INTERSPEECH 2009 BRIGHTON

Investigation into variants of Joint Factor Analysis for speaker recognition

Lukáš Burget, Pavel Matějka, Valiantsina Hubeika and Jan "Honza" Černocký

Speech@FIT, Brno University of Technology, Czech Republic

{burget|matejkap|ihubeika|cernocky}@fit.vutbr.cz

Abstract

In this paper, we have investigated into JFA used for speaker recognition. First, we performed systematic comparison of full JFA with its simplified variants and confirmed superior performance of the full JFA with both eigenchannels and eigenvoices. We investigated into sensitivity of JFA on the number of eigenvoices both for the full one and simplified variants. We studied the importance of normalization and found that gender-dependent zt-norm was crucial. The results are reported on NIST 2006 and 2008 SRE evaluation data.

Index Terms: speaker recognition, joint factor analysis.

1. Introduction

Nowadays speaker recognition systems are usually based on Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) and employ a number of techniques that improve GMM modeling capability and help fight against the main problem in speaker verification - the inter-session variability. This is caused by differences in channels, acoustic conditions and other factors varying across the speech segments being compared [2]. In several past years, systems based on Joint Factor Analysis (JFA) [4] obtained wide attention due to their ability to explicitly model the inter-session variability. However, different research labs adopted different variants JFA and it was unclear how do these variants compare in terms of recognition performance. The aim of this paper is to provide the comparison of such JFA variants and give some insight into the process of building state-of-the-art JFA system.

JFA model is a two-level generative model assuming that speech segments are generated from a GMM whose mean super-vector $\mathbf{M}$ – vector of concatenated GMM means – is first itself generated from the following distributions:

$$\mathbf{M} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} + \mathbf{U}\mathbf{x}, \qquad (1)$$

where $\mathbf{m}$ is speaker-independent mean super-vector, $\mathbf{U}$ is a subspace with high intersession variability (eigenchannels[1]), $\mathbf{V}$ is a subspace with high speaker variability (eigenvoices) and $\mathbf{D}$ is a diagonal matrix describing remaining speaker variability not covered by $\mathbf{V}$. Speaker factors $\mathbf{y}$, $\mathbf{z}$ and channel factors $\mathbf{x}$ are assumed to be normally distributed random variables. For segments of the same speaker, speaker factors are assumed to be the same, while channel factors are allowed to differ. For details, we recommend Kenny's paper [4] that served us as inspiration for building the baseline JFA systems presented in this paper.

The results in this paper are presented on NIST SRE 2006 evaluation data, especially the 1conv4w-1conv4w all-trials condition (det1 – tel-tel). The sets for other conditions (tel-mic, mic-tel, mic-mic) were defined by MIT-LL and are described in [7].

2. Baseline systems

As a baseline for the analysis presented in this paper, we have chosen two JFA systems developed for NIST SRE 2008 evaluations. The two systems differs mainly in the feature extraction.

The first system is based on features that are short time gaussianized MFCC 12 + C0 augmented with their delta, double delta and triple delta coefficients. The dimensionality of the resulting features is reduced from 52 to 39 using HLDA. HLDA classes correspond to UBM Gaussians. These features were previously used in our NIST SRE 2006 submission [2]. The system based on these features will be denoted **MFCC13⇒39**.

Inspired by the outstanding performance of the system described in [4], features used for our second baseline system are short time gaussianized MFCC 19 + energy augmented with their delta and double delta coefficients, making 60 dimensional feature vector without any dimensionality reduction . The system making use of these features will be denoted **MFCC20⇒60**.

In both cases, the features are derived with classical analysis window of 20 ms with shift of 10 ms and short-time gaussianization using window of 300 frames (3 sec). Speech/silence segmentation is performed by our Hungarian phone recognizer [1, 2], where all phoneme classes are linked to 'speech' class. Several heuristics based on short-term energy are used for two-channel telephone data to eliminate cross-talks [2].

The training of the JFA systems closely follows the description of "Large Factor Analysis model" in Patrick Kenny's paper [4]. First, UBM model with 2048 Gaussian components is trained using Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, and NIST SRE 2004 and 2005 telephone data, which is in turn used to collect zero and first order statistic for training the JFA systems. The mean super-vector $\mathbf{m}$ from (1) was set to the UBM mean and on contrary to [4] was never re-trained. The variances of Gaussian components are also taken from UBM and not re-trained in the training of JFA.

