Discriminative Adaptation for Speaker Verification

C. Longworth and M. J. F. Gales

Engineering Department, Cambridge University
Trumpington St, Cambridge, CB2 1PZ
{cl336,mjfg}@eng.cam.ac.uk

Abstract

Speaker verification is a binary classification task to determine whether a claimed speaker uttered a phrase. Current approaches to speaker verification tasks typically involve adapting a general speaker Universal Background Model (UBM), normally a Gaussian Mixture Model (GMM), to model a particular speaker. Verification is then performed by comparing the likelihoods from the speaker model to the UBM. Maximum A-Posteriori (MAP) is commonly used to adapt the UBM to a particular speaker. However speaker verification is a classification task. Thus, robust discriminative-based adaptation schemes should yield gains over the standard MAP approach. This paper describes and evaluates two discriminative approaches to speaker verification. The first is a discriminative version of MAP based on Maximum Mutual Information (MMI-MAP). The second is to use an augmented-GMM (A-GMM) as the speaker-specific model. The additional, augmented, parameters are discriminatively, and robustly, trained using a maximum margin estimation approach. The performance of these models is evaluated on the NIST 2002 SRE dataset. Though no gains were obtained using MMI-MAP, the A-GMM system gave an Equal Error Rate (EER) of 7.31%, a 30% relative reduction in EER compared to the best performing GMM system.

Index Terms: augmented statistical models, discriminative training, sequence kernels, speaker verification.

1. Introduction

Gaussian-mixture models (GMM) have become the dominant approach for modeling acoustic features in text-independent, speaker verification systems[1]. The standard approach is to train a GMM on all the available speaker data and use this as a Universal Background Model (UBM) to represent all speakers. This UBM is then adapted to the limited amount of enrolment data for a particular speaker, Maximum *A-Posteriori* adaptation is the usual approach to allow the large number of GMM components in the UBM to be robustly adapted to a speaker. However speaker verification is inherently a classification task. Hence discriminative approaches to robustly adapting the general UBM to the specific speaker have the capability to yield gains over the standard MAP approach. Most previous discriminative approaches have concentrated on the use of Support Vector Machines (SVMs) with sequence kernels that handle the dynamic nature of the speaker verification task, example kernels include generative kernels [2], the Kullback-Leibler kernel [3] and the GMM supervector kernel [4]. All these approaches generate decision boundaries in a *score-space* rather than discriminatively adapting the speaker models. This paper describes and evaluates two different discriminative approaches for speaker adaptation.

The first discriminative adaptation approach is based on a discriminative MAP scheme which has been found to work well for both task and gender adaptation in automatic speech recognition (ASR) [5]. Robust MAP estimates are obtained using the Maximum Mutual Information (MMI) criterion, rather than Maximum Likelihood (ML). This approach can be viewed as maximising the posterior of the correct speaker compared to all other speakers. The second approach uses an augmented GMM (A-GMM) as the speaker-specific model. Here the standard MAP adapted speaker model is augmented by a local exponential approximation. The parameters of this augmentation, the augmented parameters, are estimated using maximum margin training, a discriminative approach [6]. Maximum margin estimation schemes should yield robust estimates even on limited data. This form of model is closely related to the verification work in [2]. However the approach here is from an adaptation perspective, rather than using an SVM to generate a decision boundary in a generative score-space. This allows the posterior to be computed for any observation allowing simple combination with other statistical approaches [6]. Such probabilistic interpretations are not normally possible with SVMs.

This paper is structured as follows. The next section briefly describes statistical approaches to speaker verification and the two discriminative approaches investigated in this paper. In section 3, experimental results on the 2002 NIST speaker recognition evaluation dataset are presented. Finally, conclusions are drawn.

