The QUT NIST 2004 Speaker Verification System: A fused acoustic and high-level approach

Michael Mason, Robbie Vogt, Brendan Baker and Sridha Sridharan

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

Abstract

The trend towards including both acoustic and high level speech features in speaker recognition systems is addressed with the presentation of the speaker verification system developed by QUT for the NIST 2004 Speaker Recognition Evaluation. The system presented is a fusion of five subsystems including acoustic, lexical, phonetic, prosodic and durational speech features and focuses on utilising the high level feature sets in reduced training set conditions. The performance of the system on the development data resources available in the NIST SRE are presented to demonstrate the effectiveness of the fused system approach.

1. Introduction

Speaker recognition has been the subject of significant research for over a decade and for the majority of this time researchers have focused on the use of acoustic features to discriminate one speaker from another. This has lead to significant success, with verification Equal Error Rates (EER) of around 2% in matched conditions. While this performance is impressive, the use of only acoustic features is limiting because they suffer direct degradation in the presence of noise and environmental mismatch. Due to this, more recent research directions have broadened to incorporate so called high level features in an effort to make speaker recognition systems more robust.

When considering speech signals, and particularly looking at speech features which can be used to discriminate between speakers, a hierarchy can be constructed which relates different types of features to each other, based on their complexity and accessibility. At the lowest level, acoustic features like Mel Frequency Cepstral Coefficients (MFCCs) can be extracted from speech waveforms through the direct application spectral transforms and filterbanks. At the next level prosodic features, such as pitch and energy contours and speaking rates, are more difficult to extract from speech signals but are based on theoretical constructs which are independent of acoustic noise or channel mismatch. Finally at the topmost level it is possible to consider speech in terms of the phones, words and sentences which are again, in theory, separated from the acoustic environment. It is important to note that while higher level features are theoretically independent of the acoustic environment in which they are collected, the mechanisms used to collect them may still be effected by the environment - the prime example of this is that speech recognisers have reduced accuracy in noisy environments, so while the features we're referring to don't intrinsically change between a clean and noisy environment, the accuracy with which we can extract or estimate them is still effected.

The estimation of appropriate high-level speech features, the effective modelling of their characteristics and their performance when used to discriminate speakers has been the subject of a growing body of literature as well as being the focus of the 2003 National Institute of Standards and Technology (NIST) Speaker Recognition Evaulation (SRE) Extended Data Task (EDT). From these sources a number of simple generalisations can be drawn. First, significant support for the use of high-level features has been found. Second, some useful features have been identified, but there remains potential for better feature selection and modelling techniques to improve on the currently proposed systems. Third, currently the modelling of high-level features requires longer training utterances than acoustic features, due primarily to the slower relative rate of individual observations. Finally, high-level features will not provide a replacement for acoustic features, but rather the most promising systems appear to be those which combine acoustic features and multiple high-level speech features.

The annual NIST SRE has recognised these trends and based on the experience of the 2003 EDT has broadened the scope of the evaluation in 2004 by removing the explicit boundary between acoustic speaker verification and speaker verification using high-level features. In order to address this change of focus QUT has expanded its acoustic speaker verification approach to one which incorporates a number of high-level features and paper provides an overview of this system. Section 2. provides an overview of the system architecture and the data sources used throughout its development. Sections 3. through 7. describe in detail each of the subsystems developed as part of the combined system. Section 8. details the output fusion mechanism used to combine the scores from each of the five subsystems and provides the results of the system testing on the design data set.

2. System Overview

The speaker verification system developed by QUT for the NIST 2004 SRE focused on combining acoustic and higher level features to improve the performance of the system in all conditions. Of particular focus was the utilization of high-level features for situations with less training

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

data then had typically been used to train high-level feature models previously.

The system devised comprised five independent subsystems along with output score fusion. The five subsystems were identified as;

1. Acoustic Subsystem
2. Lexical Subsystem
3. Phonetic Subsystem
4. Prosodic Subsystem
5. Durational Subsystem

Each of which is described in the following sections.

