Phonetic and Lexical Speaker Recognition in Reduced Training Scenarios

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

Abstract

High-level features have been shown to be effective for speaker recognition when large amounts of training data are available for speaker model training; however the feasibility of such long lengths of training for many applications is questionable. This paper describes the evaluation of phonetic and lexical *n*-gram based speaker recognition systems for reduced training lengths. Maximum likelihood modelling is compared to a recently developed MAP adaptation modelling technique. Results obtained using a restructured NIST 2003 Speaker Recognition Extended Data Task corpora indicate that significant gains in performance for both the phonetic and lexical based speaker recognition can be achieved through use of this adaptive modelling technique. The results from fusion experiments also demonstrated that the individual system improvements obtained for the high-level features translated into an overall performance gain when used along side traditional acoustic techniques. The MAP adapted modelling process was shown to extend the usefulness of high-level features to shorter training lengths, with results indicating that even when only one conversation side was used for training, the high-level systems provide complementary classifications and improved recognition performance.

1. Introduction

In recent times, automatic speaker recognition research has expanded from utilising only the acoustic content of speech to examining the use of higher levels of speech information, commonly referred to as high-level features. These high-level features refer to information such as linguistic content, pronunciation idiosyncrasies, idiolectal word usage, prosody and speaking style. This change in research focus has been motivated by the belief that these high-level features can provide complementary information, and that the estimation of these features is more robust to changes in acoustic conditions.

A promising direction in high-level feature research has been the use of *n*-gram based models to capture speaker specific patterns in the phonetic and lexical content of speech. Doddington (2001) performed an important initial study into the use of the lexical content of speech for speaker recognition, and introduced an *n*-gram based technique for modelling a speaker's idiolect. This direction in research was continued by Andrews, Kohler, Campbell, and Godfrey (2001), who used similar *n*-gram based models to capture speaker pronunciation idiosyncrasies through analysis of automatically recognised phonetic events.

The research of Andrews *et al*. and Doddington showed word and phone *n*-gram based models to be quite promising for speaker recognition, however, good performance was really only achieved when excessive lengths of training data were provided. Reduced training scenarios resulted in under-trained models, providing little or no benefit in classifying the speaker. Consequently, the practical applicability of these techniques was greatly restricted.

Initial research into the use of high-level features has focused on characterising high-level knowledge sources and defining new feature sets. Now that several useful features for speaker recognition have been identified, an obvious next step is to further develop the classification and modelling techniques, and to analyse and improve performance of these systems under restricted testing and training conditions. In particular, techniques need to be developed to improve performance under limited training data situations.

Baker, Vogt, Mason, and Sridharan (2004) introduced an adaptive training technique for *n*-gram based speaker models. Applying a *Maximum A Posteriori* (MAP) estimation solution and adapting the *n*-gram speaker models from a background model resulted in significant gains in performance. The experiments on the NIST 2003 Extended Data Task (NIST 2003 EDT) database demonstrated that when compared against traditional Maximum Likelihood (ML) models, the same performance could be obtained with half the amount of training data by using the MAP adapted models.

The introduction of the MAP adaptation technique provided significant improvements, however, neither this technique, nor any phonetic or lexical modelling technique, has been thoroughly tested using less than 10 minutes of training speech for each speaker. For a range of potential applications of the technology, 10 or more minutes of training data is infeasible.

This paper examines the value of these high-level approaches, and the adaptive modelling approach for *n*-gram based features using reduced lengths of training speech of around 2 to 3 minutes. To facilitate this evaluation, the NIST 2003 EDT was restructured to include two new training length conditions: one conversation side, and three conversation sides.

Section 2 of this paper describes the phonetic speaker recognition system including a description of the front-end phone recognition process and both the ML and MAP mod-

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

Accepted after full paper review

elling techniques tested. Section 3 briefly describes the lexical speaker recognition system that was also developed. Results for these individual high-level systems using reduced training lengths are provided in Section 4, along with a description of the database and testing procedure used.