2. Discriminative Adaptation

The standard approaches used for speaker verification are based on Bayes' decision rule. Here to decide whether speaker s uttered $\boldsymbol{O}$ the following decision rule is applied

$$\log\left(P(\omega_s|\boldsymbol{O};\boldsymbol{\lambda})\right) \underset{reject}{\overset{accept}{\gtrless}} \beta \qquad (1)$$

β is a threshold used to set false accepts and false rejects and $\boldsymbol{\lambda}$ are the model parameters for all S speakers . As generative models, GMMs, are usually used, Bayes' rule can be applied to obtain the posterior of the class given models for all speakers. However, rather than using a combined speaker model in the denominator, which assumes a closed set, a UBM is usually trained on all the speakers and used instead. This simple approximation is faster and yields performance gains. The UBM is also used as the prior distribution for the MAP estimates of the speaker-specific parameters, $\boldsymbol{\lambda}^{(s)}$ [1]. This section describes how discriminative adaptation schemes may be used to obtain the speaker-specific models.

C. Longworth would like to thank the Schiff Foundation for funding.

2.1. Discriminative MAP

Rather than using standard ML-based MAP approaches to generate the speaker-specific parameters, $\boldsymbol{\lambda}^{(s)}$ from the UBM, it is possible to use discriminative MAP approaches such as MMI-MAP [5]. Here the following criterion is optimised

$$\mathcal{F}_{\mathtt{mmi}} = \sum_{q=1}^{Q} \left(\log \left(P(\omega(q)|\boldsymbol{O}^{(q)}; \boldsymbol{\lambda}) \right) \right) + \log \left(p(\boldsymbol{\lambda}) \right) \qquad (2)$$

where $\omega(q)$ indicates the correct speaker for utterance $\boldsymbol{O}^{(q)}$ and $p(\boldsymbol{\lambda})$ is the prior distribution of the model parameters for all speakers. By maximising $\mathcal{F}_{\mathtt{mmi}}$ it is possible to obtain robust estimates, controlled by the prior, to maximise the posterior probability of the correct speaker.

In common with other discriminative estimation schemes it is not possible to get closed form estimates for the model parameters using for example Expectation Maximisation (EM). Instead a *weak-sense* auxiliary function may be used [5]. This yields update formulae similar in fashion to EM but applicable to discriminative criteria. For MMI-MAP the the new estimate of the mean of component j for speaker s, $\hat{\boldsymbol{\mu}}_j^{(s)}$ given the current estimate $\boldsymbol{\mu}_j^{(s)}$ is (considering a single utterance $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$)[1]

$$\hat{\boldsymbol{\mu}}_j^{(s)} = \frac{\sum_{t=1}^{T} \gamma_j^{(s)}(t)\boldsymbol{o}_t - \sum_{t=1}^{T} \gamma_j^{\mathtt{den}}(t)\boldsymbol{o}_t + D_j\boldsymbol{\mu}_j^{(s)} + \tau^I \tilde{\boldsymbol{\mu}}_j^{(s)}}{\sum_{t=1}^{T} \gamma_j^{(s)}(t) - \sum_{t=1}^{T} \gamma_j^{\mathtt{den}}(t) + D_j + \tau^I} \qquad (3)$$

where τ^I is a constant controlling the influence of the prior, $\tilde{\boldsymbol{\mu}}_j^{(s)}$ is the prior mean for component j of speaker s, D_j is a component-specific smoothing term that ensures that weak-sense auxiliary function is convex. $\gamma_j^{(s)}(t)$ is the posterior probability of component j generating the observation at time t given the current model parameters for speaker s and $\gamma_j^{\mathtt{den}}(t)$ is the posterior for the *denominator* model. In contrast to the standard MAP approach the the UBM cannot be used for this, the denominator must be based on the combined speaker model $\left(\sum_{i=1}^{S} P(\omega_i)p(\boldsymbol{O}; \boldsymbol{\lambda}^{(i)})\right)$.

There are a number of parameters that need to be set. For ASR the prior term τ^I is normally set in the range 0-100 [5] and a similar range was investigated in this work. The component specific smoothing term, D_j was set to be twice the denominator occupancy, $\sum_{t=1}^{T} \gamma_j^{\mathtt{den}}(t)$. The prior for the transform parameters was set as the MAP adapted mean. Thus as $\tau^I \rightarrow \infty$ the scheme will simply become equivalent to the standard MAP approach. In addition an acoustic deweighting term is commonly used [5]. In preliminary experiments it was found that an acoustic model scaling factor of around 10^{-3} was required in contrast to ASR systems which use values of 10^{-1}. This difference was felt to be due to the nature of the combined speaker denominator model used.

