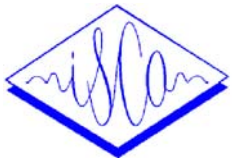

ODYSSEY04 - The Speaker and
Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Improved Phonetic and Lexical Speaker Recognition through MAP Adaptation

Brendan Baker, Robbie Vogt, Michael Mason, and Sridha Sridharan

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

Abstract

High level features such as phone and word n-grams have been shown to be effective for speaker recognition, particularly when used along side traditional acoustic speaker recognition techniques. The applicability of these high-level recognition systems is impeded by the large training data requirements needed to build robust and stable speaker models. This paper describes an extension to an existing phone n-gram based speaker recognition technique, whereby MAP adaptation is used in the speaker model training process. Results obtained for the NIST 2003 Speaker Recognition Extended Data Task indicate that a significant improvement in performance can be gained through the use of this model estimation technique. In our tests, we were able to improve performance over the baseline system, and at the same time, halved the training data requirement. Further experimentation using MAP adaptation on word n-gram models also showed improvement over baseline results, suggesting that the technique could be applied to other multinomial distribution feature sets.

1. Introduction

Traditional approaches to automatic speaker recognition only consider the acoustic properties of speech that are governed by the mechanics of the voice production system. Spontaneous and conversational speech, contains not only acoustic information, but sources of high-level information such as linguistic content, pronunciation idiosyncrasies, idiolectal word usage, prosody and speaking style. In recent years, there has seen an increased interest in the use of these high-level features of information for automatic speaker recognition. 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. This expansion has been motivated by the belief that estimation of these high-level features will be more robust to different acoustic conditions.

Doddington's [1] preliminary experiments into the use of high-level features using the SwitchBoard corpus [2] sparked this new research interest. He explored the possibility of using the lexical content of speech for speaker recognition. His approach used word n-gram statistics gathered from recorded conversational speech as features for the recognition process. This technique, although simple, gave impressive results especially when used alongside traditional acoustic features as a source of complementary information.

Motivated by the work of Doddington [1], Andrews *et al.* [3, 4] went on to successfully apply similar techniques to phone n-gram statistics in order to capture speaker pronunciation idiosyncrasies. This approach made use of parallel phone streams from multiple open-loop phone recognisers, each trained for a different language, using a similar front end to that used in Parallel Phonetic Recognition with Language Modelling (PPRLM) systems [5]. Previously, PPRLM systems have been successfully used for high-performance, extendable language identification. In the phonetic speaker recognition system, using multiple languages was found to give superior performance over any single language system. The use of multiple language phone recognisers also introduced a degree of language independence to the recognition process. This technique for speaker recognition gave encouraging results, and was found to be a useful complementary feature set when used in conjunction with short-term acoustic features.

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

Initial research into the use of high-level features has focused on characterising high-level knowledge sources and defining new features sets. Now that many of the useful features for speaker recognition have been identified, an obvious next step is to further develop the clas-

sification 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.

This paper presents the application of *Maximum A Posteriori* (MAP) adaptation [6] to the techniques described by Doddington and Andrews. This is motivated by the success of MAP adaptation in limited training conditions for Gaussian Mixture Model (GMM) based acoustic speaker recognition [7, 8, 9] and automatic speech recognition. In the MAP adaptation process, the parameters of a Universal Background Model (UBM) are used as a base model and are adjusted toward the speech of the target speaker. This has the advantages of prior information being tied into the modelling process and numeric stability of the parameters of the target speaker model. The introduction of this adaptive learning approach improved the robustness of the speaker models over the basic *Maximum Likelihood* (ML) distribution estimates.

Section 2 of this document describes the baseline phonetic speaker recognition system implemented. This is followed in Section 3, by a description of the new MAP adapted modelling technique trialed. In Section 4, results are presented for trials on the NIST 2003 Extended Data Task data set. Finally, in Section 5, the application of the MAP adaptation techinique to idiolect/lexical features is considered.

2. Phonetic N-gram Speaker Modelling

The baseline phonetic speaker recognition system is derived from the system described in [4] - 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 that are estimated through simple multinomial distributions, allowing for the capturing of recognised phonetic patterns of individual speakers. When scoring, each phone transcription is tested against *Phonetic Speaker Models* (PSM) and a *Universal Background Phonetic Models* (UBPM) using a traditional likelihood ratio test (LRT).

