Structural Joint Factor Analysis for Speaker Recognition

Marc Ferràs, Koichi Shinoda and Sadaoki Furui

Tokyo Institute of Technology, Tokyo, Japan

ferras@furui.cs.titech.ac.jp, {shinoda,furui}@cs.titech.ac.jp

Abstract

In recent years, adaptation techniques have been given a special focus in speaker recognition tasks. Addressing the separation of speaker and session variation effects, Joint Factor Analysis (JFA) has been consolidated as a powerful adaptation framework and has become ubiquitous in the last NIST Speaker Recognition Evaluations (SRE). However, its global parameter sharing strategy is not necessarily optimal when a small amount of adaptation data is available. In this paper, we address this issue by resorting to a regularization approach such as structural MAP. We introduce two variants of structural JFA (SJFA) that, depending on the amount of data, use coarser or finer parameter approximations in the adaptation process. One of these variants is shown to considerably outperform JFA. We report relative gains over 25% EER on the 2006 NIST SRE data for GMM-SVM systems using SJFA over systems using JFA.

Index Terms: joint factor analysis, structural maximum a posteriori, speaker recognition, support vector machines

10.21437/Interspeech.2011-66

1. Introduction

Speaker adaptation techniques are ubiquitous in state-of-the-art speaker recognition systems. In scenarios where many parameters need to be estimated, adaptation is often preferred over full retraining due to the smaller amount of data required for the same estimation accuracy. Two optimization criteria can be found in the literature, namely Maximum a Posteriori (MAP) and Maximum Likelihood (ML) based adaptations. Amongst the former, many variants have been proposed and some of them have been successfully applied to speaker recognition tasks, e.g. relevance MAP [1, 2], eigenchannels [3] or joint factor analysis (JFA) [4]. These methods differ in the constraints introduced in their variability models. Relevance MAP adapts any observed Gaussian independently, whereas eigenchannels assumes the session variability to be constrained to a low dimensional space. JFA uses two low-rank spaces for modeling speaker and session variability respectively. These constraints allow the speaker and session effects to be modeled separately and also reduce the number of parameters to be estimated by means of a parametric formulation.

Despite a strong parameter tying, sparsely populated Gaussian mixtures can be poorly adapted in methods such as eigenchannels or JFA. These mixtures are given a small contribution in the computation of the speaker or session latent variables and so do the corresponding Gaussian mean vectors. To overcome this issue, we proposed in [5] to use structural MAP (SMAP) adaptation [6] in a GMM-supervector based Support Vector Machine (SVM) system. SMAP structures the acoustic space in a hierarchical manner so that estimates with different precision can be used to compute the adapted parameters. It was shown in [5] that the system using SMAP estimates and nuisance attribute projection (NAP) inter-session variability compensation significantly outperformed its relevance MAP counterpart. However, this approach decouples adaptation and session compensation, the latter being applied in a post-processing stage. Furthermore, NAP uses inter-session variability information only for compensation, ignoring inter-speaker variability.

This work was supported by a Kakenhi grant 21·09805.

In this paper we propose a structural MAP framework for adaptation of joint factor analysis models of speaker and session variability, that we call "structural JFA" (SJFA). As in SMAP and hierarchical Eigenvoice adaptation [7], the main goal is to be able to keep a hierarchially structured acoustic space, now with separate speaker and session components, from which parameter estimates with different precisions can still be obtained for adaptation. Coarser speaker estimates are used for poorly observed Gaussians whereas standard JFA speaker estimates are used for Gaussians with more observations assigned. We evaluate the SJFA technique using GMM-supervector SVM-based systems on the NIST 2006 Speaker Recognitinon Evaluation (SRE) data.

This paper is structured as follows: Section 2 briefly describes joint factor analysis. 3 presents structural JFA and its formulation as well as three adaptation variants. Section 4 details the GMM-SVM speaker verification system used for experimentation. Section 5 describes the experimental protocol. The experiments and the results are shown and discussed in Section 6. Conclusions are given in Section 7.

