HAL open science

Decoupling session variability modelling and speaker characterisation

Anthony Larcher, Christophe Lévy, Driss Matrouf, Jean-François Bonastre

To cite this version:

Anthony Larcher, Christophe Lévy, Driss Matrouf, Jean-François Bonastre. Decoupling session variability modelling and speaker characterisation. INTERSPEECH, Sep 2010, Makuhari, Japan. hal-01317698

HAL Id: hal-01317698

https://hal.science/hal-01317698v1

Submitted on 19 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Decoupling session variability modelling and speaker characterisation.

Anthony Larcher, Christophe Lévy, Driss Matrouf, Jean-François Bonastre

University of Avignon
Laboratoire Informatique d'Avignon - CERI/LIA - France
{anthony.larcher, christophe.levy, driss.matrouf, jean-francois.bonastre}@univ-avignon.fr

Abstract

The Factor Analysis framework demonstrated its high power to model session variability during the past years. However, training the FA parameters implies to have a large amount of training data. When the size of the available database is limited, the number of components of the core statistical model, the UBM, is also limited as the UBM drives the dimension of the FA main matrix. As the size of the UBM gives directly the size of the speaker supervector (concatenation of the GMM mean parameters), it limits also the intrinsic capacity of the recognition system, reducing the performance expectation. This paper aims to withdraw this limitation by breaking the intrinsic link between the FA dimensionality and the UBM dimensionality. The session variability modelling is done on a smaller dimension compared to the UBM, which drives the discriminative power of the system. The first experimental results proposed in this paper, done using the NIST-SRE 2008 framework, are encouraging with a relative EER improvement of about 18% when a 512 components UBM is associated to a 32 components session variability modelling compared with a 32 components UBM associated with the same variability modelling.

Index Terms: speaker verification, GMM, EigenChannel Adaptation, Session Variability

1. Introduction

The Factor Analysis framework demonstrated its high power to model session variability during the past years [**?**], [**?**]. However, training the FA parameters implies to have a large amount of training data. When the size of the available database is limited, the number of components of the core statistical model, the UBM, is also limited as the UBM drives the dimension of the FA main matrix. As the size of the UBM gives directly the size of the speaker supervector (concatenation of the GMM mean parameters), it limits also the intrinsic capacity of the recognition system, reducing the performance expectation. This paper aims to withdraw this limitation by breaking the intrinsic link between the FA dimensionality and the UBM dimensionality. This objective is supported by one hypothesis: it seems reasonable to think that the session variability modelling could be supported by a master GMM (currently the UBM) showing a smaller number of components than the intra- inter-speaker variability modelling. We propose in this paper to model the session variability using a specific UBM with a limited amount of components and to associate this session variability modelling with a larger model, used to emphasise the speaker characterisation. By separating both aspects, the dimensionality of the session variability model could be adapted to the amount of well designed available training data for the variability modelling. The main difficulty is to keep the needed tying between the components of the variability driving model (the UBM used for variability modelling) and the components of the general UBM used for the speaker recognition core engine. As a first walk in this avenue, we propose a simple way to solve this difficulty, mainly based on the UBMs design and on a direct extension of the FA matrix. The extension corresponds to a duplication of some parts of the FA matrix. Even if this solution is clearly suboptimal, it allows to evaluate the underlined hypothesis. This paper is organised as follow. Section 2 presents the baseline UBM-GMM system, including the used FA approach. Section 3 is dedicated to the FA extension process. Section 4 presents the experimental setup and the corresponding results. The last section proposes some conclusions as well as several potential extensions of the proposed approach.

2. Speaker verification system

The baseline speaker verification system is based on a standard UBM-GMM (Universal Background Model-Gaussian Mixture Modelling) paradigm [**?**]. The acoustic features used in the system are composed of 19 Linear-Frequency Cepstral Coefficients (20ms window, 10ms shift), its derivatives, the first 11 second derivatives and the delta energy. The frequency window is restricted to 300-3400 Hz. Simple feature normalisation is applied, so that the distribution of each cepstral coefficient is 0-mean and 1-variance for a given utterance. The UBM consists of a GMM trained on telephone conversations from the Fisher English database [**?**]. Variance parameters of the UBM are floored to 50% of the global variance.