Fusion experiments were carried out, in order to determine the complementary nature of the classifications provided by the high-level features in the reduced training scenarios. The phonetic and lexical classifiers (both ML and MAP) were fused with those obtained using a traditional acoustic speaker recognition system. Results and details of these fusion experiments are outlined in Section 5.

2. Phonetic *n*-gram speaker recognition

The phonetic speaker recognition system is derived from the system described by Andrews et al. (2002) - where speaker specific information is captured by analysing sequences of phone labels produced by open-loop phone recognisers. Andrews' approach was to compare relative frequencies of *n*-gram tokens, allowing for the capturing of recognised phonetic patterns of individual speakers. The process used phone streams produced by multiple language open-loop phone streams. The transcriptions produced by 'off'-language recognisers are known as *refracted* phone transcriptions (Andrews et al. 2002). These refracted streams of phones are capable of providing speaker information which is complementary to the true language's phonetic transcription.

2.1. Front-end phone recognition

The front end of the phonetic speaker recognition system consists of six independent open-loop phone recognisers. Three state HMMs were trained for each phone using the OGI phonetically transcribed multi-lingual corpus (Muthusamy, Cole, and Oshika 1992), consisting of six different languages: English, German, Hindi, Japanese, Mandarin and Spanish. The speech recordings were parameterised by calculating 12th order Perceptual Linear Predictive (PLP) coefficients (Hermansky 1990) and energy, plus their corresponding delta and acceleration coefficients.

The six independent phone recognisers were used to decode all speaker testing and training data. The transcriptions were post processed to include *start* and *end* tokens around speech utterances. An utterance was defined as the sequence of phones occurring between two periods of silence.

2.2. Maximum likelihood speaker modelling

A baseline phonetic speaker recognition system was developed using the maximum likelihood criterion for speaker model training. The speaker models consist of simple multinomial distributions of the frequencies of phonetic *n*-gram tokens. When scoring, each phone transcription is tested against Phonetic Speaker Models (PSM) and a Universal Background Phonetic Model (UBPM) using a traditional likelihood ratio test (LRT).

The likelihood estimates for a model m, are estimated from the training data using

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

where k represents an *n*-gram token, and $C_m(k)$ is the frequency count of the token k in the training data.

To verify speaker *m*, the test segment score is calculated as the log likelihood ratio (LLR) of the speaker likelihood to background likelihood and is given by

$$\Lambda = \frac{\sum_k (w(k) \cdot log[l_m(k)/l_{ubm}(k)])}{\sum_k w(k)}, \qquad (2)$$

where $w(k)$ is a weighting function for token k, based on the count $C(k)$ of the token in the test segment and a discounting factor, d. The weighting function is calculated as

$$w(k) = C(k)^{1-d} \qquad (3)$$

The discounting factor, d, has permissible values between 0 and 1. For $d = 0$ there is no discounting. For $d = 1$ there is absolute discounting, meaning a particular *n*-gram token will contribute the same increment to the total score regardless of the number of times that *n*-gram token occurs.

Doddington (2001) and Andrews et al. (2002) found that improved performance could be achieved by ignoring infrequent *n*-grams due to the inaccuracies in modelling these infrequent events. To this end, the baseline system was developed to take a pruning threshold c_{min} as an additional parameter. *N*-grams that occur less than c_{min} times in the background training data are ignored in the scoring process.

After test segment scores are calculated for each phone stream, the scores are fused together to generate an overall score for the test segment. In the baseline system created for this study, a Multi-layer Perceptron (MLP) neural network architecture implemented using the *LNKnet* pattern classification software (Massechusetts Institute of Technology Lincoln Laboratory 2004), was used to fuse the individual scores.

2.3. MAP adapted modelling

In the baseline system, the ML criterion (Equation 1) was used to train each PSM using the set of *n*-gram frequencies extracted from the model training data. In (Baker et al. 2004), we proposed the use of an adaptive training process in order to combat data sparsity issues and improve the robustness of the models. This was achieved by tying prior information about a model's parameters into each speaker's PSM. The Bayesian learning framework and MAP estimation algorithms provided us with methods to do this.

