Speaker Verification using Hidden Markov Models in a Multilingual Text-constrained Framework

Brendan Baker and Sridha Sridharan

Speech and Audio Research Laboratory,
Queensland University of Technology,
GPO Box 2434, George St, Brisbane, AUSTRALIA, 4001.
{bj.baker, s.sridharan}@qut.edu.au

Abstract

This paper expands upon previous work, making use of a multilingual framework for text-constrained speaker verification. The framework attempts to overcome some of the restrictions found with previously developed monolingual text-constrained techniques. Pseudo-syllabic segmentation is used in order to extract regions for the constrained recognition. In this study, a comparison between Gaussian mixture models and hidden Markov models is presented for modelling these syllabic events. Results are presented for the NIST 2004 speaker recognition evaluation corpus. The results suggest that temporal patterns within the frame sequences are present and able to be exploited through use of Markovian modelling. The HMM based system is also compared against a traditional global acoustic GMM-UBM speaker verification system, with encouraging results presented.

1. Introduction

Traditionally, text-independent speaker verification systems have been based around the modelling of global or averaged acoustic characteristics. The most widely accepted approach to performing text-independent speaker verification uses Gaussian Mixture Models (GMMs) to model cepstral features across the entire acoustic space, without incorporation of timing, contextual or linguistic information.

It is well known that in conversational speech, speaker characterising information is not exclusively contained in low-level acoustic features. High-level features such as linguistic content, pronunciation idiosyncrasies and prosodic cues all carry useful speaker characterising information. In recent years, there has seen an increased interest in trying to exploit these higher levels of information for automatic speaker recognition.

Two main approaches to exploiting these higher-levels of information have emerged. The first approach seeks to build systems capable of modelling the high-level features independently of the acoustic features, in the hope that such systems provide complimentary classifications. Prominent examples of this approach include Doddington's experiments on speaker idiolect presented in [1], and the phonetic and prosodic feature based experiments carried out at the SuperSID workshop [2] held in 2003.

The second approach to exploiting high-level features uses the high-level information to aid traditional acoustic techniques. High-level information, such as word labels, can be used to segment and constrain the acoustic modelling in order to facilitate more detailed modelling. Such techniques can also provide insight into which parts of speech (or articulatory events) are most important for speaker discrimination.

Text-constrained speaker verification techniques, such as the work presented in [3] and [4] are examples of systems that try and exploit high-level information to improve acoustic modelling. Although these techniques have produced impressive results, there are considerable shortcomings that make such techniques unsuitable for many applications. The accuracy of the front-end speech recogniser transcription is influenced significantly by the language dependent grammar. This a-priori information could not be used in a language independent (or multilingual) situation. Additionally, the appropriateness of the word selection used for the constrained speaker verification is heavily dependent on the availability of sufficient examples for robust training, and scoring during testing.

In [5], a new text-constrained speaker verification technique was presented that attempts to overcome some of these restrictions. Instead of using monolingual word segmentation, a multilingual framework was proposed based on a broad phone and pseudo-syllabic segmentation process. Using GMMs to model these syllabic events, comparable performance to a standard global GMM system was achieved on the NIST 2003 speaker recognition evaluation corpus. It was also shown that performance could be improved through a selective reduction of the set of events used for modelling and scoring.

The work presented in this paper expands on the study in [5]. Hidden Markov Models (HMMs) are trialed in an attempt to capture temporal patterns in the frame se-

1-4244-0472-X/06/$20.00 © 2006 IEEE

quences within the segmented regions. Section 2 gives an outline of the database and evaluation procedure used throughout the study. Sections 3 and 4 provide a description of the general framework, and a recap of the GMM training procedure previously developed. Following this, a detailed description of the new HMM paradigm and procedures used is presented in Section 5.

Results are presented for the syllable-length framework in Section 6, and comparisons performed between the static (GMM) and temporal (HMM) paradigms for both isolated syllables and fused scores. Results are also compared against a traditional global GMM-UBM system, giving an overall indication of the new developed system's performance.

2. Database and Evaluation

The speaker verification systems developed in this study were evaluated and compared using data from the NIST 2004 Speaker Recognition Evaluation corpus [6]. The evaluation procedure was restructured into three equal sized splits. In this study, only the set of male speakers from the first split of the corpus is used for evaluation. This results in 124 speaker models, and 10794 individual test cases. The remaining data is used in development of appropriate background models. Results are presented for the one conversation side training, one conversation side test condition.