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

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

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

The test segment score is the log likelihood ratio (LLR) of the speaker likelihood to background likelihood and is given by

$$\Lambda = \frac{\sum_k (w(k) \cdot log[l_{TM}(k) - l_{BM}(k)])}{\sum_k w(k)}, \tag{2}$$

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

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

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 [1] and Andrews [4] 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 were ignored in the scoring process.

Andrews used this phonetic speaker recognition process on phone streams produced by multiple language open-loop phone streams. Andrews' described the transcriptions produced by 'off'-language recognisers as *refracted* phone transcriptions. These refracted streams of phones are capable of providing speaker information which is complementary to the true language's phonetic transcription.

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 net architecture implemented using the *LNKnet* pattern classification software [10], was used to fuse the individual scores.

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 order to combat data sparsity issues, and improve the robustness of the models, we propose to tie prior information about a model's parameters into each speaker's PSM. The Bayesian learning framework and MAP estimation algorithms provide us with methods to do this.

Lee [6] outlined a MAP estimation solution applicable to multinomial densities which we have adapted for this work. The MAP solution used for the n-gram frequencies can be expressed as

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

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

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

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

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

Equation 5 becomes

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

where α is an adaptation weight in the range (0, 1). In the limit of no adaptation data, this reverts back 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 should effectively cancel the need for the use of *ad hoc* pruning thresholds.

The scores from each language were fused using the same MLP neural network structure used for the baseline system.

4. Experiments

4.1. Database

The baseline and adapted systems were evaluated and compared using data from the NIST 2003 Speaker Recognition Evaluation Extended Data Task (NIST EDT) [11]. (For further information see [12]). The NIST EDT evaluation uses data from the SwitchBoard-II Phase 2 and 3 corpora [2].

Along with the audio data, NIST provided phone level transcriptions for five different languages, courtesy of R523 (Department of Defense). For consistency with published results, these transcriptions were used as the open-loop phone streams in our experiments.

During the development of both the baseline and MAP adaptive systems, a development data set consisting of the male speakers from splits 1-4 of the NIST EDT evaluation data was used. This development data set was used to tune the various parameters of the recognition systems, and also to train the neural network used for fusing results from multiple languages. Once the systems were calibrated, overall results were obtained using both male and female speaker sets from remaining evaluation splits (5-10) and all five language phone streams.

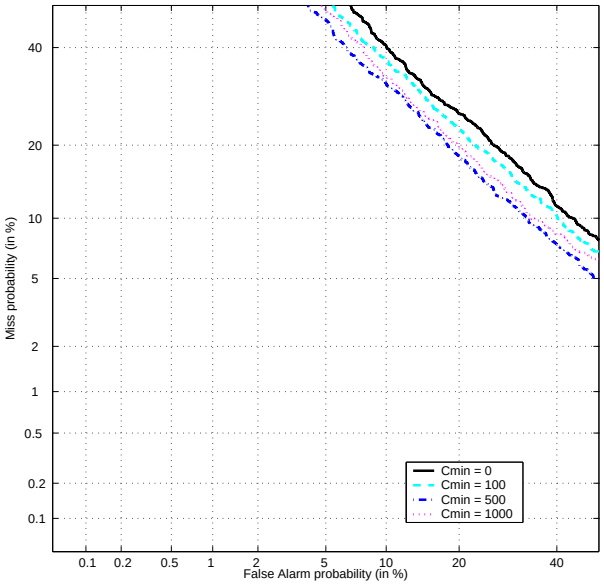

Figure 1: *DET plot for tri-phone baseline system for different values of the pruning threshold c_{min}. Results are for male tests of splits 1-4 using English phone streams only and 8 conversations for training.*

4.2. Baseline System Performance

The system was tested on the development splits using both bi-phone (n=2) and tri-phone (n=3) models. As an initial experiment, the effect of the pruning threshold c_{min} on detection performance was examined. Figure 1 shows detection-error tradeoff (DET) curves for a tri-phone system trained on 8 conversations using English phones only. The curves represent results for several values of c_{min}. Examining Figure 1, it can be seen that a pruning threshold of c_{min} = 500 for the tri-phone system offers the best performance. Using this pruning threshold improved the equal error rate (EER) to 19% from an unpruned EER of 23%. The same experiments performed on the bi-phone system, varying the pruning threshold, revealed that no benefit was gained by removing infrequent n-grams and that a pruning threshold of c_{min} = 0 was more appropriate.

Both the bi-phone and tri-phone baseline results were obtained without discounting ($d = 0$). In the baseline experiments, the use of absolute discounting ($d = 1$) for tri-phones gave a marginal performance increase, but degraded performance when used with bi-phone models.