The NIST 2004 SRE defined a series evaluation conditions discriminated by amount of training data provided and the length of the test utterance. The primary condition was identified as having a single conversation side for training and testing (A single conversation side contained 5 minutes of speech including silence). In addition to the primary condition, QUT was involved in two additional conditions using three and eight conversation sides for training, respectively, and a single side for testing.

In order to develop the whole system and to ensure meaningful and fair evaluation of its performance the marshalling and use of different databases in appropriate ways was imperative. Of key importance was the segregation of data into subsystem specific training and development data, fusion development and evaluation data. While subsystems could share training and development data, this data set and the fusion and evaluation data sets were mutually exclusive. Table 1 sets out the different data sources and identifies which set was used in and table 2 further describes which data sources were used in the training and development of each of the sub systems.

3. Acoustic Speaker Verification

The acoustic subsystem was based on the now familiar Gaussian Mixture Model (GMM) with Universal Background Models (UBM) topology pioneered by Reynolds (1997) and included handset type and test-segment normalisation (HNorm and TNorm) (Auckenthaler, Carey, and Lloyd-Thomas 2000). The unique characteristics of the system are described in the following sections addressing acoustic feature extraction, UBM and speaker model training and score normalisation

3.1. Feature Extraction

Prior to extracting appropriate acoustic features the utterances were processed to detect speech activity. For this the NIST SPQA package (National Institute for Standards and Technology 1994) was utilised, as per previous NIST SREs. The speech activity decisions were then refined further using a energy-based detection process.

The raw acoustic features extracted from the utterances were 12 Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein 1980) at a rate of one observation every 10 ms. In order to increase the robustness of the acoustic features to variations in transducers and environmental noise the raw observations were further processed using the feature mapping (Reynolds 2003) and feature warping (Pelecanos and Sridharan 2001) techniques.

3.1.1. Feature Mapping

Feature mapping is a handset type normalisation technique which operates directly in the feature-space to reduce the apparent mismatch of observations caused when speech is sampled with different transducers. This normalisation is achieved by learning a set on non-linear transforms from known context specific feature spaces to a neutral, context independent, feature space, and is derived directly from the adaption of context specific GMMs from a context neutral GMM.

Having adapted context dependent GMMs from the context independent GMM, all future utterances are scored against the context dependent models to identify their most likely context, and the mapping from that context into the neutral context in applied to all the observations from that utterances by applying the mapping

$$\mathbf{y} = \left(\mathbf{x} - \mu_i^{CD}\right)\left(\mathbf{\Sigma}_i^{CD}\right)^{-1}\mathbf{\Sigma}_i^{CI} + \mu_i^{CI} \qquad (1)$$

where i is the index of the top scoring component in the most likely context dependent model.

The feature mapping models were generated based on data driven clusters produced from the Switchboard II, phases 2 & 3 (landline) and NIST 2001 & 2002 (cellular) data. Approximately 850 segments were used for this training, totaling almost 20hrs of speech. The data driven clustering process was seeded based on the MIT handset detection labels for these databases and information provided in the Sphere headers. The observations used for clustering did not use feature warping, as feature warping is itself designed to provide some environmental normalisation, instead Cepstral Mean Substraction (CMS) was included in the acoustic feature extractor to compensate for environmental noise.

3.1.2. Feature Warping

Feature warping attempts to improve the acoustic features robustness to channel mismatch and noise by transforming the individual elements of the feature vector to conform to a normal distibution (Pelecanos and Sridharan 2001). On a observation by observation basis this is achieved by transforming the current feature vector's elements to their relative z-score based on their position in a sorted list of the set of elements from the surrounding window of the feature stream. The length of the window used trades off the accuracy of the relative position estimate against the introduced processing delay. For the observations extracted here a 5 second feature warping window was used.