Comparisons of speaker verification performance are achieved through the calculation of the Equal Error Rate (EER) and minimum Detection Cost Function (DCF). These measures are derived from Detection Error Tradeoff (DET) curves. Details of the cost function can be found in [6].

3. Framework

This study continues the work performed in [5] making use of the same pseudo-syllabic framework. The framework facilitates multilingual syllabic segmentation and subsequent model development and evaluation for speaker verification. The overall framework operation is depicted in Figure 1.

The syllabic segmentation is achieved through use of a multilingual phone recogniser. The phone recogniser contains a model set of 20 multilingual phonetic events and was trained on the OGI multi-language telephone speech corpus [7]. Further details on the phone recogniser and its construction can be found in [8].

The 20 phonetic events are subsequently mapped to broad phonetic classes. In choosing the number of broad phonetic classes, a number of factors were taken into consideration. The number of broad phonetic classes chosen dictates the number of syllabic units used in the subsequent speaker modelling. By defining a larger set of broad phonetic classes, less acoustic events are pooled together, which potentially leads to more detailed speaker modelling. Smaller sets, however, ensure that sufficient data is available for training. Also, the accuracy and consistency of the front-end of the framework is somewhat governed by the number of these broad phonetic classes. By using a small number of broad phonetic groupings, more consistent/accurate transcriptions can be expected. In this study, a choice of four broad phonetic classes was made in order to ensure recognition accuracy and consistency at the front-end, whilst at the same time, limiting the syllabic set to a practical size.

The four broad phonetic classes used in this study are given in Table 1. It should be noted that the framework allows for the number of classes and phonetic groupings to be redefined at a later stage in order to add or remove phonetic detail.

BPC	Articulatory Description
c1	Vowels and Dipthongs
c2	Nasals, Liquids and Glides
c3	Fricatives
c4	Stops and Pauses

Table 1: Broad Phonetic Classes

Following the phone recognition and broad phone mapping, the phonetic transcription is subsequently converted into a transcription containing broad class phone-triplets, that act as a representation for the syllabic event. These events are referred to as pseudo-syllabic because they do not necessarily reflect the process expected from true syllabic segmentation and the subsequent boundary information this would provide. The average length of these segmentation units does, however, closely resemble that expected from syllabic segmentation. Throughout the remainder of this study, the set of broad syllabic (or pseudo-syllabic) events is denoted by ψ.

Given that each pseudo-syllable contains three phones, and the number of possible broad phonetic classes is 4, the resulting number of syllables in the set ψ is 64. As stated previously, this small set size ensures that sufficient training data is available for each syllable, a problem that presents itself when using large dictionaries such as those used within word-based segmentation and constrained modelling techniques. It should also be noted that in time stamping each instance of ψ, overlapping windows are used. In contrast to true syllabic segmentation, this method provides more instances of training data for each sequence of three phones. The syllabification process is also illustrated in Figure 1.

After the syllabic segmentation, the boundary information can be used to extract features and train individual

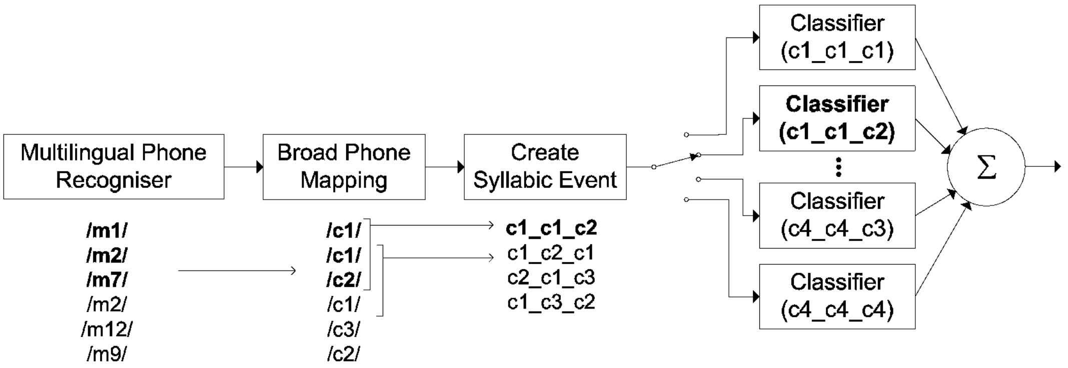

Figure 1: Syllable-length Framework.

classifiers for each syllabic event. In this way, a classifier is available for each syllable and its success can be examined in isolation or in conjunction with other syllabic classifiers. Figure 1 illustrates that the individual classifiers can be combined via a simple summing of the classifier outputs, however more sophisticated combining techniques can easily be applied. The framework also allows for the evaluation of differing feature sets and flexibility in the choice of modelling paradigms.

