Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification

Brendan Baker, Robbie Vogt and Sridha Sridharan

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

10.21437/Interspeech.2005-648

Abstract

This paper examines the usefulness of a multilingual broad syllable-based framework for text-independent speaker verification. Syllabic segmentation is used in order to obtain a convenient unit for constrained and more detailed model generation. Gaussian mixture models are chosen as a suitable modelling paradigm for initial testing of the framework. Promising results are presented for the NIST 2003 speaker recognition evaluation corpus. The syllable-based modelling technique is shown to outperform a state-of-the-art baseline GMM system. A simple selective reduction of the syllable set is also shown to give further improvement in performance. Overall, the syllable based framework presents itself as valid alternative to text-constrained speaker verification systems, with the advantage of being multilingual. The framework allows for future testing of alternative modelling paradigms, feature sets and qualitative analysis.

1. Introduction

In recent years, there has seen an increased interest in the use high-level features of information for automatic speaker recognition [1]. Research has expanded from only using the acoustic content of speech to trying to utilise all levels of speaker specific information in order to achieve superior speaker recognition performance.

For modelling of acoustic features, most text-independent speaker verification systems utilise Gaussian Mixture Modelling (GMM). Most implementations of this technique produce models for the entire acoustic space, without incorporation of high-level contextual or linguistic information. A priori information defining which articulatory events have occurred could provide direct benefits for more detailed modelling. Such information could also provide insight into which parts of the feature space are important for discriminating between speakers. Importantly, segmentation provides a natural framework for determining this association.

Some recent studies have looked at improving performance for text-independent speaker verification by attempting to convert the task into a text-dependent one [2, 3]. This was achieved by constraining the verification process to a limited set of words. By building models for these constrained events, more detailed models were able to be generated. In the study performed by Boakye et al. [3], utterances were pre-segmented at the word level using a state-of-the-art English ASR system. From this, a selection of words were modelled using Hidden Markov Models (HMMs). This technique proved to be quite successful and was found to provide complementary classifications to those obtained using a standard GMM based system.

Although the text-constrained approaches have shown improvements, there are still a number of shortcomings that make them unsuitable for many applications. The accuracy of the ASR transcription is influenced significantly by the language dependent grammar. This a-priori information could not be used in a language independent (or multi-lingual) situation. Additionally, the appropriateness of the word selection is also dependent on the availability of sufficient examples for robust training, and scoring during testing.

Experiments were performed by Martin *et al*. [4, 5], using the syllable as a unit for segmentation and subsequent model development for language identification. The systems outlined in these studies used a multi-lingual broad phone set to build psuedo-syllabic events (broad phone triplets). Unlike the word based techniques used for speaker verification, this segmentation process was not language dependent. The studies showed that not only did the segmentation allow for more detailed modelling, but qualitative analysis could also be performed examining the discriminative ability of particular articulatory events.

This paper presents a preliminary study examining the use of the syllable-length segmentation process as outlined in [4, 5], for the task of speaker verification. The appropriateness of the syllable as a unit for segmentation is explored through the use of the framework and subsequent modelling using Gaussian mixture models. Sections 2 and 3 give an outline of the evaluation procedure used throughout the study as well as describing the baseline system. This is followed by an explanation of the new syllable length framework in Section 4.

Section 5 presents results for modelling across broad phonetic events, as well as presenting results showing the effect of speech activity detection on verification performance. Finally, Section 6 presents results for the final syllable based modelling. Comparisons are made showing the effect of increasing model complexity and some simple optimisations are achieved through selective pruning of the syllables used for verification.

2. Database and Evaluation

The speaker verification systems developed in this study were evaluated and compared using data from the NIST 2003 Speaker Recognition Evaluation Extended Data Task corpus [6]. The evaluation data is a subset of the Switchboard-II Phase 2 and 3 corpora. The NIST 2003 EDT evaluation procedure [6] was restructured to include a new one conversation side training length condition (derived from the existing four side training condition). In this study, only the set of male speakers from splits 1-4 of the evaluation corpus are used. The remaining data is used in development of appropriate background models.