Although in previous studies tri-phone models have been chosen [3, 4], our experiments on the NIST 2003 Extended Task Data resulted in bi-phone models out performing tri-phone models by a considerable amount. Figure 2 shows a comparison of the tuned baseline bi-phone and tri-phone systems for 8 and 4 conversations training data. The plot shows that in the tests carried out, the tri-phone models perform worse for both training conditions.

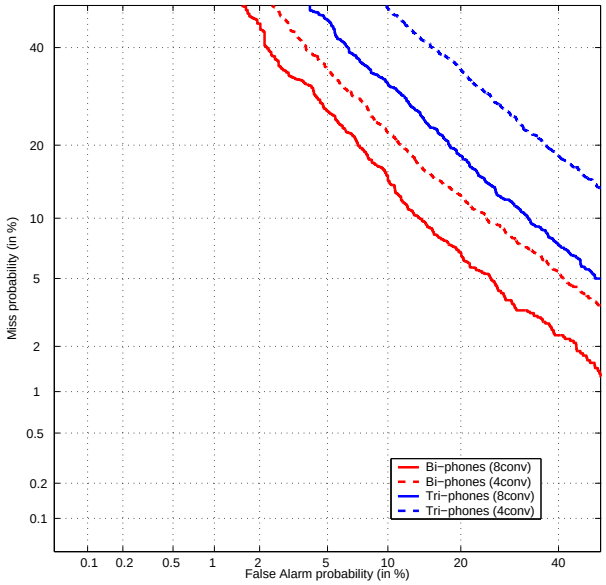

Figure 2: *DET plot comparing tri-phone models (using $c_{min} = 500$) and bi-phone models for 8 and 4 conversations training. Results are for male tests of splits 1-4 using English phone streams only.*

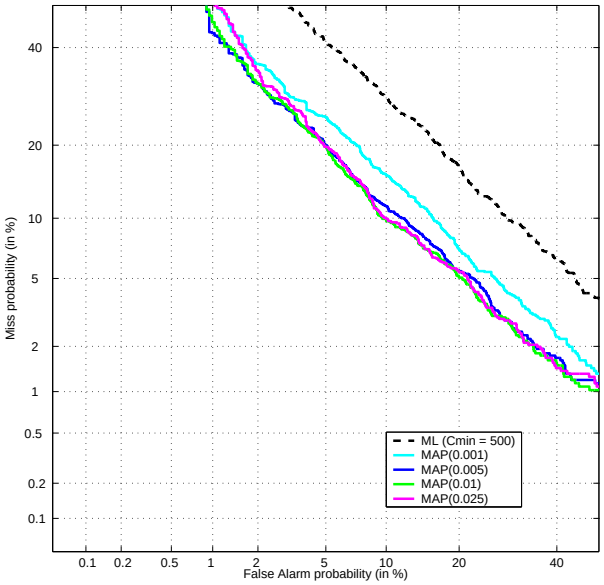

Figure 3: *Comparison of MAP adapted models for different values of α. Results are for male tests of splits 1-4 using English phone streams only with 8 conversations training.*

Also of particular note, is the quite substantial increase in EER for the tri-phone system when the training data is reduced from 8 to 4 conversations. The EER of the system using the tri-phone models increases from 18% for 8 conversations to 27% for 4 conversations, compared to an increase from 12% to 15% for the bi-phone models.

The inferior performance of the tri-phone models, in particular the poor performance with 4 conversations training suggests that many of the tri-phone tokens in the target models may be under-trained. The focus of the rest of the experiments was to try and improve the stability and performance of the tri-phone models, in particular when trained with reduced amounts of data.

4.3. MAP Adapted System Performance

Tri-phone models were generated using the MAP adaptation estimates discussed in Section 3. Initial experiments were carried out using the development data splits. Different MAP weightings α were trialed, and results compared against the baseline maximum likelihood tri-phone system. Figure 3 shows a DET plot comparing the adapted model systems using different values of α for 8 conversations training data. The dashed curve represents the baseline maximum likelihood tri-phone system performance. The plot shows that using the adaptive models increased performance considerably, and that all MAP configurations tested gave improved performance over ML models. Best performance was achieved when using a MAP adaptation weight α in the range (0.005, 0.025). With a MAP weighting of $\alpha = 0.01$, a relative improvement of 44% was achieved over the ML estimates.