4. Modelling Syllabic Events using GMMs

The speaker verification framework outlined above was first presented in [5], with initial trials performed using Gaussian mixture models (GMMs) to model the pseudo-syllabic events. The same system structure and feature extraction process was utilised in this study, and models trained for the NIST 2004 evaluation database.

A standard GMM-UBM training methodology [9] is used to train a model for each syllabic event. Short-term cepstral-based feature vectors consisting of 12 MFCC's and 12 corresponding delta coefficients are used. The audio is band filtered between 300Hz and 3.2KHz, followed by an energy based speech activity detection (SAD) process. After features have been extracted, feature warping is also applied [10].

A background set of syllabic GMMs, denoted by $GMM_{BM}(\psi)$, is first built using the segmentation information produced by the front end recogniser for the second and third splits of the evaluation corpus. In order to match the number of free parameters used within the HMM based system presented later in this paper, 80 mixture components are used to model each syllabic event's background model. MAP adaptation is then used to adapt the syllabic models for each speaker j from the UBM using the appropriate target model transcriptions and feature vector sets, producing a set of 64 target GMMs denoted by $GMM_{TM_j}(\psi)$. The speaker specific models are derived from the background models using an iterative MAP adaptation process as outlined in [10].

For each test utterance, a set of scores, one for each syllable, is produced. Each of these scores is calculated by using the relevant frames contained within all instances of that particular syllable in the utterance. The likelihoods of these frames given a target model and a background model are determined, and the ratio of the likelihoods subsequently calculated. This ratio is normalised by the total number of frames belonging to the syllable in the utterance, effectively producing an expected log likelihood ratio (ELLR) for each syllabic event.

A number of options are available for combining the scores from the classifiers belonging to the 64 syllabic events. A previous study [5] showed that giving equal weighting to each classifier is an effective strategy, and has been adopted again in this study.

5. Modelling Syllabic Events using HMMs

The main purpose of this study was to determine whether HMMs could model the segmented regions more accurately than GMMs. Instead of modelling only static averages, it was hoped that HMMs could detect and exploit temporal patterns within the frame sequences. This section outlines the use of HMMs for modelling the segmented syllabic events in an attempt to capture these temporal patterns . The feature set modelled replicates those used previously for the GMM modelling. Each syllabic event, ψ, is modelled, producing a set of HMMs denoted as $HMM(\psi)$.

In the same fashion as the GMM based system, a background set of syllabic HMMs, denoted by $HMM_{BM}(\psi)$, is first built using the segmentation information produced by the front end recogniser for the second and third splits of the evaluation corpus. MAP adaptation is once again used to adapt speaker models for speaker j from the UBM set using the appropriate target model transcriptions and feature vector sets producing a set of 64 target HMMs denoted by $HMM_{TM_j}(\psi)$. The MAP adaptation procedure was chosen over global or indirect adaptation alternatives such as maximum likelihood linear regression (MLLR). Evidence from GMM based speaker verification experiments has suggested that updating components without explicit adaptation data for those components generally degrades performance. By using MAP, unseen components are not modified.

A 7-state left to right state HMM topology is used in this initial study. Seven was thought an appropriate number of states to model each syllabic event given that each phone in the front-end recogniser was defined as a 3-state HMM, and that in building the phone triplets (syllabic units) it could be assumed that the entry and exit states of the middle phone overlap with those of its surrounding phones. Each emitting state distribution is modelled using 16 mixture components, resulting in a total of 80 (16 x 5) mixture components to model each syllabic event. This matches the number of free parameters (or mixture components) described earlier for the GMM system.

In scoring a test utterance, a set of ELLR scores, one for each syllabic event, is calculated. These scores are derived by performing a forced alignment of each isolated syllable in the test utterance against target and background models for that syllable. The forced alignment target and background scores are then collated for each syllable, normalised for the number of frames, and the log likelihood ratio calculated. In the same fashion as the GMM-based system, to combine scores from the 64 syllabic classifiers, equal weighting is given to each classifier.

6. Results

Models for both the GMM and HMM topologies were trained and evaluations carried out. Results were first gathered for each syllabic event tested in isolation. Table 2 gives the EER of the top 10 performing syllables from the GMM system along side EERs achieved using HMMs for the same syllables.