3.2. UBM and Speaker Model Training

Gender dependent UBMs were trained based on the Switchboard II, phase 2 & 3 (landline) and NIST 2001 & 2002 (cellular) data with 512 mixture components for the single sided training condition and 1024 mixture components for the three and eight sided training conditions.

	Subsystem Training and Development	Fusion Development and Evaluation
Switchboard II, phase 1 (swb2p1)	X	
Switchboard II, phase 2 & 3 (swb2p23)	X	
NIST 2001 & NIST 2002 (cellular data only)	X	
MIT Handset Detection Data for swb2p23	X	
BYBLOS Transcripts of Switchboard II, phases 2,3 & 4	X	
OGI Multi-lingual database	X	
NIST 2003 EDT - splits 1–4		X

Table 1: Data sources assigned to subsystem training and development, fusion development and evaluation sets.

	Acoustic	Lexical	Phonetic	Prosodic	Durational
Switchboard II, phase 1 (swb2p1)	X				
Switchboard II, phase 2 & 3 (swb2p23)	X		X	X	X
NIST 2001 & NIST 2002 (cellular data only)	X				
MIT Handset Detection Data for swb2p23	X				
BYBLOS Transcripts of Switchboard II, phases 2,3 & 4		X			
OGI Multi-lingual database			X		

Table 2: Data sources used in subsystem development.

The UBMs were trained with the Expectation Maximisation (EM) algorithm, initial model seeding was performed with a vector quantisation algorithm (Pelecanos, Myers, Sridharan, and Chandran 2000). The speaker models were trained using iterative mean only MAP adaption with a "relevance factor" $\tau = 8$ (Vogt, Pelecanos, and Sridharan 2003). For efficiency the top-N component, Expected Log-Likelihood Ratio (ELLR) method was used, with N=5 for all conditions.

3.3. Score Normalisation

In addition to the feature mapping and feature warping of the acoustic features two output score normalisation techniques were included; Handset Normalisation and Test segment Normalisation (Auckenthaler, Carey, and Lloyd-Thomas 2000).

HNorm is designed to remove the biases in a trained speaker model's response to differing telephone handset types. This is achieved by measuring the response of each speaker model to a set of impostor trials from each handset context. In testing, each score is then normalised by the mean and standard deviation of the handset context of the trial. The contextual segments used for determining HNorm statistics were also used to train the feature mapping models.

As the name suggests, TNorm compensates for variations within a test utterance such as length, noise levels and linguistic content. In a similar fashion to HNorm, each score is normalised based on the statistics of an impostor population. These statistics are obtained through scoring the test segment against a set of independently trained impostor speaker models. For each gender and training length combination 100 speaker models were trained on Switchboard-II, Phase 1 data.

4. Lexical Speaker Verification

Lexical speaker verification aims to exploit the differences of speakers personal lexicon (idolect) — a concept pioneered by Doddington (2001). To do this speaker specific n-gram models are trained from the enrollment data and test utterances are scored against a claimants model and a universal background.

4.1. N-gram Modelling

The likelihood estimate of a given observed n-gram, k, is estimated from the n-gram model, m, using

$$l_m(k) = \frac{C_m(k)}{\sum_{n=1}^{N} C_m(n)} \tag{2}$$

where $C_m(k)$ is the frequency counts of the n-gram token in the training data. For a sequence of n-grams the log-likelihood ratio of the speaker model, Θ, to the background model, Ω, is given by

$$\Lambda = \frac{\sum_k \left(w(k) \cdot \log\left(l_\Theta(k)/l_\Omega(k)\right)\right)}{\sum_k w(k)} \tag{3}$$

where the weighting function, $w(k)$, is calculated from the count of token k in the test utterance and a discounting factor, $d \in [0, 1]$, from

$$w(k) = C(k)^{1-d} \tag{4}$$

Equation 2 represents the maximum likelihood training criteria for the *n*-gram model and provides reasonable estimates of *n*-gram likelihoods for scenarios when the training data set is large enough to approximate the observation ensemble. When enrolling individual speakers however this sets an unreasonably high requirement for training

data. In order to reduce this data requirement the MAP estimation solution for multinomial densities proposed by Lee and Gauvain (1996) has been adapted to the *n*-gram models proposed by Baker, Vogt, Mason, and Sridharan (2004).