A reduced training set experiment was also conducted using only 4 conversations for training. Figure 4 shows the resulting DET curves. One can see from this plot that a vast improvement over the ML models is achieved when the MAP adapted models are used. Once again, all MAP models outperformed ML models, with best performance being achieved using a MAP adaptation weight α in the range (0.005, 0.01). Using the MAP adapted models with $\alpha = 0.01$ gave a 52% relative performance improvement over the baseline ML estimate system.

The excellent results obtained using MAP adapted models under the 4 conversation training condition support the hypothesis of data sparsity issues for tri-phone models using ML training. Results also show that the MAP system is relatively insensitive to different values of the MAP adaptation weighting α, indicating that MAP adaptation is a robust method for modelling.

4.4. Overall Fused Results

The overall performance of the baseline and adapted systems using all languages was evaluated using the remaining splits (5-10) of the evaluation data. Scores for all languages were fused together using a multi-layer perceptron neural network. The MLP was constructed using the LNKnet [10] software package and was trained using the scores obtained on the development splits (1-4).

Overall results were obtained for the tuned baseline systems using bi-phone and tri-phone ML models, and the newly developed system using MAP adapted models.

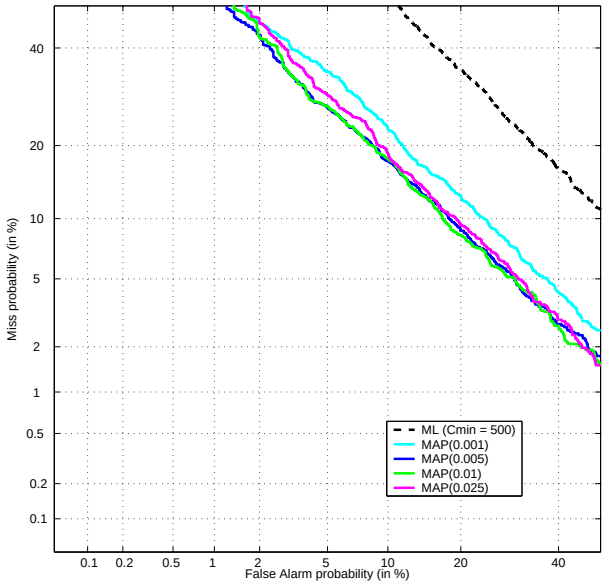

Figure 4: *Comparison of MAP adapted models for different values of α. Results are for male tests of splits 1-4 using English phone streams only with 4 conversations training.*

For the final system, a MAP weighting of $\alpha = 0.01$ was chosen. Figure 5 shows the resulting DET curves. Results are shown for both the 8 and 4 conversation training conditions.

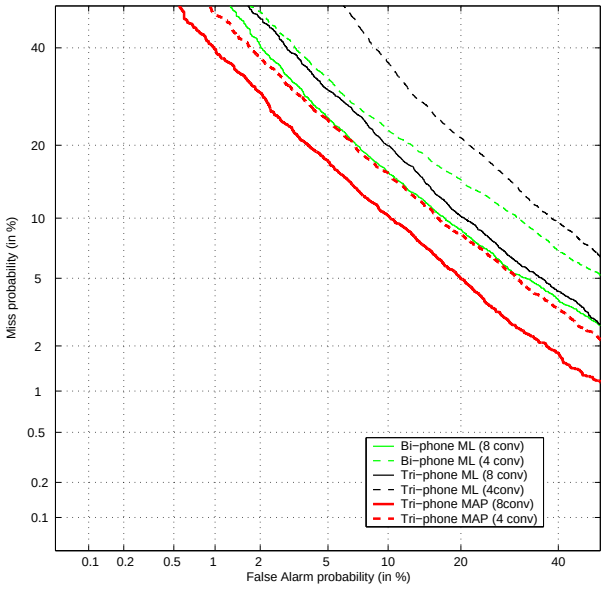

Figure 5: *Overall Fused Language Results.*

Table 1 shows a summary of these results in terms of Equal Error Rate and the minimum Detection Cost Function (DCF) criterion [11] used in the NIST 2003 evaluation. The results show that the MAP adapted models had superior performance over both ML systems in terms of detection cost and EER, for both training conditions. It is also interesting to note that the MAP adapted models when trained with 4 conversations, outperformed the ML models using 8 conversations for training. We have been able to effectively halve the training data requirements, whilst still improving on overall recognition accuracy, prior to fusing these features with a traditional acoustic speaker recognition system.