Lee and Gauvain (1996) outlined a MAP estimation solution applicable to multinomial densities which was adapted for this work. The MAP solution used for the n-gram frequencies can be expressed as

$$\widetilde{l}_m(k) = \frac{\widetilde{C}_m(k)}{\sum_{n=1}^{N} \widetilde{C}_m(n)} \qquad (4)$$

The MAP re-estimated count is calculated using the speaker specific n-gram frequencies from the training data, along with the hyper-parameters $v(k)$. This re-estimated count can be expressed as

$$\widetilde{C}_m(k) = C_m(k) + v(k) - 1, \qquad (5)$$

which optimally combines the n-gram frequency counts from the training data with prior knowledge of the model parameter distributions expressed in $v(k)$. If we take the UBPM as an estimation of the *a priori* n-gram frequency expectations, $v(k)$ becomes simply a weighted expression of the UBPM. By imposing the condition

$$v(k) = \alpha C_{ubm}(k) + 1, \tag{6}$$

Equation 5 becomes

$$\widetilde{C}_m(k) = C_m(k) + \alpha C_{ubm}(k), \tag{7}$$

where α is an adaptation weight in the range [0, 1]. In the limit of no adaptation data, this reverts to the background model, while converging to the ML solution for infinite training data. This MAP adaptation solution ensures numeric stability in the models and effectively cancels the need for *ad hoc* pruning thresholds (Baker et al. 2004).

3. Lexical speaker recognition

A word-based speaker recognition system was created based on that described by Doddington (2001). The approach uses word *n*-gram statistics gathered from ASR transcriptions of the speech as features for the speaker recognition process.

3.1. ASR transcriptions

For this study, transcriptions produced by the BBN realtime Byblos system were used. Before *n*-gram statistics were gathered, the transcriptions were pre-processed to add *start* and *end* tags to sentence boundaries based on pauses in the speech.

3.2. Speaker modelling

Speaker modelling and scoring is performed in the same manner as the phonetic technique described in Section 2 substituting phonetic *n*-gram tokens for word *n*-gram tokens and using only a single token stream (English word transcriptions). Both ML and MAP adapted modelling techniques were evaluated for the lexical system in this study.

4. Experiments

4.1. Database

The developed speaker recognition systems were evaluated and compared using data from the NIST 2003 Speaker Recognition Evaluation Extended Data Task corpus. (For further information see (National Institute of Standards and Technology 2003)). The evaluation data is a subset of the Switchboard-II Phase 2 and 3 corpora (Linguistic Data Consortium 1997). The aim of this paper was to examine the performance of the *n*-gram based high-level features in reduced training scenarios. To this end, the NIST 2003 EDT evaluation procedure was restructured to include two new training length conditions: one conversation side, and three conversation sides. These correspond to approximately 2.5 minutes and 7.5 minutes of training data respectively. The training and testing lists for these new conditions were derived from the existing four conversation side lists. Modifications were also made to the evaluation to include more impostor trials.

During the development of both the phonetic and lexical systems, a development data set consisting of splits 1-4 of the NIST 2003 EDT evaluation data was used. This development data set was used to tune the various parameters of the recognition systems, and to train the neural network used for fusing results from multiple phone streams in the phonetic speaker recognition system. Once the systems were calibrated, overall results were obtained using the remaining evaluation splits (5-10).

4.2. Phonetic System Performance

The phonetic speaker recognition system was evaluated using both ML and MAP adapted models. Our previous experiments (Baker et al. 2004) have shown that when using ML models, best performance is obtained for triphone models with absolute discounting ($d = 1$) and a pruning threshold of $c_{min} = 500$. For our MAP adapted models, a MAP weighting of $\alpha = 0.01$ was used along with absolute discounting. No pruning is necessary for the MAP adapted models.