The MAP solution adapts the speaker specific *n*-gram model from the background *n*-gram model. The speaker specific counts can be express as

$$\tilde{C}_m(k) = C_m(k) + \alpha C_\Omega(k) \quad (5)$$

where $C_m(k)$ is the *n*-gram count from the speaker specific training data, $C_\Omega(k)$ is the background model *n*-gram count and α is the adaption relevance factor on the interval $[0, 1]$.

4.2. Subsystem Configuration

The QUT lexical subsystem uses the fused output of two lexical speaker verification systems; one modelling Uni-grams and the second modelling Bi-grams. The background models were trained from the Byblos ASR transcriptions of Switchboard II, Phases 2,3 & 4.

Uni-gram and Bi-gram scores were fused at the model output level using a Multi-Layer Perceptron (MLP) in order to produce an overall lexical score.

5. Phonetic Speaker Verification

Phonetic speaker verification exploits the personal variations in pronunciation and individuals' tendency to vocalise a variety of phones in similar ways. The original proposal by Andrews, Kohler, Campbell, and Godfrey (2001) was to analyse the phone labels produced by an open-loop speech recogniser and build an *n*-gram model similar to the multinomial *n*-gram models proposed for lexical speaker recognition by Doddington (2001). In addition to modelling the phone transcriptions produced by a recogniser trained on the same language as the test utterance, Andrews, Kohler, Campbell, Godfrey, and Hernndez-Cordero (2002) also demonstrated that phone transcriptions produced by recognising a test utterance in a number of 'off'-langauge phone recognisers provided further complementary information. The process of recognising an utterance using multiple recognisers is referred to as phonetic refraction.

As with the *n*-gram models produced by the ML training process initially proposed for lexical speaker recognition, model sparsity issues required a large amount of speaker specific data in order to be reliable. By using the MAP adaption of speaker models from a well trained background model these data requirements can be similarly reduced (Baker, Vogt, Mason, and Sridharan 2004).

5.1. Subsystem Configuration

The QUT phonetic subsystem developed for NIST 2003 SRE used MAP adapted tri-gram modelling of phone sequences from six languages. The Open-Loop Phone Recognisers (OLPR) used to produce the streams were trained on the OGI multi-lingual database and the six languages used were English, German, Hindi, Japanese, Mandarin and Spanish.

The background tri-gram model was trained on the Switchboard II, Phases 2 & 3 corpus and speaker specific models were MAP adapted using a relevance factor $\alpha = 0.01$. The tri-gram model scores for each language were fused, using an MLP, at the output level to produce an overall phonetic score.

6. Prosodic Speaker Verification

Prosodic speaker verification exploits interpersonal variations in pitch and volume patterns to discriminate between speakers. The prosodic speaker verification subsystem implemented by QUT for the NIST 2004 SRE was motivated by the successful prosodic features and modelling techniques proposed by Adami and Hermansky (2003).

The prosodic features are estimated by representing the pitch contour of an utterance as a series of straight line segments and identifying each of these segments as single prosodic event. Each event can then be characterised in terms of the slope of the pitch contour, the slope of the associated energy contour and by its duration. Unvoiced events are quantised simply in terms of their duration.

As with the lexical and phonetic modelling techniques it is possible to model the *n*-gram distributions of these prosodic features to produce a background model and individual speaker models. The system presented here used bi-gram models and trained the background models on the data from Switchboard II, Phases 2 & 3. Unlike the lexical and phonetic model the use of higher order MAP adapted speaker models did not provide any performance improvement over the ML estimated models produced directly from the speaker enrollment data.