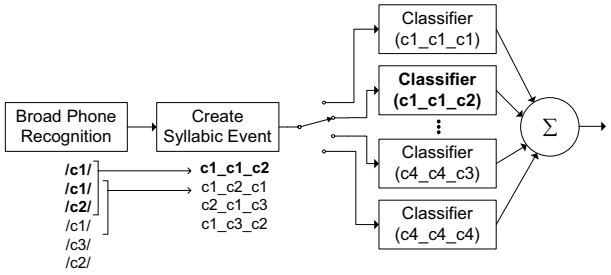

Figure 1: Syllable-length Framework.

Comparisons of speaker verification performance in this study are achieved through the calculation of the equal error rate (EER) and minimum detection cost function (DCF). These measures are derived from detection error trade-off (DET) curves. Details of the cost function can be found in [6].

3. Baseline System

A baseline acoustic speaker verification system was used as a benchmark for comparisons in this study. The baseline system used is a standard GMM-UBM system [7] using short-term cepstral-based feature vectors consisting of 12 MFCC's and 12 corresponding delta coefficients. 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 [8].

The UBM is a 512 mixture component Gaussian mixture model. Speaker models are derived from the UBM using an iterative MAP adaptation process [8]. For the experiments carried out in this study, no handset or test segment score normalisation techniques were used. The baseline system obtains an EER of 14.2% and a minimum DCF value of 0.0460.

4. Broad Phonetic and Syllabic Framework

In this study, the syllable is used to provide a segmentation framework for sub-sequent model development for the speaker verification task. The syllabic segmentation is achieved by recognising Broad Phonetic Classes (BPC) using a multilingual broad phone recogniser, and then concatenating these phones to form pseudo-syllabic events.

The overall system operation is depicted in Figure 1. The front end phone recognition system produces a sequence of broad phonetic events. In this study, four broad phonetic classes are used in order to limit the total number of possible syllables. These are: *c1* - Vowels and Dipthongs; *c2* - Nasals, Liquids and Glides; *c3* - Fricatives; and *c4* - Stops and Pauses. Further details on the construction of this multilingual broad phone recogniser can be found in [5]. After phone recognition, 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. Throughout the remainder of this study, the set of broad 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. This small set size ensures that sufficient training data is available for each syllable. It also should be noted that in time stamping each instance of ψ, overlapping windows were 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 illustrated in Figure 1.

After the syllabic segmentation, the boundary information can be use to extract features and train individual 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. The framework also allows for the evaluation of differing feature sets and flexibility in the choice of modelling paradigms.

5. Modelling of Broad Phonetic Events

Before attempting to model a speaker's acoustics for the syllabic events, a preliminary set of experiments were performed to examine the performance when a GMM-UBM modelling technique is used to model the acoustic space for each of the four broad phonetic categories (BPC).

This experiment effectively segments the feature space into four parts, and as such, it was thought appropriate to reduce the number of mixture components used to model each segment UBM to a quarter of that used by the baseline system (128 mixture components). The same features (12 cepstral coefficients and deltas) were used as previously defined for the baseline system. This broad phone based GMM-UBM system is denoted as $GMM_{128}(BPC)$ throughout the rest of this study.

Classification results were obtained for the individual phonetic categories using the evaluation database. Table 1 shows the performance of the four phonetic categories when tested in isolation. It can be seen from these results that there are significant differences in the performance of each of the broad phones. Of particular note, is the poor performance of the stop/short pause model (*c4*). Examination of the number of frames assigned to each phonetic class at testing gave some insight into what may be causing the poor performance of the *c4* class. Over two times as many frames were assigned to the *c4* than any other class. This high frame count suggested that this phone model was recognising more than just stops and short pauses (as was intended), but instead was also recognising frames from long periods of non-active speech or silence (segments that contain no speaker characterising information).

BPC	EER	Min DCF
c1	15.60%	0.0533
c2	17.39%	0.0564
c3	21.14%	0.0700
c4	38.01%	0.0992

Table 1: Performance of the four broad phone models

Rather than relying on the phone recogniser to throw out these frames, a speech activity detector can be employed. In order to try and improve the systems performance, an energy based speech activity detector was utilised. The majority of frames removed through the activity detection process were found to be from *c3* and *c4* phonetic groupings. For the broad phone class *c4*, 65% of the frames originally assigned to that phonetic category were removed by the activity detector. Table

BPC	EER	Min DCF
c1	15.25%	0.0499
c2	13.86%	0.0462
c3	13.63%	0.0522
c4	25.19%	0.0822

Table 2: Performance of the four broad phone models with SAD

2 compares the performance of the individual phonetic categories after SAD. It can be clearly seen that an improvement in performance has been achieved as a result of SAD. The most dramatic of improvements was as expected, found with *c4*, however the other phonetic categories also improved as a result of the activity detection.