A set of 2810 conversation from 124 male speakers from the NIST SRE04 were used to train a 40-rank eigenchannel matrix to model the session variability. According to the Latent Factor Analysis (LFA) modelling [**?**], speaker models are formed of three different components: a speaker and session independent background model, a speaker dependent and a session dependent components [**?**], [**?**]. The resulting model can be written as:

$$m_{(h,s)} = m + Dy_s + Ux_{(h,s)} \tag{1}$$

where $m_{(h,s)}$ is the session-speaker dependent mean supervector, D is $S \times S$ diagonal matrix (S is the dimension of the supervector), y_s the speaker vector, U is the eigenchannel matrix of low rank R (a $S \times R$ matrix) and $x_{(h,s)}$ are the session factors. Both y_s and $x_{(h,s)}$ are normally distributed among $\mathcal{N}(0,1)$. D satisfies the following equation $I = \tau D^t \Sigma^{-1} D$ where τ is the relevance factor required in the standard MAP adaptation. For scoring normalisation, when applied, 180 male speaker segments from the Fisher English database are used for zt-norm.

3. EigenChannel expansion

The aim of the work presented in this paper is to propose an approach able to optimise independently the LFA parameters and the core speaker recognition parameters, including the dimensionality of the underlined models. This objective is supported by three motivations. Firstly, only a few amount of data is available in most cases, as specific data are needed to train the U matrix of LFA. A restricted amount of training data could be a limiting factor for the dimensionality of this matrix. Moreover, it is well known that increasing the dimensionality of the UBM increases the accuracy. With the UBM usual sizes (about 2048 components), there are a lot of practical situations where it is difficult to estimate correctly the U matrix parameters. Secondly, session variability and speaker characterisation are two different phenomena and there is no straightforward hypothesis implying that they show a similar dimensionality, independently of the available amount of training data. Finally, as the objectives of both parts of the system are different, it is reasonable to think that the optimisation of both underlined statistical models could be driven by different optimisation criteria.

In this article we focus on the dimensionality aspect. We propose a method able to train the LFA U matrix with a low dimensionality (using c Gaussian components) and then to expand it to an higher dimensionality (corresponding to C Gaussian components), allowing to have a general UBM with a higher dimensionality than the session variability modelling. The proposed approach is illustrated by figure 1. The process is composed of several steps using a top-down-top approach:

- *High-Dimension*-UBM training;
- UBM size reduction to the targeted (lower) number of components;
- U matrix training (using the reduced "UBM");
- U matrix expansion to obtain a matrix corresponding to dimensionality of the *High-Dimension*-UBM.

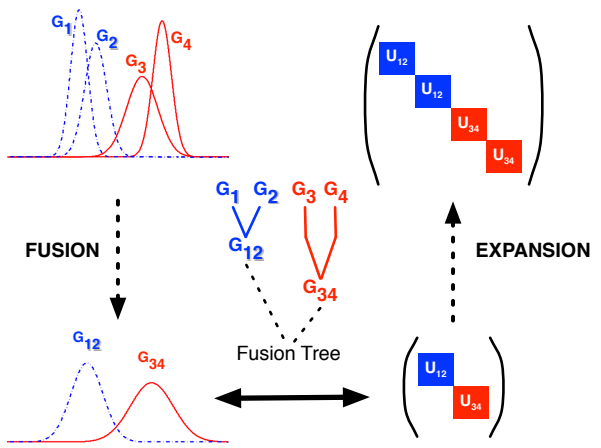


Figure 1: Global view of the EigenChannel matrix expansion process between 2-distributions and 4-distributions UBMs.