7. Durational Speaker Verification

The durational speaker verification system presented here was designed to discriminate speakers based of the variation present in the rate with which individuals produce different phones and is based on the system proposed by Kajarekar, Ferrer, Venkataraman, Sonmez, Shriberg, Stolcke, Bratt, and Gadde (2003).

The durational features of interest were estimated by quantising the state occupancy durations of three state HMM phone recognition transcripts (Kajarekar, Ferrer, Venkataraman, Sonmez, Shriberg, Stolcke, Bratt, and Gadde 2003). Speaker specific histogram models of these features were MAP adapted from background models trained using the QUT OLPR transcripts of Switchboard II, Phase 2 & 3.

8. Output Fusion

The output scores from the five subsystems were combined to produce a single verification decision by a multi-layer perceptron (MLP) trained using the LNKnet software package (MIT Lincoln Laboratory 1994). A separate MLP was trained for each of the three evaluation conditions.

In addition to the five subsystem output scores, which were normalised for mean and standard deviation, a sixth input, representing the gender of the target, was used for fusion.

A k-fold training and validation process was used during the development of the MLPs and the output class priors weighted to account for the bias in expected evaluation conditions. Splits 1-4 of the NIST2003 SRE EDT were used to train the fusion.

Accepted after full paper review

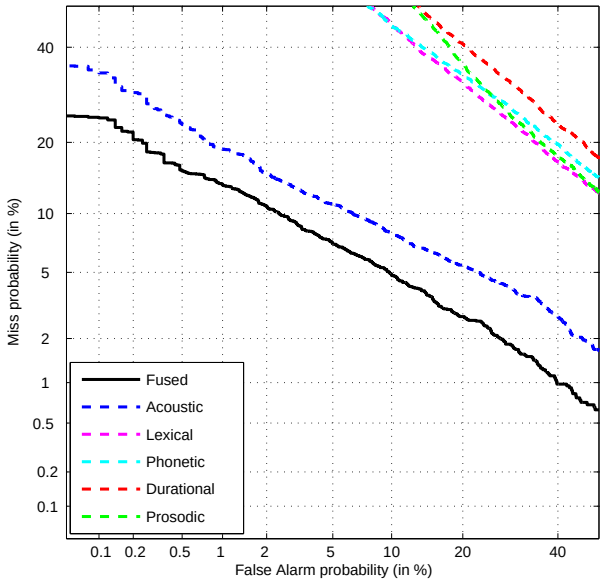

Figure 1: DET for single sided training condition.

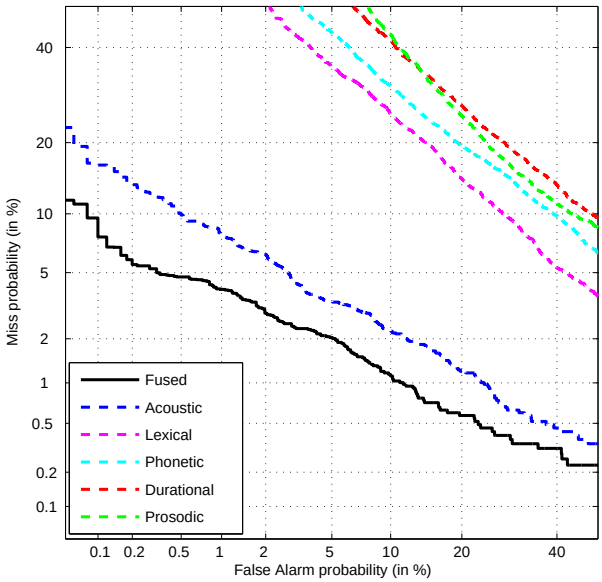

Figure 2: DET for three sided training condition.

9. Results and Discussion

Table 3 summarises the EERs and the minimum Detection Cost Function (DCF) values for the individual and fused scores and figures 1–3 depict the DET curves of the systems. From these results we can see that the fused systems achieved relative EER improvements of between 28% (single sided training) to 43% when using eight conversation sides for training. The improvements in minimum DCF were even more impressive ranging from 26% to 65% — this is highlighted in figure 4.