Results were obtained for the newly defined one and three conversation side training length conditions. Figures 1 and 2 show detection-error tradeoff (DET) curve comparisons of the ML and MAP systems for three and one side training conditions respectively. In Figure 1 it can be seen that a vast improvement over the ML model is achieved when the MAP adapted models are used. Using the adapted models gave a 34% relative improvement in terms of equal error rate (EER). This improvement trend is continued in the one side training condition and is illustrated by Figure 2. For this condition, using the MAP adapted models decreased the EER from 41% to 28%, equivalent to a 30% relative improvement.

4.3. Lexical system performance

Similar tests were performed on the lexical speaker recognition system. For the lexical system the best performance was provided by bigram models. Maximum likelihood and MAP adapted models were compared with the following parameters:

- ML models: $d = 1$, $c_{min} = 0$
- MAP models: $d = 1$, $c_{min} = 0$, $\alpha = 0.01$

Results were obtained for the one and three conversation side training length conditions. Figure 3 demonstrates the improvement gained by using MAP adapted models for the three side condition. Using MAP adapted models gave a 18% relative improvement in EER over ML modelling. For the one side training length condition (see Figure 4), a 13% relative improvement was gained through the use of adapted models.

5. Fusion with acoustic system

The lexical and phonetic system results indicate that significant performance gains can be made in reduced training length scenarios through adaptive modelling. The improvements gained in the individual performance of both

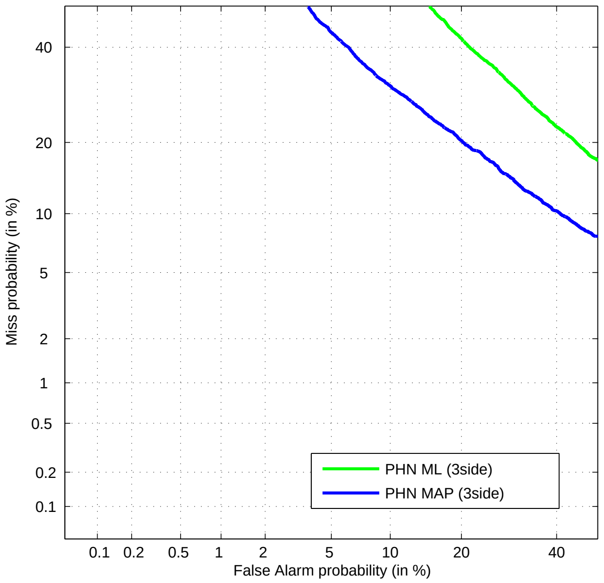


Figure 1: *DET plot comparing Phonetic ML and MAP modelling techniques for the three conversation sides training length condition.*

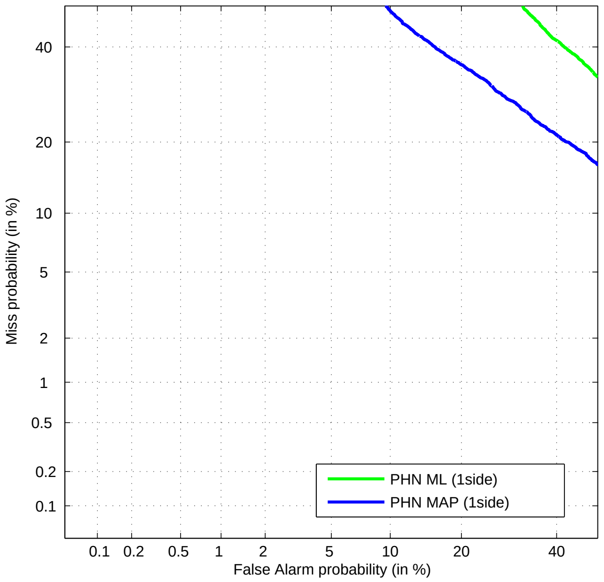


Figure 2: *DET plot comparing Phonetic ML and MAP modelling techniques for the one conversation side training length condition.*

the phonetic and lexical speaker recognition systems, however, are of little use unless they also translate into a performance gain when used in conjunction with an acoustic system.