First, for each JFA system, 300 eigenvoices (matrix $\mathbf{V}$) are trained using EM algorithm [4] on the same data as UBM. For the estimated eigenvoices, MAP estimates of speaker factors are obtained and fixed for the following training of eigenchannels. A set of 100 eigenchannels is trained on NIST SRE 2004

This work was partly supported by European projects MOBIO (FP7-214324) and AMIDA (FP6-033812), by Grant Agency of Czech Republic project No. 102/08/0707, and by Czech Ministry of Education project No. MSM0021630528. We would like to thank MIT-LL team for creating the sets with the microphone conditions and Patrick Kenny for fruitful discussions helpful advices

[1] We refer to "eigenvoices" and "eigenchannels" following the terminology defined in [4] although these sub-spaces are estimated using EM-algorithm, not PCA.

and 2005 telephone data. Another set of 100 eigenchannels is trained on SRE 2005 auxiliary microphone data to allow the system to deal with the microphone speech segments. Both sets are stacked to form the final matrix $\mathbf{U}$. On contrary to Kenny's paper [4], the diagonal matrix describing the remaining speaker super-vector variability (matrix $\mathbf{D}$ in (1)) is estimated on top of eigenvoices and eigenchannels. A small disjoint set of NIST SRE 2004 speakers (recordings of only 44 females and 13 males) is used for training of $\mathbf{D}$ using fixed MAP point estimates of speaker and channel factors. To obtain speaker models, MAP point estimates of all the factors are estimated on enrollment segments using Gauss-Seidel-like iterative method [6]. For details about the training data and its splits for training the different sets of hyperparameters see [7]. In all the experiments described in this paper, the standard 10-best Expected Log Likelihood Ratio frame-by-frame scoring was used. It was based on the MAP point estimates of the channel factors[2].

Unless stated otherwise, all results were obtained with scores normalized using zt-norm. We have used 221 females and 149 males z-norm segments, 200 females and 159 males t-norm models, together 729 segments taken each from one speaker of NIST SRE 2004 and 2005 data.

In the case of systems developed for NIST SRE 2008 evaluations, single gender-independent (GI) system **MFCC13⇒39** was trained and evaluated using the data of both genders, while two gender-dependent (GD) systems **MFCC20⇒60** were trained and evaluated using the data of only the corresponding gender. However, note that gender dependent zt-norm was applied in both cases (i.e. even for system **MFCC13⇒39**, only z-norm segments and t-norm models of corresponding gender were used to normalize scores). The performance of these systems is demonstrated in Fig 1. On the left, we can see that the larger (GD, feature dimensionality 60) system **MFCC20⇒60** outperforms the smaller (GI, feature dimensionality 39) system **MFCC13⇒39** when evaluating on tel-tel condition. To see, whether the improvement comes from using GD models or from using different features, we have also trained GI version of **MFCC20⇒60** system, which is also shown in the figure. It seems that most of the improvement comes from the features with more detailed spectral resolution as the performances of both GD and GI versions are comparable. However, for low false-alarm region, which is the region of main interest in NIST evaluations, performance of the GD system is superior. Conversely, **MFCC13⇒39** system performs better on mic-mic trials shown on the right panel in Fig 1. The most probable reason for it is that large **MFCC20⇒60** system is overtrained to telephone data, which is the only type of data used for training UBM and speaker subspace hyperparameters. This hypothesis is also supported by the improved performance of **MFCC20⇒60** system when halving the number of system parameters by using GI instead of GD version. Unless stated otherwise, the GI version of **MFCC20⇒60** system will be used in the following experiments.

3. Analysis of JFA

3.1. Variants of Joined Factor Analysis

In the past years, different research labs adopted simplified variants of full JFA dropping some of the terms in (1) and using different methods for the hyperparameter estimation. In this

[2]Note that in [10], we have shown that similar or better results can be obtained with different approximate scoring schemes, while significantly speeding up the scoring process.