2. Joint Factor Analysis

The aim of joint factor analysis (JFA) [4, 8] is to separate the speaker and session effects during adaptation of a Universal Background Model (UBM) to some speaker data. For this purpose, JFA uses a parametric model of the form

$$\mathbf{m}' = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} + \mathbf{U}\mathbf{x} \tag{1}$$

where $\mathbf{m}'$ and $\mathbf{m}$ are the adapted and UBM mean supervectors, $\mathbf{V}\mathbf{y}$ and $\mathbf{D}\mathbf{z}$ are the low-rank and full-rank diagonal speaker components respectively and $\mathbf{U}\mathbf{x}$ is the low-rank session component. In the training phase, the matrices $\mathbf{V}$, $\mathbf{D}$ and $\mathbf{U}$ are estimated under the MAP criterion. The latent variables are assumed to be independent and distributed following a normal pdf $y \sim \mathcal{N}(0, 1)$. In the adaptation phase, the speaker factors $\mathbf{y}$,$\mathbf{z}$ are estimated so that the speaker adapted model can be computed using the speaker terms of Eq. 1.

Given a multi-speaker and multi-session database, the sufficient statistics for speaker s and session h can be computed as

$$N_s^g = \sum_{t \in s} \gamma_g(t),\ N_{h,s}^g = \sum_{t \in (h,s)} \gamma_g(t) \tag{2}$$

$$\mathbf{X}_s^g = \sum_{t \in s} \gamma_g(t)(\mathbf{x}_t - \mu^g),\ \mathbf{X}_{h,s}^g = \sum_{t \in (h,s)} \gamma_g(t)(\mathbf{x}_t - \mu^g) \tag{3}$$

where N_s^g, $N_{h,s}^g$ are the average number of frames assigned to Gaussian g of session h of speaker s and $\mathbf{X}_s^g$, $\mathbf{X}_{h,s}^g$ are its corresponding first order statistics.

In our JFA implementation, the factor loading matrices $\mathbf{V}$, $\mathbf{D}$ and $\mathbf{U}$ are successively computed, performing several iterations of conditional maximum likelihood estimation for each matrix. At each training iteration, the centered sufficient statistics, having the undesired/non-modeled effects removed from the sufficient statistics using the model of Eq. 1, are computed as

Copyright © 2011 ISCA

$$\overline{\mathbf{X}}_y^g = \mathbf{X}_s^g - N_s^g(\mathbf{m}^g + \mathbf{D}^g\mathbf{z}) - \sum_{h\in s} N_{h,s}^g \mathbf{U}^g\mathbf{x} \quad (4)$$

$$\overline{\mathbf{X}}_z^g = \mathbf{X}_s^g - N_s^g(\mathbf{m}^g + \mathbf{V}^g\mathbf{y}) - \sum_{h\in s} N_{h,s}^g \mathbf{U}^g\mathbf{x} \quad (5)$$

$$\overline{\mathbf{X}}_x^g = \mathbf{X}_{h,s}^g - N_{h,s}^g(\mathbf{m}^g + \mathbf{V}^g\mathbf{y} + \mathbf{D}^g\mathbf{z}) \quad (6)$$

with $\overline{\mathbf{X}}_y^g$, $\overline{\mathbf{X}}_z^g$ and $\overline{\mathbf{X}}_x^g$ being used in the estimation of speaker factors $\mathbf{y}$,$\mathbf{z}$ and session factors $\mathbf{x}$ respectively. Using these statistics, the corresponding latent variables are estimated for every speaker/session in the training database as follows. For the $\mathbf{y}$ speaker factors,

$$\mathbf{y} = \mathbf{L}^{-1}\mathbf{b} \quad (7)$$

$$\mathbf{L} = \mathbf{I} + \sum_{g=1}^{G} N_s^g \mathbf{V}^{g,T}\mathbf{\Sigma}^{g,-1}\mathbf{V}^g \quad (8)$$