2.2. Augmented Gaussian mixture models

Augmented statistical models were proposed in [7] as a method of incorporating complex dependencies in a statistical model. Starting from a base distribution, $\hat{p}(\boldsymbol{O}; \boldsymbol{\lambda})$, a local exponential approximation to that base distribution is produced. Thus for an set of observations $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$, $\boldsymbol{o}_t \in \mathbb{R}^d$, augmented statistical models take the form

$$p(\boldsymbol{O}; \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \frac{1}{Z}\check{p}(\boldsymbol{O}; \boldsymbol{\lambda}) \exp\left(\boldsymbol{\alpha}^{\mathsf{T}} \nabla_{\boldsymbol{\lambda}}^{(1,\rho)} \ln \check{p}(\boldsymbol{O}; \boldsymbol{\lambda}) \right) \qquad (4)$$

where $\nabla_{\boldsymbol{\lambda}}^{(1,\rho)} \ln \check{p}(\boldsymbol{O}; \boldsymbol{\lambda})$ are the derivatives of order 1 through ρ of the log likelihood with respect to the base model parameters $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ is a set of additional parameters, referred to as the augmented parameters, controlling the influence of each derivative on the final likelihood. Z is a normalisation term, required to ensure that $p(\boldsymbol{O}; \boldsymbol{\lambda}, \boldsymbol{\alpha})$ is a valid probabilistic distribution and is defined as

$$Z = \int \check{p}(\boldsymbol{O}; \boldsymbol{\lambda}) \exp\left(\boldsymbol{\alpha}^{\mathsf{T}} \nabla_{\boldsymbol{\lambda}}^{(1,\rho)} \ln \check{p}(\boldsymbol{O}; \boldsymbol{\lambda}) \right) d\,\boldsymbol{O} \qquad (5)$$

The model parameters are normally estimated in two stages. First the base distribution parameters, $\boldsymbol{\lambda}$, are estimated then the augmented parameters $\boldsymbol{\alpha}$ are found. In general the estimation of the augmented parameters is highly complex. However for binary classification tasks the estimation of these model parameters is related to finding a linear decision boundary in a generative score-space [6].

Augmented statistical models are used in this work for the speaker-specific model, $p(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)}, \boldsymbol{\alpha}^{(s)})$. It is necessary to first define the base distribution, $\check{p}(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)})$. As in standard verification a GMM is used as the base-distribution to yield an A-GMM. The parameters for the base distribution of speaker s are the standard MAP adapted model set. Thus if the augmented model parameters are set to zero, $\boldsymbol{\alpha}^{(s)} = \boldsymbol{0}$, the system defaults to the standard MAP approach. This base distribution is then augmented, in this paper, with first order derivatives with respect to the means only. These derivatives can be calculated as

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}_j^{(s)}} \log(\check{p}(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)})) = \sum_{t=1}^{T} \gamma_j^{(s)}(t)\boldsymbol{\Sigma}_j^{(s)-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_j^{(s)}) \qquad (6)$$

$\boldsymbol{\mu}_j^{(s)}$ and $\boldsymbol{\Sigma}_j^{(s)}$ are the MAP-adapted mean and covariance matrix for component j for speaker s.

To obtain the speaker-specific augmented parameters, $\boldsymbol{\alpha}^{(s)}$, equation 1 is used. The posterior is approximated by

$$P(\omega_s|\boldsymbol{O}; \boldsymbol{\lambda}) \approx \frac{P(\omega_s)p(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)}, \boldsymbol{\alpha}^{(s)})}{P(\omega_u)p(\boldsymbol{O}; \boldsymbol{\lambda}^{(u)})}$$