3.1. *High-Dimension*-UBM training

The *High-Dimension*-UBM training follows the classical EM-ML algorithm. This model (C components) is trained on the Fisher database.

3.2. UBM size reduction

This step aims to decrease the number of components from the *High-Dimension*-UBM until the targeted size of the U matrix. Reduction is achieved by merging the two nearest distributions ($\mathcal{N}_1(\mu_1, \Sigma_1, w_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2, w_2)$) according to the distance defined by Equation 2.

$$D(\mathcal{N}_1, \mathcal{N}_2) = \frac{w_1}{w_1 + w_2} \log(\frac{\sqrt{\Sigma}}{\sqrt{\Sigma_1}}) + \frac{w_2}{w_1 + w_2} \log(\frac{\sqrt{\Sigma}}{\sqrt{\Sigma_2}}) \quad (2)$$

where Σ corresponds to the variance of the Gaussian component that stems from $\mathcal{N}_1$ and $\mathcal{N}_2$, as defined by the Equation 5.

The Gaussian $g'(c', \mu', \Sigma')$ resulting from $g_i(c_i, \mu_i, \Sigma_i)$ and $g_j(c_j, \mu_j, \Sigma_j)$ merging is defined by:

$$c' = c_i + c_j \quad (3)$$

$$\mu' = \frac{c_i * \mu_i + c_j * \mu_j}{c_i + c_j} \quad (4)$$

$$\Sigma' = \frac{c_i}{c_i + c_j}\Sigma_i + \frac{c_j}{c_i + c_j}\Sigma_j + \frac{c_i * c_j}{(c_i + c_j)^2}(\mu_i - \mu_j)(\mu_i - \mu_j)^{tr} \quad (5)$$

The process is reiterated to reach the targeted number of components (c). All the merging steps are stored as a tree (*cf.* Figure 1).

3.3. Eigen Channel U matrix training

The LFA parameters are estimated as presented in [**?**] using the c Gaussian components UBM obtained from the previous step. For each Gaussian (denoted G_{xy} in Figure 1), the block matrix (U_{xy}) is train to model the corresponding session variability. This is the first step of the down-top phase.

3.4. Eigen Channel U matrix expansion

This step aims to expand the previously trained matrix (with c blocks) to obtain the *High-Dimension* U matrix (with C blocks, and $C >> c$).

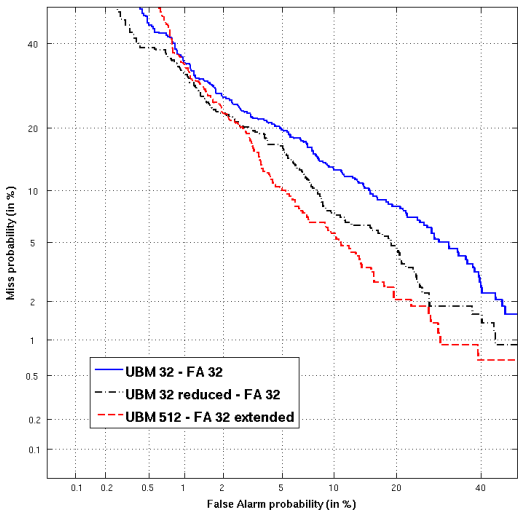
The expansion process is based on a direct duplication of blocks of the *small* LFA matrix, driven by the merging tree saved during the top-down phase. For each step, the related block is duplicated. For example, in Figure 1, G_{12} is the component resulting from the merging of G_1 and G_2 and we duplicate U_{12} for both Gaussian components; it is the same for U_{34} which is associated to G_3 and G_4. The duplication is achieved when the U matrix reaches the *High-Dimension*-UBM size (with C components).

4. Experiments

All the results reported in this paper are evaluated on the NIST SRE08 [**?**] short2-short3 condition. This condition takes one session of the target speaker for enrolment and one session for testing. Short2-short3 is divided into several conditions and we are only interested in the male condition 7 with trials involving only English language telephone speech in training and test. In this condition, 470 target speakers and 638 tests segments are used to perform 6616 verification tests. Results are reported in terms of Equal Error Rate (%EER) and described by DCF curves.