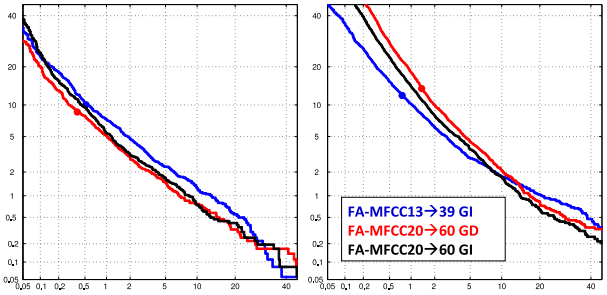

Figure 1: Performance of JFA systems based on different features and gender dependent or gender independent variants. Results on NIST 2006 data. Left panel: tel-tel trials, right panel: mic-mic trials.

section, we present a comparison of some of the JFA variants and we show that the baseline (full JFA) systems provide superior performance. Systems with only 50 eigenchannels are used in these experiments to allow for fair comparison as this was found to be the optimal number of eigenchannels for the simplified JFA variants described here.

3.1.1. Relevance MAP adaptation

The standard relevance MAP adaptation [9] can be actually seen as a special simplest case of JFA. Dropping the terms with eigenvoices and eigenchannels in equation (1), we obtain $\mathbf{M} = \mathbf{m} + \mathbf{Dz}$. For relevance MAP we simply set $\mathbf{D}^2 = \mathbf{\Sigma}/\tau$, where $\mathbf{\Sigma}$ is diagonal matrix with super-vector of UBM variances in the diagonal and τ is the relevance factor. For point MAP estimates of factors $\mathbf{z}$, it is then easy to show that the speaker model represented by $\mathbf{M}$ is equivalent to that obtained with standard relevance MAP re-estimation formulae [9].

3.1.2. Eigenchannel adaptation

The systems with *eigenchannel adaptation* [3, 2] use relevance MAP for enrolling speaker model. In the test phase, each speaker model is MAP adapted to the channel of test utterance by estimating the channel factors $\mathbf{x}$. Unlike the case of other JFA variants, PCA is used to estimate the eigenchannel matrix $\mathbf{U}$ instead of the EM algorithm. No eigenvoices are considered by this system. See [2] for thorough description of *eigenchannel adaptation* and its comparison with a system without channel compensation.

3.1.3. JFA without eigenvoices with relevance-MAP-like $\mathbf{D}$

In [6, 8], JFA systems without eigenvoices are described, where only the eigenchannel matrix $\mathbf{U}$ is trained using EM algorithm on top of the $\mathbf{D}$ matrix, which is set as in the case of the relevance MAP. On contrary to the system system based on *eigenchannel adaptation*, here, the inter-session variability is considered also for enrollment. In both [6] and [8], given the enrollment segment, MAP point estimates of factors $\mathbf{z}$ and $\mathbf{x}$ are estimated jointly using Gauss-Seidel-like iterative method. The processing of a test segment is the same as for *eigenchannel adaptation*.

3.1.4. JFA without eigenvoices with $\mathbf{D}$ *matrix trained on data*

As an alternative to the previous JFA variant, the $\mathbf{D}$ matrix in systems without eigenvoices can be also trained using EM algorithm (see the system with zero speaker factors in [4]). In

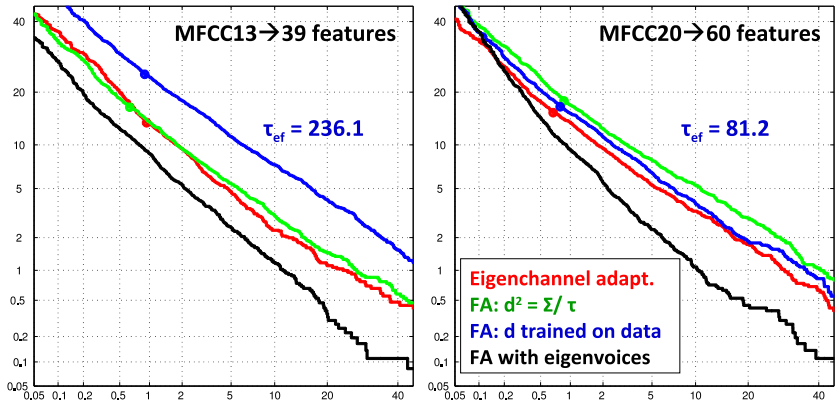

Figure 2: Flavors of JFA. Results on NIST 2006 data.

our experiment with this system, $\mathbf{D}$ matrix is trained first (unlike for the baseline system) and it is fixed for the following training of the eigenchannel matrix $\mathbf{U}$. Note also that all the data that are used for training eigenvoices in the baseline system, are now used for training $\mathbf{D}$. In the case of relevance MAP, the relevance factor τ has an intuitive interpretation. It specifies the number of frames in the adaptation data associated with a given UBM Gaussian component, which makes the MAP adaptation to shift the Gaussian component right in a half way between its original position and mean of the adaptation data. Training the matrix $\mathbf{D}$ from the data can be seen as training specific relevance factor for each coefficient of each Gaussian component. As proposed by Kenny, effective relevance factor $\tau_{ef} = trace(\mathbf{\Sigma})/trace(\mathbf{D}^2)$ can be used in this case, which can be loosely interpreted as a number of frames needed in average for each component to make the adaptation effective.