Figure 5 depicts the system performances of the acoustic only systems, systems based on the fusion of high-level features only and the full fusion of acoustic and high-level features. From the graph the effect of using high-level features can be considered in terms of training data requirements. By comparing the performance of the eight sided high-level system to the single sided acoustic system it can be seen that high-level features require much more training data then acoustic features to perform at comparable EER.

More positively though we can compare the performance of the eight sided acoustic system to the three sided fully fused system and can see that the fusion has provided us with approximately equivalent performance with less then half the data requirement.

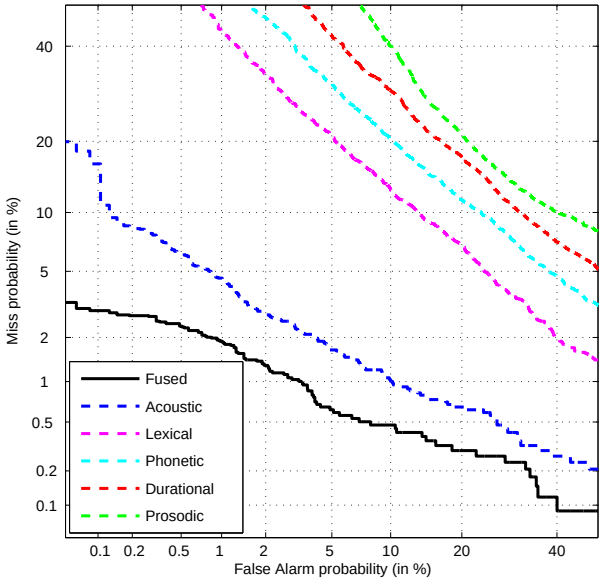

Figure 3: DET for eight sided training condition.

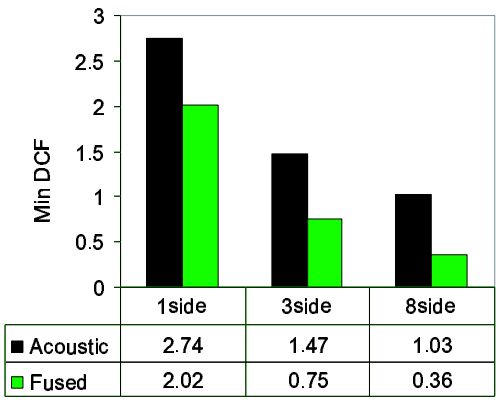

Figure 4: Relative minimum DCF values for the one, three and eight sided acoustic and fused systems.

10. Conclusion

This paper has presented a description of the QUT speaker verification system as it was developed for the NIST 2004 SRE. From the development data the performance of the system showed a relative improvement of 26% over the purely acoustic performance when using only a single conversation side for training.

The fused system trained with three conversation sides performed at approximately the same level as an acoustic system trained with eight sides, supporting the claim that the fusion of high-level features and acoustic features can provide true reductions in the training data requirements of speaker verification systems.

Accepted after full paper review

	1conv		3conv		8conv	
	DCF	EER	DCF	EER	DCF	EER
Acoustic	0.0274	0.0878	0.0147	0.0376	0.0103	0.0254
Lexical	0.0856	0.2617	0.0689	0.1702	0.0519	0.1144
Phonetic	0.0891	0.2775	0.0766	0.1968	0.0653	0.1480
Prosodic	0.0991	0.2726	0.0966	0.2244	0.0933	0.2057
Durational	0.0973	0.3076	0.0888	0.2344	0.0793	0.1847
Fused	0.0202	0.0634	0.0075	0.0255	0.0036	0.0145