Figure 2 shows the results of a first experiment aiming to compare three systems based on a 32 components GMM-session variability modelling:

- The first system uses a 32 components UBM issued from a 512 components UBM thanks to the iterative component fusion process explained in section 3. The session variability matrix is classically learned using this reduced UBM (with dimensions corresponding to a 32 components UBM). The UBM is sub-obtimal as it is directly issued from the 512 components UBM (we don't apply any EM-ML iterations on this reduced UBM)
- The second system uses the same matrix but expanded, as explained in section 3, and associated with the 512 components UBM. It is important to notice that the expanded matrix is obtained only by the duplication of blocks of the original matrix: the number of session variability parameters is the same in first and second systems.
- The third system is proposed for comparison. It is issued from a classical 32 components UBM trained using EM-ML algorithm and a corresponding session variability matrix. The training of this matrix is driven by this new UBM. The number of session variability parameters remains unchanged compared to the two other systems.



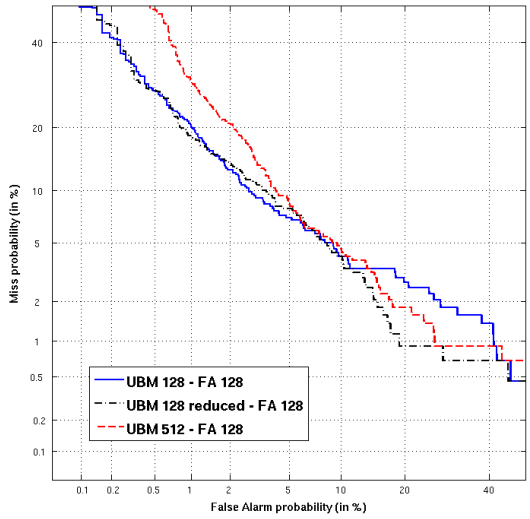
SS

Figure 2: Results for the expanded eigenchannel matrix, from 32 to 512 Gaussian distributions, without score normalisation. DET Curves for male in NIST SRE08 short2-short3 condition. English language telephone speech trials (det7).

No score normalisation is applied during this experiment. The second system performs slightly better than the first one, with an EER of about 7% to be compared to 8.6% for the first system, even if the gain is not so clear in the low false alarm part of the DET curves. This result reinforces our main hypothesis: it seems interesting to use different dimensions for session variability modelling (FA matrix) and for speaker discrimination (UBM) as using a larger UBM improves the performance, even if the FA matrix (the session variability model) remains unchanged.

The first system performs significantly better than the third system with an EER of 8.6% vs about 11% and the performance difference is clearly visible on all the DET curve. This result indicates that building a small UBM (32 components) by training a larger one (512 components) and fusing the resulting components is better than training directly the small UBM. This result confirms that the component number reduction process doesn't degrade the quality of the resulting model. When our expectation was to observe similar performance between the two systems, it is amazing to observe that the reduce model performs better than a comparable UBM, in terms of number of components, trained directly using EM-ML algorithm. As the FA matrices are trained using the corresponding UBMs (32 components, trained directly or obtained by the component fusion process), this result indicates that Maximum Likelihood criterion is not always the best criterion to estimate an UBM dedicated to FA training.

Figure 3 presents the results of an experiment similar to the previous one except the higher dimension of the session variability modelling, which is now based on a 128 components UBM. The two systems based on 128 components UBMs (one trained using EM-ML and one issued from the component reduction process applied on a 512 components UBM) show comparable performance. This observation confirms the fact that the component fusion approach gives good quality models for FA matrix estimation.