3.1.5. Results with JFA variants

The results on NIST 2006 data obtained with the JFA variants described above are shown in Fig 2. All the JFA variants without eigenvoices provide comparable performance for both types of features MFCC13⇒39 features and MFCC20⇒60. The simple eigenchannel adaptation seems to be somewhat more robust, though. The exception is the system with $\mathbf{D}$ trained on features MFCC13⇒39, which fails to perform well. The effective relevance factor $\tau_{ef} = 236.1$ for this system is significantly higher than for MFCC20⇒60 ($\tau_{ef} = 81.2$), which probably prevented the system to effectively adapt to enrollment data. The reason for this failure is still unclear and deserves further investigation. Finally, the full JFA system with eigenvoices significantly outperforms all the other JFA configurations on both feature sets.

3.2. Sensitivity of JFA to the number of eigenchannels

In Fig. 3, the three solid lines show again the performance of three JFA variants from the previous section, where 50 eigenchannels were trained for each system. The dashed lines show the change in the performance with increased number of 100 eigenchannels. We observe degradation in performance for the two variants without eigenvoices, namely the eigenchannel adaptation and the JFA with $\mathbf{D}$ trained on data. These systems seem not to be able to robustly estimate the increased number of eigenchannels. However, in the case of full JFA system, we benefit from more eigenchannels significantly after explaining the speaker variability in the model space by eigenvoices.

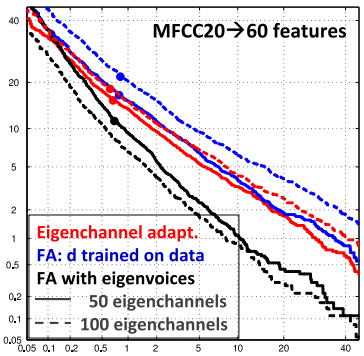

Figure 3: The effect of number of eigenchannels for JFA with $\mathbf{D}$ trained on data and full JFA. Results on NIST 2006 data.

3.3. Effect of zt-norm

The importance of using zt-norm for getting good performance with JFA systems was previously reported in [6, 5]. On contrary, our experience was that omitting zt-norm was not critical for eigenchannel adaptation based system. To verify these contradictory findings, we evaluated both eigenchannel adaptation and full JFA system with and without using zt-norm. As can be seen in Fig. 4, without zt-norm, both eigenchannel adaptation and full JFA system provide very similar performance. However, while only small gain was obtained with zt-norm for eigenchannel adaptation, dramatic improvement was obtained for full JFA system. Note again that gender-dependent zt-norm was used in both cases, which is crucial for good performance even for GI version of full JFA system. With gender-independent zt-norm (results are not shown in the figure), no significant gain was obtained for eigenchannel adaptation [2] and significant degradation in performance was observed for full JFA system compared to the system without zt-norm.

3.4. Training eigenchannels for different channel conditions

As described in section 2, our baseline JFA systems were primarily developed for telephone data. All the hyperparameters are trained on telephone data, only 100 additional eigenchannels were trained on microphone data. This strategy was already found to be effective [4] to allow the system to deal with the microphone speech segments. In Fig. 5, results are presented for all four conditions, where enrollment and test seg-

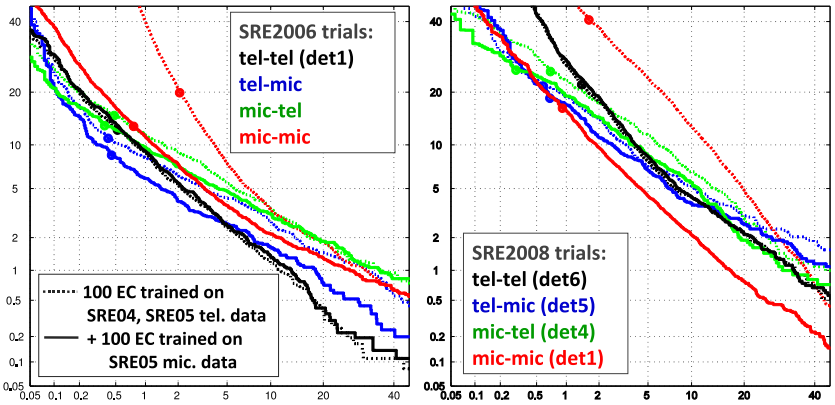

Figure 5: Performance of **MFCC13⇒39** system for different channel conditions.