$p(\boldsymbol{O}; \boldsymbol{\lambda}^{(u)})$ is the UBM distribution and $P(\omega_u)$ is the prior distribution for the UBM[2]. Substituting this into equation 1 yields a linear decision boundary in a score-space

$$\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)}) + b \;\; \overset{accept}{\underset{reject}{\gtrless}} \;\; \beta \qquad (7)$$

where the direction, $\mathbf{w}$, and bias, b, are given by

$$b = \log\left(\frac{P(\omega_s)}{P(\omega_u)Z^{(s)}} \right), \qquad \mathbf{w} = \left[\begin{array}{c} 1 \\ \boldsymbol{\alpha}^{(s)} \end{array} \right] \qquad (8)$$

$Z^{(s)}$ is the normalisation term associated with augmented speaker model. The score space is a generative score-space [6], $\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)})$, defined as

$$\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)}) = \left[\begin{array}{c} \log \check{p}(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)}) - \log p(\boldsymbol{O}; \boldsymbol{\lambda}^{(u)}) \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}^{(s)}} \log \check{p}(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)}) \end{array} \right] \qquad (9)$$

[1] If $D_j = 0$ and $\gamma_j^{\mathtt{den}}(t) = 0$ this yields the standard MAP update.

[2] Normally priors on the models are incorporated in the threshold β with $P(\omega_u) = 1$. However for A-GMM training they need to be maintained as part of the model since training optimises the ratio of the two priors.

This score-space is similar to that used in [2]. However, here the score-space is defined from a modeling perspective rather than as a dynamic kernel. One approach to robustly estimating linear decision boundaries in high-dimensional space (the dimensions of the score-space is $Md + 1$ where M is the number of components in the speaker-specific GMM) is to use maximum margin training. This may be viewed as building a Support Vector Machine (SVM) using the score-space $\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda}^{(s)})$. As training an SVM is a distance-based learning approach an appropriate metric is required. In this work the distance between two observations, $K(\boldsymbol{O}^{(i)}, \boldsymbol{O}^{(j)}; \boldsymbol{\lambda}^{(s)})$ is defined as

$$K(\boldsymbol{O}^{(i)}, \boldsymbol{O}^{(j)}; \boldsymbol{\lambda}^{(s)}) = \frac{1}{T_i T_j} \boldsymbol{\phi}(\mathbf{O}^{(i)}; \boldsymbol{\lambda}^{(s)})^{\mathsf{T}} \mathbf{G}^{-1} \boldsymbol{\phi}(\mathbf{O}^{(j)}; \boldsymbol{\lambda}^{(s)}) \quad (10)$$

where $\mathbf{G}$ is the total diagonal covariance of all the score-space examples, T_i and T_j are the lengths of utterances $\boldsymbol{O}^{(i)}$ and $\boldsymbol{O}^{(j)}$ respectively. This is the form of metric described in [6].

3. Experimental Results

The performance of the discriminative adaptation approaches was evaluated on the 2002 NIST SRE one-speaker detection task. This tasks contains telephone cellular data with speech from 139 male and 191 female speakers. The utterances are split so that only one side of the conversation is present. For training there was a single utterance from each speaker for enrolment of up to 120 seconds in length. There are 3570 test utterances, each of known gender. Each test utterance is scored against 11 different potential speakers of the correct gender, one of which is usually the true speaker. The utterances were parameterised using a frame rate of 10ms and a window size of 30ms. A 31 dimensional feature vector was extracted from each frame using a bandwidth of 0-3.8 KHz. The feature vector consisted of 15 static Mel-PLP coefficients, 15 delta coefficients and the delta energy. Static energy or acceleration features were not extracted as it has been previously reported [8] that they contain no speaker-discriminative information. Lastly, Cepstral Mean Subtraction was performed on each utterance to introduce robustness to stationary channel noise. Cepstral feature warping [9] was also carried out using a 3 second window to introduce additional noise robustness.

Gender-dependent UBMs were trained using Baum-Welch reestimation from an initial single-component, diagonal-covariance model doubling the number of mixture components after every 4 iterations. The UBMs were trained using all training utterances of the appropriate gender. This approach differs from that in [8] where the UBMs were trained on a different dataset. Baseline speaker-dependent GMMs were trained using 2 iterations of MAP adaptation using the appropriate UBM as a static prior. The MAP adaptation constant, τ^{map}, was initially kept fixed at 25 and only the model means were adapted. Preliminary experiments were carried out to investigate whether additionally adapting variances and mixture weights provided any improvement in verification, however as in [8] no significant improvement in performance was observed.