$$\mathbf{b} = \sum_{g=1}^{G} \mathbf{V}^{g,T}\mathbf{\Sigma}^{g,-1}\overline{\mathbf{X}}_s^g \quad (9)$$

with G being the number of Gaussian mixtures in the model. The factor loading matrix $\mathbf{V}$ can then be computed row-by-row by means of matrix regression. For row i,

$$\mathbf{V}^g(i) = \left(\sum_s \overline{\mathbf{X}}_s^g(i)\mathbf{y}^T\right)\left(\sum_s N_s^g(\mathbf{L}^{-1} + \mathbf{y}\mathbf{y}^T)\right)^{-1} \quad (10)$$

Equations analogous to Eq. 7 and 10 can be obtained for $\mathbf{D}$, $\mathbf{z}$ and $\mathbf{U}$, $\mathbf{x}$ using their corresponding factor loading matrices and centered sufficient statistics.

Eqs. 8 and 9 reveal the parameter tying strategy of JFA. Both numerator and denominator involve a weighted average over all mixtures in the model, i.e. weighted by the observation count N_s^g[1]. Poorly observed Gaussians are given a small contribution[2] into Eq. 7 and, conversely, Gaussian with many observations dominate the latent variable estimates.

In the adaptation phase, the factor loading matrices are assumed to be fixed and Eq. 7 is used to jointly estimate the speaker factors $\mathbf{y}$ and session factors $\mathbf{x}$, followed by the estimation of the residual speaker factors $\mathbf{z}$. In this study, we sequentially estimate factors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. This strategy resulted in a significant performance improvement over joint estimation of $\mathbf{y}$ and $\mathbf{x}$.

3. Structural Joint Factor Analysis

Structural JFA adaptation aims at regularizing the MAP estimates obtained with JFA by using coarser estimates for poorly observed mixtures. These coarse estimates are obtained from a cluster of Gaussians, inevitably resulting in suboptimal MAP estimates during adaptation. Interestingly, these suboptimal estimates are expected to generalize better than the optimal ones.

Prior to any estimation step, SJFA first structures the acoustic space in a hierarchical manner, which is achieved by using a tree structure to cluster the UBM Gaussian mixtures. Each node/cluster keeps a Gaussian component summarizing the probability density function of all the mixtures that depend on it, thus assigning the UBM mixtures to the leaf nodes. If the tree structure is appropriate, a node's parent node should gather Gaussian mixtures that are close to the former node. We use a data-driven approach to hierarchical clustering based on the Kullback Leibler (KL) divergence as the distance measure with a predefined structure, i.e. number of layers and nodes/layer. Please refer to [6, 5] for details about the clustering process. We assume that this tree structure is available in the remaining of the paper.

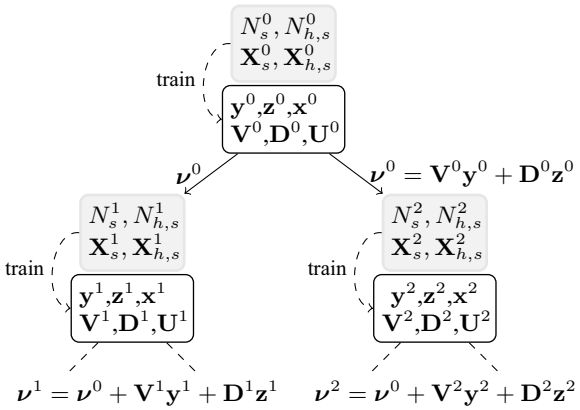

Figure 1: Diagram of speaker mean vector propagation through a tree structure in SJFA training. At each node, all the latent variables and factor loading matrices are estimated using the node's sufficient statistics. Only the speaker adapted component is propagated to the child nodes.

The next step is to populate the sufficient statistics for each node in the tree. For this purpose, starting from the speaker and session wise sufficient statistics stored in the leaf nodes, i.e. UBM Gaussians, each node in the tree gathers the sufficient statistics of all its child nodes. In a bottom-up manner, all statistics are collected recursively until the root node ends up gathering all observations of all Gaussian mixtures.