Syllabic Event	GMM EER	HMM EER
c1_c2_c1	14.27%	13.15%
c1_c3_c1	14.99%	14.27%
c2_c1_c2	15.22%	14.16%
c4_c1_c2	15.78%	14.38%
c2_c1_c4	16.44%	14.72%
c2_c1_c3	16.44%	14.60%
c1_c4_c1	16.61%	14.60%
c3_c1_c2	17.00%	14.44%
c4_c1_c3	18.79%	18.50%
c1_c2_c4	19.96%	17.89%

Table 2: Performance comparison of isolated syllabic events using GMMs versus HMMs (*c1* = Vowels and Dipthongs; *c2* = Nasals, Liquids and Glides; *c3* = Fricatives; and *c4* = Stops and Pauses.)

Examining these results, it is quite clear that for each of these syllabic events, using a temporal model via a HMM topology instead of the static GMM results in an improvement in performance. Approximately, a 1% to 2% absolute gain in EER is achieved through substitution of the modelling technique for each of these events. This trend was observed throughout the remaining isolated syllable results, apart from approximately the twenty worst performing syllables. Further analysis of these bottom performing syllables revealed a high correlation between poor performance and low rates of occurrence of the syllable. In other words, the performance of the HMM structure appears to degrade with sparse training data.

In [5], it was shown that the worst performing syllables do not contribute heavily to overall performance once the syllable scores are combined. Therefore, it was thought likely that once scores from all syllables were fused, the HMM based system would continue to outperform its GMM counterpart. To test this hypothesis, the 64 sets of scores from the GMM and HMM systems were combined respectively. Equal weighting was given to each syllabic event's classifier score. Although this weighting technique is almost certainly not optimal for fusing the syllable scores, it allows for the simplest comparison of two modelling paradigms.

Figure 2 shows a comparison of the HMM and GMM systems when all 64 syllable scores were fused using this equal weighting technique. The DET plot shows that the HMM based system is clearly ahead in performance for almost the entire curve. As stated earlier, after observing the isolated syllable results, this performance improvement was expected. The gain achieved through use of the HMM is significant. The HMM based system achieves a 11.5% relative improvement in terms of EER and 10.8% relative minimum DCF over the GMM system.

A comparison against a standard global GMM-UBM speaker recognition system was also performed. This baseline speaker verification system is a traditional GMM-UBM system [9], using the same short-term cepstral-based feature vectors as the syllable-based systems. The UBM is a 512 mixture component Gaussian mixture model. Speaker models are derived from the UBM using an iterative MAP adaptation process [10]. No handset or test segment normalisation techniques are used for fairness of comparison. The baseline system obtains an EER of 13.04% and a minimum DCF value of 0.0392 on the evaluation data.

It should be noted that a second level of speech activity detection that uses a dynamic energy threshold is incorporated into the global GMM system to further remove non-active speech sections of the audio stream. In internal tests, this second level of SAD has shown to have a significant positive effect on performance (approximately 3% absolute in terms of EER for this database).

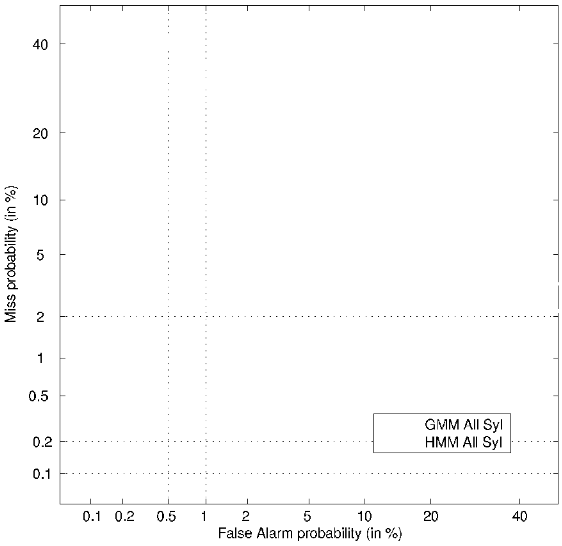

Figure 2: DET plot comparing the GMM and HMM syllable-constrained systems using scores from all 64 syllabic events.

This second layer is not incorporated into the syllable-length framework, as it was thought that the phone recogniser's internal silence model would be able to sufficiently handle these remnant non-active areas of speech. Further analysis of this effect should be performed in the future.