Performance was primarily evaluated using the Equal Error Rate (EER). This is the value of the False Alarm and the Miss probabilities when the operating threshold is adjusted such that they are equal. The EER score provides a threshold-independent score for which the costs of misses and false alarms are equal. DET graphs [10] are also used in this paper to compare classifiers. These are similar to Receiver Operating Characteristics (ROC) curves that plot miss against false alarm probability except that DET graphs utilise an exponentially warped scale to improve readability. In the 2002 NIST SRE, performance was evaluated by means of a detection cost function (DCF) . This is the weighted sum of the False alarm and Miss probabilities at a defined threshold. The normalised cost [11] takes the form

$$DCF = P_{Miss} + 9.9 P_{FalseAlarm} \quad (11)$$

In these experiments no attempt was made to obtain an appropriate operating threshold. To aid comparison with other work we report minDCF scores alongside EER. Here, minDCF is the minimum DCF obtained a-posteriori by adjusting the decision threshold.

Initially the MMI-MAP scheme was evaluated. As discussed in section 2.1, the MAP adapted speaker models were used as the prior. Only a single iteration of MMI-MAP was performed with various values of τ^I. In these preliminary results only the male-speakers were evaluated and a 128 component UBM was used. To be consistent with the training criterion verification was based on the composite speaker model in the denominator, rather than the UBM. This caused a slight degradation in the baseline performance, about 0.15% in EER.

τ^I	50	100	200	500	∞
EER (%)	12.92	12.81	12.83	12.51	12.33

Table 1: Performance of MMI-MAP with varying τ^I

The MMI-MAP results are shown in table 1. Overall the performance is highly disappointing, as for no value of τ^I does MMI-MAP outperform the baseline ($\tau^I = \infty$). In training the posterior probability of the correct speaker did increase, however this gain did not generalise to the test speakers. This lack of generalisation differs from results obtained on the YOHO database where gains were observed with MMI-MAP.

A-GMM-based speaker-models were then trained with the baseline MAP-adapted GMMs as the base-distributions. Maximum Margin training was implemented using SVM^{light} [12]. As impostors are required for SVM training, utterances from all competing speakers of the same gender were used during training. To prevent the classifier unduly penalising true utterances, each true utterance was weighted until the "sizes" of the true and impostor training sets were equal. The performance of the A-GMM based system was evaluated on both male and female data. Initially the value of τ^{map} used in standard MAP was set at 25 for all systems.

# Components	GMM		A-GMM	
	EER(%)	minDCF	EER(%)	minDCF
128	12.17	0.5014	8.62	0.3714
256	11.24	0.4704	7.88	0.3467
512	11.13	0.4638	7.48	0.3371
1024	11.37	0.4669	7.31	0.3281

Table 2: Performance for A-GMM and GMM acoustic models

Table 2 shows the performance of both the the baseline GMM and the A-GMMs. The performance of the GMM systems only improved marginally as the number of components was increased from 256 up to 1024. This lack of performance gain is in contrast to that observed in [8]. The A-GMM systems shows large gains over the baseline GMM system for all sizes of model. The best performing system was based on 1024 components and gave an EER of 7.31%.

In [8] the value of $\tau^{\mathtt{map}}$ was reduced as the number of components increased. This is because the amount of data associated with each component decreases as the number of components increases. Thus with a fixed $\tau^{\mathtt{map}}$ many components will not shift far from the prior. Table 3 shows the performance of the 1024 component UBM system with both GMMs and A-GMMs. Again for all values of $\tau^{\mathtt{map}}$ the A-GMM out-performed the baseline GMM. However as the value $\tau^{\mathtt{map}}$ increased the performance gain of the A-GMM over the GMM gradually increased (note the performance of the GMM initially improved and then got worse). This can be explained as the augmented parameters $\boldsymbol{\alpha}^{(s)}$ are estimated in the score-space defined by the MAP-adapted base distribution. If the base distribution is too "close" to the training example, a "biased" score-space, and scores, compared to the test data will be obtained. This bias is present for both the GMM and A-GMM, but it affects the performance more in the A-GMM case because of the large dimension of the score-space. Thus the best value of $\tau^{\mathtt{map}}$ is expected to be smaller for the GMM system than the A-GMM system, as seen in table 3. When $\tau^{\mathtt{map}} = 0$ the parameter estimates for the speaker models are based only on the maximum-likelihood estimate given the adaptation data, this is similar to the approach used in [2].