The underlying concept behind SJFA is to successively obtain JFA estimates for each node in the tree from root node to leaf nodes. A hierarchical variability model is necessary to relate the JFA estimates at each node in the tree. We propose a recursive variability model derived from Eq. 1 with parent node mean vectors replacing the UBM mean vectors, hence using hierarchical priors in the same way as SMAP. For each non-leaf node n with parent node $p(n)$ in the tree,

$$\boldsymbol{\nu}^n = \boldsymbol{\nu}^{p(n)} + \mathbf{V}^n\mathbf{y} + \mathbf{D}^n\mathbf{z} + \mathbf{U}^n\mathbf{x} \quad (11)$$

where $\boldsymbol{\nu}^n$ and $\boldsymbol{\nu}^{p(n)}$ are the cumulative mean vectors down to node n and $p(n)$ respectively[3]. The further down the tree, the more precise the estimates are, as nodes gather less Gaussians per cluster. In this sense, SJFA can be seen as a modeling speaker and session effects in a multiresolution manner. The rest of Eq. 11 is the same as JFA except that the factor loading matrices involve one Gaussian mixture only, hence all the tree loading factor matrices have as many rows as the number of cepstral coefficients. The process is depicted in Fig. 1.

SJFA training is performed node-by-node in a top-down manner. In each node, factor loading matrices are estimated as in JFA: several iterations of latent variable and loading matrix estimation for each $\mathbf{V}$, $\mathbf{D}$ and $\mathbf{U}$. Once the node's matrices are estimated the prior $\boldsymbol{\nu}^n$ for each speaker is computed and propagated down to all of its child nodes. The hierarchical training is ready when all the non-leaf nodes have their factor loading matrices, speaker and session factors as well as speaker priors $\boldsymbol{\nu}^n$ estimated on the whole training database.

The leaf node factor loading matrices are obtained by training a standard JFA model on the UBM Gaussian mixtures, with

[1] Note that $\overline{\mathbf{X}}_y^g = N_s^g(E^g[\mathbf{x}] - \mathbf{m}^g)$, so it is proportional to N_s^g.

[2] Assuming one Gaussian and one dimension, Eq. 7 simplifies to $y = \frac{v\sigma^{-1}\overline{X}_s}{1+v^2\sigma^{-1}N_s} = \frac{vN_s(E[x]-\mu)}{\sigma+v^2N_s}$. Then, $\lim_{N_s\to 0}\frac{vN_s(E[x]-\mu)}{\sigma+v^2N_s} = 0$

[3] The recursion starts with $\boldsymbol{\nu}^{p(n)} = \mathbf{0}$ when n is the root node.

several modifications that take advantage of the structural JFA available estimates. In this regard, we propose two variants:

1. **Speaker Priors (SP)**: A standard JFA model is trained replacing the UBM mean vectors $\mathbf{m}^n$ of Eq. 1 by the corresponding speaker mean vector of its parent node as

$$\mathbf{m}^{n\prime} = \boldsymbol{\nu}^{n-1} + \mathbf{V}^n\mathbf{y} + \mathbf{D}^n\mathbf{z} + \mathbf{U}^n\mathbf{x} \qquad (12)$$

   where $\boldsymbol{\nu}^{n-1}$ is the parent node's speaker estimates for Gaussian n. The speaker- and session-dependent terms of Eq. 12 have a different origin for each speaker. Coarse speaker mean vector estimates are embedded in $\boldsymbol{\nu}^{n-1}$ itself, leaving less variance to be captured by $\mathbf{Vy}$, $\mathbf{Dz}$ and $\mathbf{Ux}$. The number of speaker/session factors in the tree JFA analyses can be different from those at the leaf nodes.

