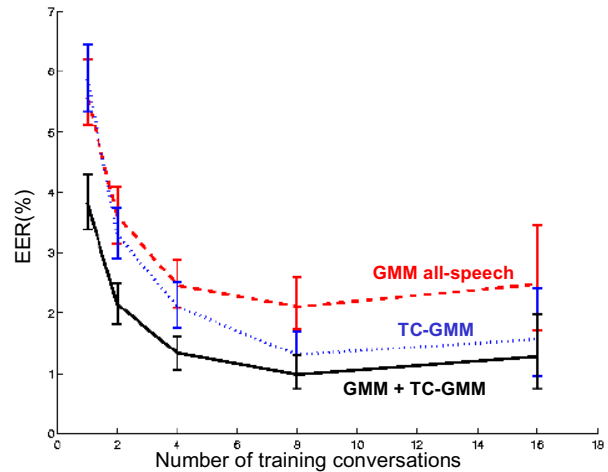


Figure 6 EER as a function of number of training conversation for all-speech, joint NIST word set and combination of the two systems. Results are from the complete cross-validation set.

6. REFERENCES

- [1] F. Weber, B. Peskin, M. Newman, A. Corrada-Emmanuel and L. Gillick, "Speaker Recognition on Single- and Multispeaker Data," *Digital Signal Processing* 10(1-3): 75-92
- [2] 1997 SRI LVCSR Based Speaker Recognition System, 1997 NIST Speaker Recognition Evaluation Workshop
- [3] G. Doddington, et. al, "The NIST Speaker Recognition Evaluation - Overview, Methodology, Systems, Results, Perspective," *Speech Communication* 31 (2-3), pp. 225-254, 2000
- [4] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models." *Digital Signal Processing* 10(1-3): 19-41
- [5] G. Doddington, "Some Experiments on Idiolectal Differences Among Speakers," <http://www.nist.gov/speech/tests/spk/2001/doc/>, January 2001
- [6] W. Andrews, M. Kohler, and J. Campbell, "Acoustic, Idiolectal, and Phonetic Speaker Recognition," *2001: A Speaker Odyssey Workshop*, pp 55-63, 2001
- [7] J. Eatock and J. Mason, "A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes," *ICASSP*, pp. 133-136, 1994