A further experiment was also carried out using the broad phone classifiers to compare two classifier combination techniques. The first combination technique calculates an overall expected log likelihood ratio score for the test utterance. The frame count is used to weight each classifiers individual score before summing. The summed score is then averaged over the entire frame count, resulting in an *overall ELLR* score. The second technique gives *equal weighting* to each of the classifiers, no matter how many frames were used in the scoring process for that classifier. The overall score is calculated as the sum of the individual classifier ELLR scores.

Table 3 shows the resulting EER using the two classifier fusion techniques with and without SAD. From this table, it can be seen that for the non-SAD system, catastrophic fusion occurred using both combination techniques. This highlights the importance of the activity detection process. The equal weighting technique, however, seems to have reduced the degradation caused by the poor performance of *c4* in the non-SAD system. The equal weighting method has the ability to de-emphasise those classifiers that have poorer performance but larger frame counts.

The difference in performance between the two fusion techniques for the activity detected task isn't as great, however the equal weighting technique still outperforms the ELLR method. A considerable improvement over the individual broad phone classifiers is achieved through fusion. The combined SAD system using equal weights gave an EER of 12.77% compared to 13.63% for the best performing individual phonetic group, *c3*.

Fusion Technique	EER No SAD	EER With SAD
Overall ELLR	26.98%	13.75%
Equal Weighting	20.68%	12.77%

Table 3: Performance of combined broad phonetic system with and without speech activity detection

6. Modelling of Broad Syllabic Events

The main purpose of this study was to examine the usefulness of broad syllabic events as a segmental unit on which to model speaker characteristics. The time boundaries of the phones were used to segment the data into the pseudo-syllabic events as described in Section 4. The features extracted were once again modelled using a GMM.

The fact that the syllabic events overlap, effectively means that overall, more frames are available for training the classifiers (approximately 3 times as many frames). With this in mind, it was thought that 32 mixture components would be a good starting point for modelling the syllabic events. This system is denoted as $GMM_{32}(\psi)$

The 32 mixture component syllable based models were trialled on the evaluation corpus with individual syllables achieving quite varied performances. Results ranged from around 14.6% EER ($c2\_c1\_c2$) to 50% EER ($c3\_c3\_c3$) with an average EER of 27.17%. The fact that some of the individual syllable classifiers obtained comparable performance to that of the acoustic baseline was a pleasing intial result.

Further experiments examined 64 and 128 mixture component models, designated as $GMM_{64}(\psi)$ and $GMM_{128}(\psi)$ respectively, with a similar spread of individual syllable performances observed. Table 4 summarises the performance measures for the 32, 64 and 128 mixture component systems. Increasing the number of components seems to improve performance in the low false alarm region, but has the opposite effect in the low miss probability region. In terms of minimum detection cost function, better performance in the low false alarm region is of greater importance, however the choice is dependent on the application. For the remainder of the study, the 64 mixture component system $GMM_{64}(\psi)$ was used.

# Mix	EER	Min DCF
32	13.29%	0.0475
64	13.63%	0.0458
128	13.63%	0.0451

Table 4: Performance comparison for different numbers of mixture components for syllable based modelling

In an attempt to optimise the system, it was decided to reduce the set of syllabic events using EER as a selection criteria. As an experiment, rather than using all 64 syllabic events, only the classifications from the top 32 and top 16 syllables, in terms of EER, were combined together. The equal weighting technique was used to combine the scores. Figure 2 shows a comparison of the performance of these two pruned systems against the complete syllable set system, and the 512 mixture component acoustic baseline system. Table 5 summarises the performance measures obtained for each of these systems. Although this optimisation technique is simple, the results indicate that it has been quite effective. Reducing the syllable set gave improvements in both EER and minimum DCF measures. The best performing syllabic system is that using only the top 16 syllables, with an 11% relative improvement in EER achieved over the acoustic baseline.

As mentioned previously, this pruning of the syllable set was only an initial and fairly crude optimisation. Other factors that influence the discriminative ability, such as rate of occurrence of the events, should be taken into consideration. Further investigation into the internal structure of the better performing syllabic events should also be carried out.