   In the adaptation phase, the same training steps must be applied. The speaker and session factors are estimated and the speaker mean vectors are computed for each node in the tree in a top-down manner. The standard JFA model is then adapted using the tree-based speaker mean vectors for each Gaussian and the standard JFA factor estimates.

2. **Back-off (BO)**: The standard JFA model of Eq. 1 is used. In the adaptation phase, for each Gaussian mixture, the closest ancestor node in the hierarchy with at least N_{th} observations is chosen for adaptation. Its observation counts, factor loading matrices, centered statistics and covariance matrices are used in Eqs. 8, 9 and 10. For the estimation of speaker factors $\mathbf{y}$,

$$\mathbf{L} = \mathbf{I} + \sum_{g=1}^{G} N_s^{b_\eta(g)} \mathbf{V}^{b_\eta(g),T} \boldsymbol{\Sigma}^{b_\eta(g),-1} \mathbf{V}_\eta^b(g) \qquad (13)$$

$$\mathbf{b} = \sum_{g=1}^{G} \mathbf{V}^{b_\eta(g),T} \boldsymbol{\Sigma}^{b_\eta(g),-1} \overline{\mathbf{X}}_s^{b_\eta(g)} \qquad (14)$$

   where $b_\eta(g)$ is the closest node, starting from g itself, with a path to Gaussian g with at least η observations. This variant makes sure that all Gaussians are given a minimum contribution to the final latent variables. Note that the number of speaker/session factors in the tree JFA must be the same as that used in the standard JFA.

Regarding the computational cost, SJFA requires the estimation of loading matrices and latent variables from the top to the bottom layers of the tree. Although the process cannot be easily parallelized, the estimation can be run quite fast given that only one Gaussian is assigned to each non-leaf node. The number of latent variables is a key factor affecting the computational cost as, the more the latent variables, the larger the loading matrices that must be factored using the Choleski decomposition.

4. GMM-SVM System

The system is implemented using the Hidden Markov Model toolkit (HTK) and LIBSVM as the SVM classifier. The front-end extracts 15 Perceptual Linear Prediction (PLP) features with normalized energy, plus their Δ and $\Delta\Delta$ coefficients, every 10ms using a window of 30ms. Feature warping [9] using a 3s window is also applied. We use the ETSI 050 22 advanced front-end for voice activity detection.

For speaker adaptation, we use SJFA mean vector estimates in a GMM-SVM speaker verification system. The system uses a UBM with 512 Gaussians trained using around 400h (1100 speakers, 12000 sessions) of effective speech data taken from the NIST SRE'04 and SRE'05 and the Switchboard II phase 1 databases. The speaker-adapted mean vector parameters are stacked into a supervector that is classified with Support Vector Machines (SVM) using the NIST SRE'04 training data as impostors. The GMM supervector linear kernel [10], an upper bound of the Kullback-Leibler divergence, was used to compute the similarity between two speakers. For two speech segments s_a and s_b with the corresponding adapted models, it can be written by

$$k(s^a, s^b) = \sum_{m=1}^{M} \left(\sqrt{\lambda_m}\boldsymbol{\Sigma}_m^{-\frac{1}{2}}\boldsymbol{\mu}_m^a\right)^T \left(\sqrt{\lambda_m}\boldsymbol{\Sigma}_m^{-\frac{1}{2}}\boldsymbol{\mu}_m^b\right) \qquad (15)$$

which can be computed as a linear kernel if we let GMM supervectors be normalized as

$$\mathrm{m} = (\sqrt{\lambda_1}\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\mu}_1^T, \ldots, \sqrt{\lambda_M}\boldsymbol{\Sigma}_M^{-\frac{1}{2}}\boldsymbol{\mu}_M^T)^T. \qquad (16)$$

JFA and SJFA (non-leaf or leaf node) training use the conditional maximum-likelihood estimation to sequentially estimate the factor loading matrices. A maximum of 10 iterations per factor loading matrix is used, stopping training whenever the relative likelihood increase falls below 0.001%. The same initialization is used for all training runs. For adaptation, session and speaker factors are estimated in this order, which outperformed joint estimation in informal experiments.