Table 3: Individual subsystem and fused overall system performance

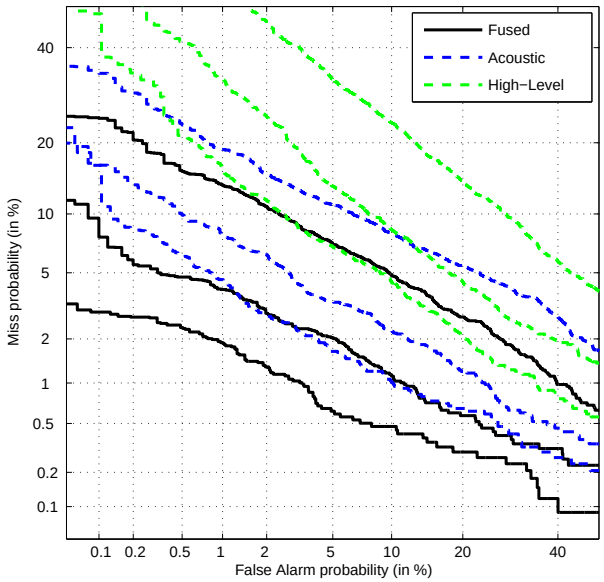

Figure 5: DET for one, three and eight sided training conditions, comparing the acoustic only system, the fused high-level features and the fusion of all features.

11. Acknowledgments

This research was supported by the Office of Naval Research (ONR) under grant N000140310662.

References

Adami, A. G. and H. Hermansky (2003). Segmentation of speech for speaker and language recognition. In *Proc. of Eurospeech*, pp. 841–844.

Andrews, W., M. Kohler, J. Campbell, and J. Godfrey (2001). Phonetic, idiolectal and acoustic speaker recognition. In *2001: A Speaker Odyssey*, pp. 55–63.

Andrews, W., M. Kohler, J. Campbell, J. Godfrey, and J. Hernndez-Cordero (2002). Gender-dependent phonetic refraction for speaker recognition. In *Proc. of ICASSP*, Volume 1, pp. 149–152.

Auckenthaler, R., M. Carey, and H. Lloyd-Thomas (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing 10*, 42–54.

Baker, B., R. Vogt, M. Mason, and S. Sridharan (2004). Improved phonetic and lexical speaker recognition through MAP adaptation. In *Proc. of Odyssey*, pp. 91–96.

Davis, S. and P. Mermelstein (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics Speech and Signal Processing ASSP-28*, 357–366.

Doddington, G. (2001). Speaker recognition based on idiolectal differences between speakers. In *Proc. of Eurospeech*, Volume 4, pp. 2521–2524.

Kajarekar, S., L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, and R. R. Gadde (2003). Speaker recognition using prosodic and lexical features. In *Proc. IEEE Speech Recognition and Understanding Workshop*.

Lee, C. and J. Gauvain (1996). Bayesian adaptive learning and MAP estimation of HMM. In *Automatic speech and speaker recognition : Advanced topics*, pp. 83–107. Boston, Massachusetts, USA: Kluwer Academic Publishers.

MIT Lincoln Laboratory (1994). LNKnet pattern classification software. http://www.ll.mit.edu/IST/lnknet.

National Institute for Standards and Technology (1994). SPeech Quality Assurance (SPQA) package (version 2.3). http://www.nist.gov/speech/tools/.

Pelecanos, J., S. Myers, S. Sridharan, and V. Chandran (2000). Vector quantization based gaussian modelling for speaker verification. In *Proc. of International Conference on Patteren Recognition*, Volume 3, pp. 298–301.

Pelecanos, J. and S. Sridharan (2001). Feature warping for robust speaker verification. In *2001: A Speaker Odyssey*, pp. 213–218.

Reynolds, D. (1997). Comparison of background normalization methods for text-independent speaker verification. In *Proc. of Eurospeech*, Volume 2, pp. 963–966.

Reynolds, D. (2003). Channel robust speaker verification via feature mapping. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 2, pp. II–53–6.

Vogt, R., J. Pelecanos, and S. Sridharan (2003). Dependence of GMM adaptation on feature postprocessing for speaker recognition. In *Proc. of Eurospeech*, pp. 3013–3016.

Accepted after full paper review