7. Conclusions and Future Work

In this study, a syllable length framework was proposed for the task of speaker verification. The syllabic segmentation outlined was achieved through the use of a multilingual broad phone recogniser. Broad phones boundaries were then used to provide boundaries for broad pseudo-syllabic events (broad phone triplets).

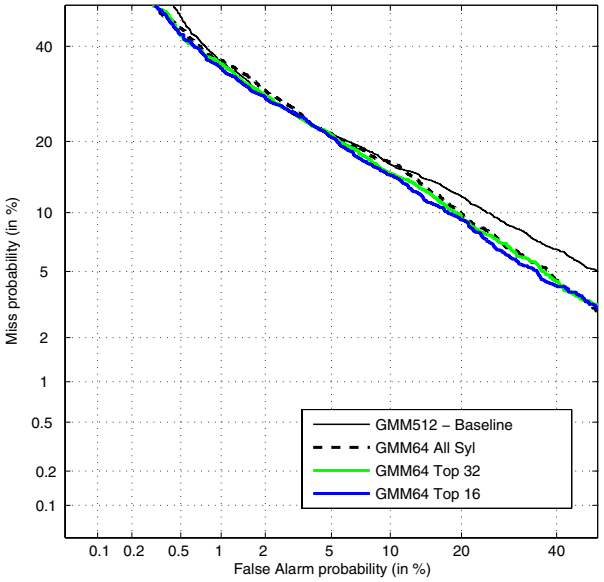

Figure 2: DET plot comparing the complete and reduced set syllable based systems with the acoustic baseline

Verification System	EER	Min DCF
GMM_{512}	14.20%	0.0460
$GMM_{64}(\psi)$	13.63%	0.0458
$GMM_{64}(\psi_{top32})$	13.29%	0.0452
$GMM_{64}(\psi_{top16})$	12.59%	0.0443

Table 5: Performance comparison of the optimised syllable systems

Initial experiments modelling the acoustic data for the four broad phonetic categories highlighted a need for speech activity detection as a pre-processing step. These experiments also showed that comparable performance to a state-of-the-art acoustic baseline could be achieved through modelling of just the vowel/dipthong and nasal/glide phonetic events.

Two techniques were examined for the combining of individual classification scores. The equal weighting technique was found to give superior performance to that of the overall ELLR. This seemed due to its ability to de-emphasise poor performing classifiers that had high frame score counts.

Several experiments were carried out using the syllabic segmentation and modelling framework. These experiments indicated that 32 mixture components are sufficient to build accurate GMM based models for the syllabic events. Individual performances of the syllable based classifiers ranged from 14.5% EER to 50% EER.

A simple selective reduction of the syllable set was also shown to give further improvement in performance. The best performance was obtained when only the top 16 syllables (in terms of individual EER) were combined to obtain an overall score. This system obtained a 11% relative improvement over the acoustic baseline, and a substantial reduction in minimum DCF.

Overall, the syllable-length framework has presented itself as a promising mechanism for speaker speaker verification. The preliminary experiments carried out in this study have shown the syllable-based technique to be a suitable replacement for previously presented word constrained speaker verification, overcoming many of the shortcomings found with such systems.

The framework presented allows for considerable future development. Other modelling paradigms should be trialled, such as HMMs, which may be able to capture more of the temporal characteristics of speech. The framework also allows for the analysis of different feature sets, such as pitch and energy based features.

Not only can other modelling paradigms and feature sets be applied, but the framework also allows for further qualitative analysis of which syllabic events provide the most discriminative information. Judicious selection of the syllabic events and more sophisticated classifier combination techniques will most likely lead to further improvements.

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 supplying the front-end broad phone recogniser.

9. References

[1] 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*, vol. 4, 2003, pp. 784–787.

[2] 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*, vol. 1, 2002, pp. 677–680.

[3] 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.

[4] T. Martin, E. Wong, B. Baker, M. Mason, and S. Sridharan, "Pitch and energy trajectory modelling in a syllable length temporal framework for language identification," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 289–296.

[5] T. Martin, B. Baker, E. Wong, and S. Sridharan, "A syllable-length framework for language identification," *in print Computer Speech and Language*, 2005.

[6] M. Przybocki and A. Martin, *"The NIST Year 2003 Speaker Recognition Evaluation Plan"*. http://www.nist.gov/speech/tests/spk/2003/doc/, February 2003.

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

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