5. Experimental Protocol

We followed the 2006 NIST SRE protocol[4], involving telephone speech data, a large number of speakers and strong acoustic channel mismatch. The speaker verification system is asked to decide whether speech from a given target speaker is present in an unknown speech segment. We scored the 22316 English trials of the core condition, consisting in speech segments of 5 minutes with an average of 2 minutes of effective speech per conversation side.

For evaluation, we use the Detection Cost Function (DCF) defined as a cost function weighting the false alarm and miss error probabilities $\mathrm{DCF}_{\mathrm{Norm}} = P_{\mathrm{Miss}} + 9.9 \times P_{\mathrm{FalseAlarm}}$ according to the defined decision costs. We report the Minimal DCF (MDC) based on the optimal a posteriori threshold. Since this operating point favors false alarms, we also provide the Equal Error Rate (EER) as an alternative performance measure.

6. Experiments and Results

The experiments focus on the comparison of JFA-based versus SJFA-based GMM-SVM systems for the same number of Gaussians and joint factor analysis training data. We report results for SJFA using two differentiated tree structure topologies with several hundreds of nodes each[5], one rather long with several layers (8/4/4/4 nodes/layer) and another rather wide with many nodes per layer but only two layers (15/15 nodes/layer). Systems using SJFA with speaker priors and the back-off strategy are named with the acronyms SP and BO respectively. The latter also include the minimum observation count used.

The first part of Table 1 shows results for the JFA system, with EER around 4%, which is consistent with the results obtained by other researchers [11] for the same type of system on the NIST'06 evaluation data. Note that this error rate is significantly larger than those obtained by a GMM-UBM using JFA adaptation. This is typically attributed to the correlation among Gaussians when using JFA adaptation, i.e. all Gaussian

[4]The NIST 2006 SRE evaluation plan, http://www.nist.gov/speech/tests/spk/

[5]Our best results on the SMAP-based GMM-SVM system [5] were obtained with hundreds of nodes.

System	# tree nodes	MDC	EER (%)
JFA	-	0.0199	4.24
SJFA/SP 8/4/4/4	305+512	0.0232	4.45
SJFA/BO 8/4/4/4 η=2	305+512	0.0168	3.17
SJFA/BO 8/4/4/4 η=5	305+512	**0.0159**	3.08
SJFA/BO 8/4/4/4 η=10	305+512	0.0160	3.08
SJFA/SP 15/15	210+512	0.0231	4.73
SJFA/BO 15/15 η=2	210+512	0.0168	3.21
SJFA/BO 15/15 η=5	210+512	**0.0159**	**3.03**
SJFA/BO 15/15 η=10	210+512	0.0160	3.08

Table 1: MDC and EER for GMM-SVM systems using JFA and SJFA adaptation on the 2006 NIST SRE evaluation data. For SJFA systems, the number of child nodes at each tree layer as well as the number of non-leaf and leaf nodes in the tree structure are specified. The lowest DCF and EER are shown in boldface.

mean vectors are obtained using the same speaker factors. The Kullback-Leibler kernel does not account for this correlation and can therefore be suboptimal.

Systems using hierarchical speaker priors (SP) perform poorly compared to systems using standard JFA. In the SP variant, part of the speaker variability is transferred to the speaker priors, while the terms $\mathbf{Vy}$ and $\mathbf{Dz}$ capture the remaining speaker variability. This is particularly clear during training in terms of likelihood: likelihood before training the global $\mathbf{V}$ and $\mathbf{D}$ matrices is larger in SJFA/SP compared to standard JFA, i.e. before training the global speaker subspace, the hierarchichal speaker priors have already improved the initial likelihood. However, likelihood after training $\mathbf{V}$ and $\mathbf{D}$ is exactly the same for SJFA/SP and standard JFA, indicating that the global analysis in the former captures the speaker variability not captured by the priors. However, this equivalence is not likely to hold in the adaptation phase due to the lack of data, as the speaker priors can be just roughly estimated and its residual cannot be compensated by $\mathbf{V}$ and $\mathbf{D}$, since these were trained to cover a part of the speaker space only.