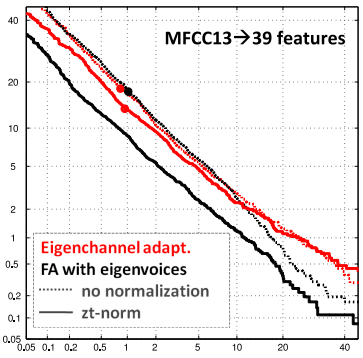

Figure 4: The effect of zt-norm. Results on NIST 2006 data.

ments are recorded either over telephone or microphone. On the left, results are presented for NIST SRE 2006 data described in section 1. On the right, results on corresponding conditions from NIST SRE 2008 evaluations[3] are presented for comparison. The dotted lines represents performance of systems with only 100 eigenchannels trained on SRE04, SRE05 telephone data while systems represented by solid lines make also use of the additional 100 eigenchannels trained also on SRE05 microphone data. We can see that augmenting the original 100 eigenchannels by those trained on microphone data brought negligible degradation for tel-tel condition and large improvement particularly on mic-mic condition. An interesting observation is that, when dropping eigenchannels trained on microphone data, much smaller degradation in performance is obtained for conditions with either enrollment or test segment recorded over telephone compared to the case where both the segments are recoded over microphone.

4. Conclusions

In this paper, we have investigated into different variants of JFA used for speaker recognition. We have shown that the full JFA with both eigenchannels and eigenvoices outperforms all simplified variants. The presence of eigenvoices allows for use of increased number of eigenchannels, which would otherwise lead to over-training of the system. We found that **gender-dependent** zt-norm was crucial for good performance of the full JFA system. This suggests, that further conditioning on other dominant speaker characteristics might be beneficial and calls for further investigation.

Although our system was primarily trained on and tuned for telephone data, JFA subsystems can be simply augmented with eigenchannels trained on microphone data, which makes the system performing well also on microphone conditions.

[3] `http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf`

5. References

[1] P. Schwarz, P. Matějka and J. Černocký: Hierarchical Structures of Neural Networks for Phoneme Recognition, In Proceedings of ICASSP 2006, May 2006, Toulouse, France

[2] L. Burget, P. Matějka, P. Schwarz, O. Glembek and J. Černocký: Analysis of feature extraction and channel compensation in GMM speaker recognition system, In: IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 7, 2007, pp. 1979-1986.

[3] N. Brümmer: Spescom DataVoice NIST 2004 system description, in Proc. NIST Speaker Recognition Evaluation 2004, Toledo, Spain, June 2004.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel: A Study of Inter-Speaker Variability in Speaker Verification, IEEE Transactions on Audio, Speech and Language Processing, July 2008.

[5] Kenny, P., Boulianne, G., Ouellet, P. and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition, In IEEE Transactions on Audio, Speech and Language Processing 15 (4), pp. 1435–1447, May 2007.

[6] R. Vogt, and S. Sridharan: Explicit Modelling of Session Variability for Speaker Verification. Computer Speech & Language 22(1), 2008, pp. 17-38.

[7] L. Burget et al.: BUT system description: NIST SRE 2008, In: Proc. 2008 NIST Speaker Recognition Evaluation Workshop, Montreal, Canada, 2008, `http://www.fit.vutbr.cz/research/view_pub.php?id=8745`

[8] D. Matrouf, N. Sheffer, B. Fauve, and J-F. Bonastre: A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification, in Proc. ICSLP 2007, Antwerp, Belgium, pp. 1242–1245, August 2007.

[9] D. Reynolds, T. Quatieri, and R. Dunn: Speaker verification using adapted Gaussian mixture models, Digital Signal Processing, vol. 10, pp. 19–41, 2000.

[10] O. Glembek et al.: Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis, In Proc. ICASSP 2009, Taipei, Taiwan, April 2009