In reduced training scenarios particularly, it is expected that most of the classification strength will be provided by the acoustic methods. The value of high-level features, therefore, is in the complementary information they provide. To this end, a set of fusion experiments were performed in order to evaluate the complementary nature of the phonetic and lexical speaker classifications in such conditions.

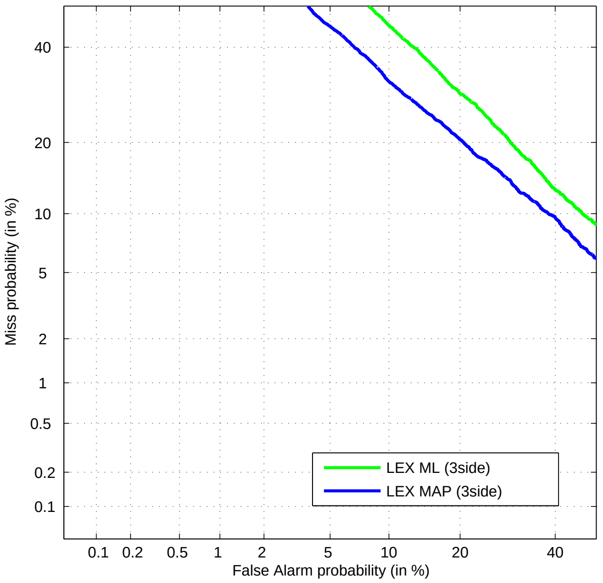


Figure 3: *DET plot comparing Lexical ML and MAP modelling techniques for the three conversation sides training length condition.*

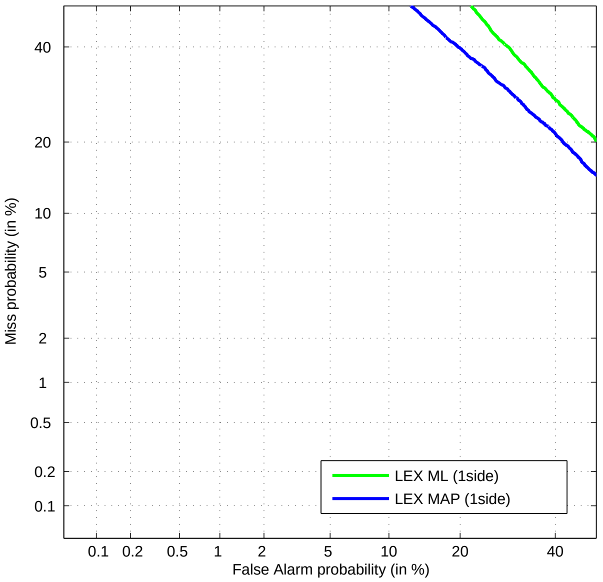


Figure 4: *DET plot comparing Lexical ML and MAP modelling techniques for the one conversation side training length condition.*

5.1. Acoustic system

The acoustic speaker recognition system used is a standard GMM-UBM system (Reynolds 1997) using short-term cepstral-based feature vectors consisting of 12 MFCC's and 12 corresponding delta coefficients. Before the features are extracted, 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 (Pelecanos and Sridharan 2001).

The UBM is a 512 mixture component Gaussian mixture model. Speaker models are derived from the UBM using an iterative MAP adaptation process (Pelecanos, Vogt, and Sridharan 2002). The verification score for each test utterance is calculated as the expected log-likelihood ratio of the claimant and the UBM. For the experiments carried out in this study, no handset or test segment score normalisation techniques were used.

5.2. Fusion

Fusion of the GMM-UBM acoustic system and high-level feature systems was performed using a Multi-layer Perceptron (MLP) neural network. Two fusion combinations were trialled for each training length condition. The first fused system, denoted in the plots by *AC+HL(ML)*, consisted of the acoustic system scores combined with those obtained from the tuned ML phonetic and lexical systems. The second combined the classifications from the acoustic system with the MAP adapted high level systems, and is denoted by *AC+HL(MAP)*.