Figure 3 and Table 3 compare the performance of the HMM syllable-length system with the GMM baseline. The DET plot seems to indicate that the HMM system achieves superior performance at most operating points, with fairly equivalent results at the minimum DCF point.

System	EER	Min DCF
GMM baseline	13.04%	0.0392
HMM(ψ)	11.15%	0.0396

Table 3: Performance comparison of baseline GMM system and the syllable-length HMM system

7. Conclusions and Future Work

In this study, a comparison was carried out between the use of GMMs and HMMs to model syllable length units extracted using a multilingual framework for text-constrained speaker verifications. Initial results performing verification on isolated syllabic events suggested that the temporal modelling provided by the HMM paradigm is able to more accurately model the segmented regions than the previously trialed GMMs. This improvement in performance was also observed when the scores from the entire syllable set were combined. The combined HMM

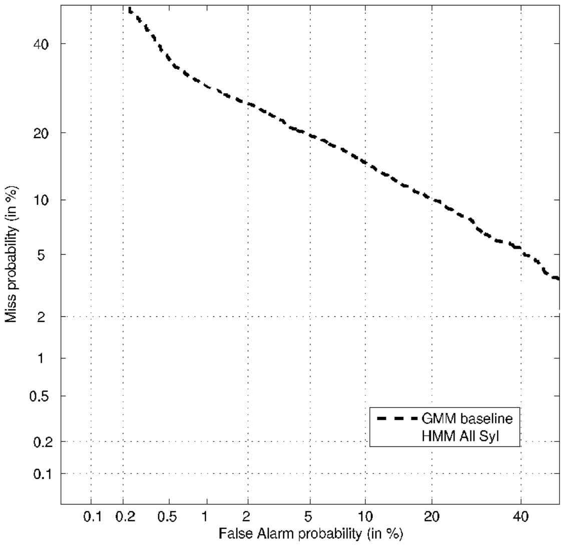

Figure 3: DET plot comparing a state-of-the-art GMM-UBM system and the syllable-length HMM system.

system achieved a relative improvement of approximately 11% in terms of both EER and minimum DCF over the GMM system. These results seem to indicate that speaker specific patterns exist within the frame sequences of the segmented regions, and that the HMM topology is able to exploit these patterns.

When compared against a traditional acoustic GMM-UBM system, the HMM based system gave encouraging results. Improvements were observed for the majority of operating points, with equivalent performance achieved at the minimum DCF point.

Significant improvements are possible in both the front end of the framework and the modelling technique used. There is a significant mismatch between the channel conditions of the corpus used to train phone recogniser and the speaker verification corpus. Effort should be placed on adapting the front-end phone recogniser closer to the target domain. Further experimentation should also be performed in order to optimise HMM topology and training procedure.

The developed framework also allows for the introduction (or reduction) of further phonetic detail into the syllabic units. Further examination of the isolated syllable results may give some clues as to where finer phonetic detail would be most useful. More sophisticated classifier combination techniques, such as neural networks and support vector machines, should also be examined for combining the individual syllabic event classifier outputs.

8. Acknowledgements

This research was supported by the Australian Research Council (ARC) Discovery Grant Project ID: DP0453278. The authors would also like to acknowledge the help of Terrence Martin in building the front-end broad phone recogniser.

9. References

[1] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, Denmark, 2001, vol. 4, pp. 2517–2520.

[2] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003, vol. 4, pp. 784–787.

[3] D. Sturim, D. Reynolds, R. Dunn, and T. Quatieri, "Speaker verification using text-constrained Gaussian mixture models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 1, pp. 677–680.

[4] K. Boakye and B. Peskin, "Text-constrained speaker recognition on a text-independent task," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 129–134.

[5] B. Baker, R. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, Lisbon, 2005, pp. 2429–2432.

[6] M. Przybocki and A. Martin, *"The NIST Year 2004 Speaker Recognition Evaluation Plan"*, http://www.nist.gov/speech/tests/spk/2004/, January 2004.

[7] Y. Muthusamy, R. Cole, and B. Oshika, "The OGI multi-language telephone speech corpus," in *International Conference on Spoken Language Processing*, 1992, pp. 895–898.

[8] T. Martin, B. Baker, E. Wong, and S. Sridharan, "A syllable-scale framework for language identification," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 276–302, April-July 2006.

[9] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 1997, vol. 2, pp. 963–966.

[10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, 2001, pp. 213–218.