All systems using the back-off variant of SJFA (BO) considerably outperform standard JFA. Despite the difference in the their topology, both 8/4/4/4 and 15/15 tree structures obtain very similar gains, reaching 20% MDC and 28% EER for a threshold of $\eta = 5$ frames. This is somewhat surprising if we account for the large number of observations, far beyond η, typically observed in the non-leaf nodes of structure 15/15. Regarding the threshold, doubling or halving it results in a slight performance loss that still largely outperforms the systems using standard JFA. As shown in Fig. 2, the gains are neatly obtained at all operating points in the DET curve.

It is worthwhile noting that these gains are obtained by partly leaving the optimality criterion aside. Using nodes other than leaf nodes in the estimation process implies that maximum likelihood estimates are no longer obtained: the likelihood on the training or adaptation data decreases. However, if we assume that the principle of regularization holds, it is expected that the SJFA estimates generalize better to *other* data of the same speaker.

7. Conclusion

We have shown that the concept behind structural MAP adaptation can be effectively used to separate speaker and channel effects in a speaker recognition task. We have introduced two variants for the estimation of a hierarchical Joint Factor Analysis variability model for adaptation of a GMM-UBM. Used in a GMM supervector system with a SVM classifier, one of the variants obtained relative gains up to 20% MDC and 28% EER versus a system using standard JFA. These results also show that the global tying strategy of JFA is not necessarily optimal under recognition performance criteria.

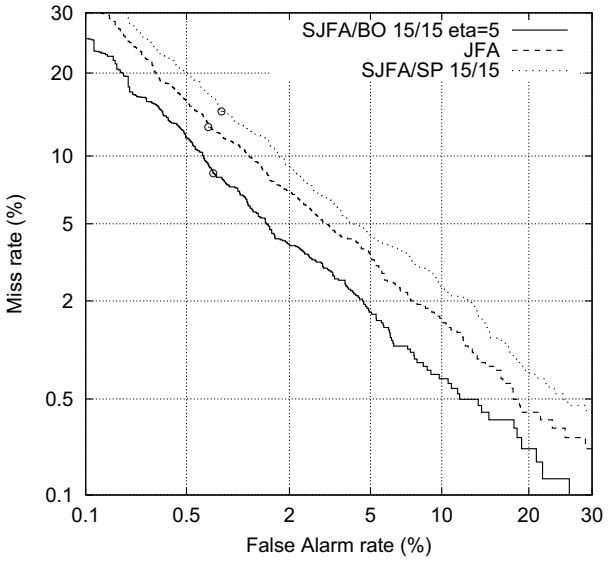

Figure 2: DET curve for GMM-SVM systems using SJFA/BO 15/15, standard JFA and SJFA/SP 15/15 adaptations. MDC operating points are shown by circles.

8. References

[1] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, April 1994.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[3] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 3117–3120.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2007.

[5] M. Ferras, K. Shinoda, and S. Furui, "Structural MAP Adaptation in GMM-Supervector based Speaker Recognition," in *Proc. IEEE ICASSP*, 2011.

[6] K. Shinoda and C.-H. Lee, "A Structural Bayes Approach to Speaker Adaptation," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 276–287, March 2001.

[7] Y. Onishi and K.-I. Iso, "Speaker adaptation by hierarchical eigenvoice," in *Proc. IEEE ICASSP*, Hong Kong, Hong Kong, 2003.

[8] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2008.

[9] J. Pelecanos and S. Sridharan, "Feature Warping for Speaker Verification," in *Proceedings of the IEEE Speaker Odyssey Workshop*, 2001.

[10] W. M. Campbell, D. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[11] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Proc. IEEE ICASSP*, Taipei, Taiwan, 2009.