The MLP training and testing was performed using the LNKnet pattern classification software package (Massechusetts Institute of Technology Lincoln Laboratory 2004), using the development splits (1-4) for training, and the remaining splits for evaluation. Three inputs, consisting of the classification scores from the acoustic, phonetic and lexical systems, were fed into a single hidden-layer MLP. Simple mean and variance normalisation was performed to the features before fusion. Additionally, the priors were adjusted to specifically minimise the detection cost function (DCF) criterion specified for the NIST evaluation (National Institute of Standards and Technology 2003).

Figure 5 compares the DET curves for a baseline acoustic system, and the two fusion combinations for the three side training length condition. It can be seen that there is generally improved performance for the AC+HL(ML) system over the acoustic baseline, with a 12.3% relative improvement in EER achieved. This is with the exception of the high false alarm region, where performance degrades and is behind that of the acoustic system. It can also been seen that the fused system incorporating MAP adapted models gave an even larger gain in performance. The curve shows that the AC+HL(MAP) system is consistently ahead of both the acoustic and AC+HL(ML) fused system, with a 23.8% relative improvement in EER achieved over the acoustic baseline.

Similar trends in performance were found when the training length was further reduced to one conversation side. Figure 6 depicts the DET curves for the acoustic baseline and the two fused systems for the one side training length condition. The AC+HL(ML) system only gave a slight improvement over the acoustic baseline. Significant gains, however, were achieved when using the MAP adapted fused system. For the AC+HL(MAP) system, a 13.6% relative improvement in EER was achieved over the acoustic system. The minimum detection cost function (DCF) was also measured for each of the systems. In Figure 7, a comparison of the minimum DCF values obtained for the acoustic system and the two fusion combinations is given for both the one and three side training conditions. For both training conditions, the fused systems gave better minimum DCF results than the acoustic baseline. The best performing system in terms of minimum DCF, was the AC+HL(MAP) MAP adapted fused system, with 21.8% and 11.7% relative improvements over the baseline for the three side and one side training length conditions respectively.

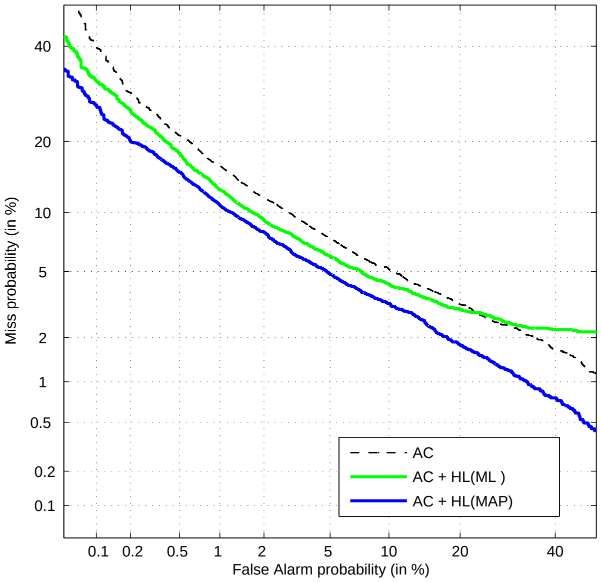


Figure 5: *DET plot for the three side training condition comparing a) a baseline acoustic system b) a fused acoustic and ML high-level system c) a fused acoustic and MAP adapted high-level system.*