	8conv		4conv	
	EER	DCF	EER	DCF
Bi-phone ML	12.94%	0.0599	16.57%	0.0669
Tri-phone ML	14.44%	0.0660	20.73%	0.0899
Tri-phone MAP	**10.12**%	**0.0485**	**12.70**%	**0.0566**

Table 1: *Overall Phonetic Speaker Recognition Results*

5. Lexical Experiments

As a further test, the MAP adaptation modelling technique was trialed on a system using word n-gram features, to see if adaptation could improve the performance of other n-gram based models. The system is similar to that described in [1]. Results were obtained for a baseline ML bi-gram system and using MAP adapted models. A MAP weighting $\alpha = 0.01$ was used. Results were obtained using NIST 2003 SRE EDT evaluation data. Figure 6 shows DET curves for the two systems for both 8 and 4 conversations training data.

The results show that an improvement is gained by using MAP adaptation for our lexical models. The EER is reduced to 14% (from 17%) for the 8 conversation training condition, and to 19% (from 23%) for 4 conversations.

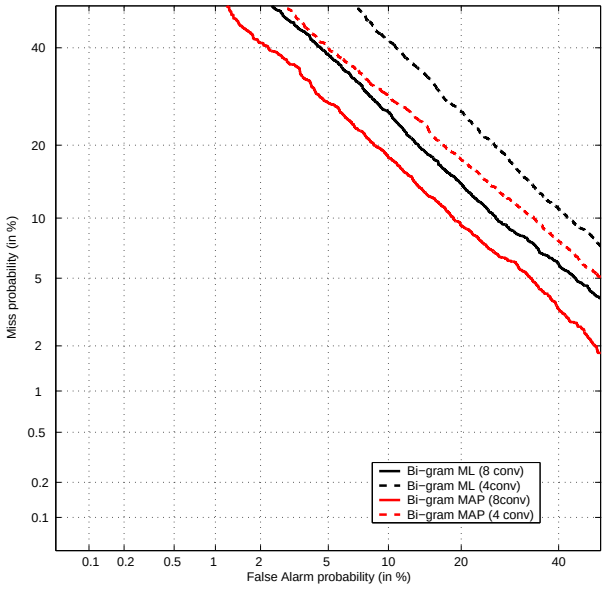

Figure 6: *Lexical results*

These results suggest that the MAP adaptation technique outlined in this paper could be applied to other multinomial distribution based features to improve the numeric stability and robustness of the models.

6. Conclusions

Motivated by model sparsity issues identified in reduced training conditions for the NIST 2003 EDT, a new modelling technique was developed for the phonetic n-gram system. Under this technique, MAP adaptation was used to incorporate prior distribution information (estimated from a UBPM) into the speaker models. Trials using this technique showed a marked improvement in performance over baseline results under all training conditions. The newly developed MAP adapted tri-phone models trained with 4 conversations was also shown to outperform the best performing baseline ML system using 8 conversations for training. In other words, a halving of the training data requirement for the models was achieved, whilst at the same time, an improvement in speaker recognition performance was gained.

Results gathered from tests of the adaptation technique on word n-gram based models also suggest that the technique can be successfully applied to other multinomial distribution based classifiers.

7. Acknowledgements

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

8. References

[1] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Eurospeech*, Denmark, 2001, vol. 4, pp. 2517–2520.

[2] "SWITCHBOARD: A user's manual," Linguistic Data Consortium, http://www.ldc.upenn.edu/readme_files/switchboard.readme.html.

[3] W. Andrews, M. Kohler, J. Campbell, and J. Godfrey, "Phonetic, idiolectal, and acoustic speaker recognition," in *Speaker Odyssey Workshop*, 2001.

[4] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 1, pp. 149 –152.

[5] M. Zissman and E. Singer, "Automatic language identification of telephone speech messages usinjg phoneme recognition and n-gram modelling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994, vol. 1, pp. 305–308.

[6] C. Lee and J. Gauvain, "Bayesian adaptive learning and map estimation of hmm," in *Auotmatic speech and speaker recognition : Advanced topics*, pp. 83–107. Kluwer Academic Publishers, Boston, Massachusetts, USA, 1996.

[7] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[8] D. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 4, pp. 4072–4075.

[9] R. Vogt, J. Pelecanos, and S. Sridharan, "Dependence of gmm adaptation on feature post-processing for speaker recognition," in *Eurospeech*, Switzerland, 2003, pp. 3013–3016.

[10] "LNKnet Pattern Classification Software," MIT Lincoln Laboratory, http://www.ll.mit.edu/IST/lnknet/.

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

[12] "NIST speech group website," http://www.nist.gov/speech/.