$\tau^{\mathtt{map}}$	GMM		A-GMM	
	EER(%)	minDCF	EER(%)	minDCF
0	11.94	0.4724	9.54	0.4317
10	10.43	0.4378	7.75	0.3585
25	11.37	0.4669	7.31	0.3281
50	12.33	0.5029	7.44	0.3272

Table 3: Performance for 1024-component acoustic models

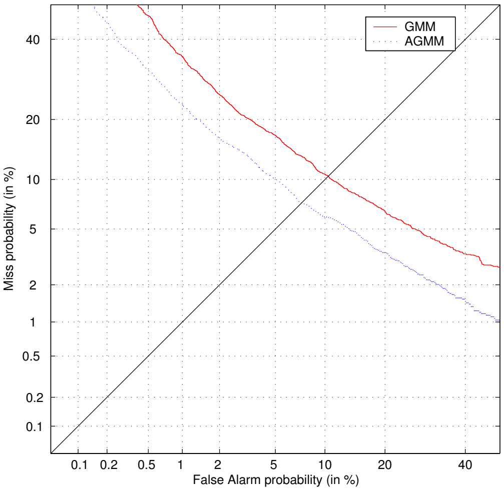

Figure 1: DET curve comparing best GMM and A-GMM systems

The best performing GMM system used 1024 components and $\tau^{\mathtt{map}} = 10$. The best A-GMM also was based on a 1024 component UBM, but with $\tau^{\mathtt{map}} = 25$. The DET curves for these two systems are shown in figure 1. Compared to the best baseline GMM system, the A-GMM achieved a 30% relative reduction in EER.

4. Conclusions

This paper has described two discriminative approaches to speaker verification. The first approach is based on discriminative-MAP (MMI-MAP) to adapt a UBM to a particular speaker. This allows the posterior of the correct speaker to be directly increased. This scheme has been found to be useful for ASR tasks, but yielded no performance gain on the 2002 NIST SRE task. Though gains in the posterior of the correct speaker were obtained in training these did not generalise well to the test data. In contrast, the maximum-margin trained Augmented GMM, A-GMM, acoustic models were found to generalise well. The A-GMM based system gave a 30% relative reduction in EER compared to the best performing GMM system. The best system performance achieved was 7.31% EER.

5. References

[1] D.A. Reynolds, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[2] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Transactions Speech and Audio Processing*, 2004.

[3] P. Moreno, P. Ho, and B. Vasconcelos, "A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications," *Advances in Neural Information Processing Systems*, 2004.

[4] W.M. Campbell, D.Sturim, D.A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006.

[5] D. Povey, M.J.F. Gales, D.Y. Kim, and P.C. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation," in *Eurospeech*, 2003.

[6] M.J.F. Gales and M.I. Layton, "SVMs, score-spaces and maximum margin statistical models," in *Beyond HMM workshop, ATR*, 2004, http://mi.eng.cam.ac.uk/˜mjfg/BeyondHMM.pdf.

[7] N.D. Smith, *Using Augmented Statistical Models and Score Spaces for Classification*, Ph.D. thesis, University of Cambridge, September 2003.

[8] C. Barras and J.L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. ICASSP*, 2003.

[9] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition - 2001: A Speaker Oddyssey*, 2001.

[10] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przyboci, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997.

[11] A. Martin, "The NIST year 2002 speaker recognition evaluation plan," 2002, Available from http://www.nist.gov/speech/tests/spk/2002/doc.

[12] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges and A. Smola, Ed. MIT Press, 1999.