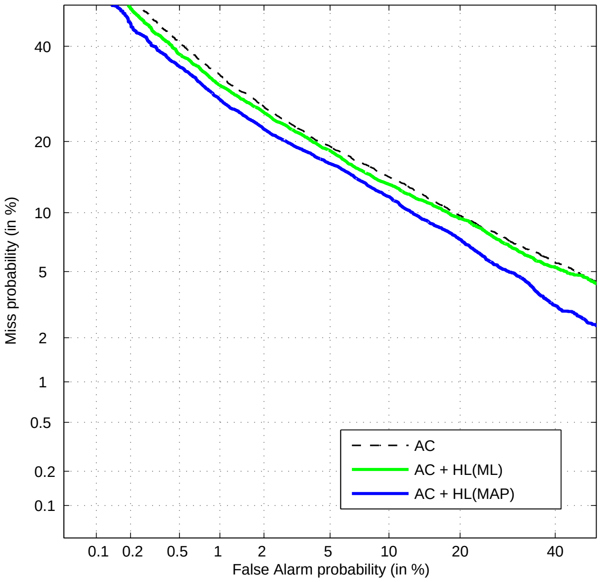
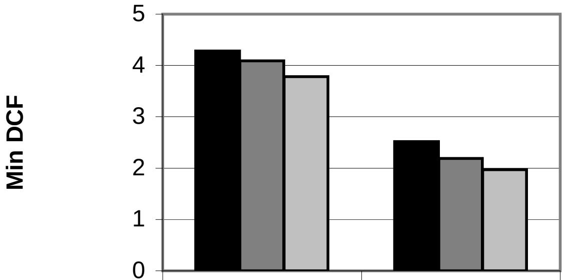


Figure 6: *DET plot for the one side training condition comparing a) a baseline acoustic system b) a fused acoustic and ML high-level system c) a fused acoustic and MAP adapted high-level system.*

6. Conclusions

Phonetic and lexical *n*-gram based speaker systems were evaluated using substantially reduced lengths of training speech. Traditional ML modelling and a previously developed adaptive modelling technique for *n*-gram based



	1side	3side
■ AC	4.28	2.52
▩ AC + HL (ML)	4.09	2.19
□ AC + HL (MAP)	3.78	1.97

Figure 7: *Minimum DCF values ($\times 10^{-2}$) for the acoustic baseline and the fused acoustic and high-level systems for the one and three side training length conditions.*

features were tested using a restructured NIST 2003 EDT protocol that included two new reduced training length conditions. Results indicated that the MAP adaptation process reduced model sparsity effects and showed a marked improvement in performance over ML models for both phonetic and lexical techniques.

The individual improvements in performance obtained for the *n*-gram based features were also found to translate into overall gains in performance when used along side acoustic classifications. Fusion experiments performed combining acoustic classifications with the high-level classifications showed that even with as little as one conversation side of training data, the enhanced high-level systems provided complementary classifications and improved recognition performance.

7. Acknowledgements

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

References

Andrews, W., M. Kohler, J. Campbell, and J. Godfrey (2001). Phonetic, idiolectal, and acoustic speaker recognition. In *A Speaker Odyssey, The Speaker Recognition Workshop*.

Andrews, W., M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero (2002). Gender-dependent phonetic refraction for speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 1, pp. 149 –152.

Baker, B., R. Vogt, M. Mason, and S. Sridharan (2004). Improved phonetic and lexical speaker recognition through MAP adaptation. In *Odyssey: The Speaker and Language Recognition Workshop*, pp. 94–99.

Doddington, G. (2001). Speaker recognition based on idiolectal differences between speakers. In *Eurospeech*, Volume 4, Denmark, pp. 2517–2520.

Hermansky, H. (1990). Perceptual linear predicive (PLP) analysis of speech. *The Journal of the Acoustical Society of America 87*(4), 1738–1752.

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

Linguistic Data Consortium (1997). SWITCHBOARD: A user's manual. http://www.ldc.upenn.edu/readme_files/switchboard.readme.html.

Massechusetts Institute of Technology Lincoln Laboratory (2004). LNKnet Pattern Classification Software. http://www.ll.mit.edu/IST/lnknet/.

Muthusamy, Y., R. Cole, and B. Oshika (1992). The OGI multi-language telephone speech corpus. In *International Conference on Spoken Language Processing*.

National Institute of Standards and Technology (2003). NIST speech group website. http://www.nist.gov/speech.

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

Pelecanos, J., R. Vogt, and S. Sridharan (2002). A study on standard and iterative MAP adaptation for speaker recognition. In *International Conference on Speech Science and Technology*, pp. 190–195.

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