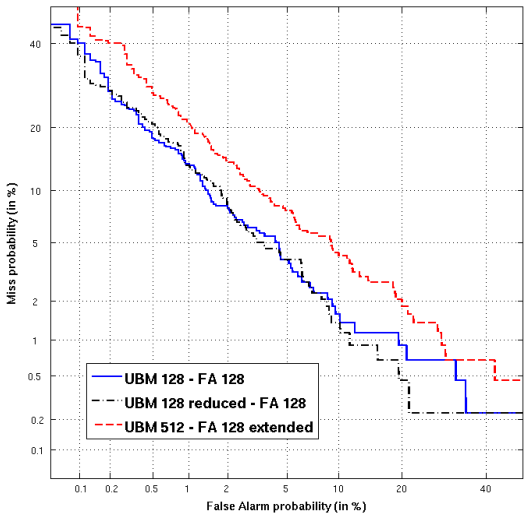


SS

Figure 3: Results for the expanded eigenchannel matrix, from 128 to 512 Gaussian distributions, without score normalisation. DET Curves for male in NIST SRE08 short2-short3 condition. English language telephone speech trials (det7).

The system associating a 512 components UBM with the 128-based FA matrix (expanded as before) obtains an EER very similar than the ones of the other systems (about 6% for the three systems) but it seems to perform lower than the other systems on the rest of the DET curve. Figure 4 presents the results of the same experiment when applying a score normalisation (z-norm applied after t-norm). Using the score normalisation, it appears very clearly that the 512/128 system performs lower than the two other systems, with an EER of about 6% to be compared with EERs of about 4.6% for the others. This difference is noticeable on all the DET curve. This quite disappointing result could come from the over-simple FA matrix expansion used in this work: the expanded FA matrix is only a duplication of blocks of the original matrix. With a larger number of components (128 components for the second experiment to be compared with 32 for the first one), it seems that the tying between the original UBM components (512) and one of the per-

component part of the FA matrix, *i.e.* with the underlined 128 components UBM, is not good enough to authorize a simple duplication process.



ss

Figure 4: Results for the expanded eigenchannel matrix, from 128 to 512 Gaussian distributions, with ZT-normalisation. DET Curves for male in NIST SRE08 short2-short3 condition. English language telephone speech trials (det7).

5. Conclusions

The direct modelling of the session variability, thanks to Factor Analysis framework (and the correlated approaches like Joint Factor Analysis or Nuisance Attribute Projection) is embedded in the main part of state-of-the-art speaker recognition systems, due to the huge performance improvement linked to it.

This UBM-GMM-FA architecture shows an important slide effect: the variability modelling and the core speaker recognition system share the same UBM. The role of the UBM is in fact different for the two subtasks. For the session variability modelling, the UBM is used to drive the variability modelling, working mainly as an automatic frame labelling system. Due -at least- to the training data constraints, it seems reasonable to think that the dimension of this UBM should be limited. For the core speaker recognition engine, the UBM dimension gives directly the number of parameters used to describe a given speaker, i.e. the intrinsic capacity of the recogniser. It is well known that a large UBM is needed in this case. Furthermore, using an unique model implies to use the same training criterion, which is suboptimal considering that the objectives are different.

This paper investigated a part of this problem and proposed a strategy in order to work with two different UBMs, one for the session variability modelling and one for the core system. It studied mainly the effects linked to the number of components in both models. Even if the strategy used in order to work with a different number of components in the two parts of the system was very simple -a simple duplication of blocks of the FA matrix- the results showed the possibilities of this approach and the interest of decoupling the UBM models. After this first step in this avenue, it is easy to propose several additional works. Firstly, a more complex strategy to deal with the difference in terms number of components in the UBMs should be investigated. A first solution could be to apply the FA at the frame level [**?**] in order to separate completely the session variability problem and the core speaker recognition engine. This solution seems risky as one of the main advantages of the global UBM-GMM architecture is the structuring role of the UBM [**?**]. An interesting alternative consists in the combination of all the FA components in order to propose a well suited set of FA parameters for a given (core engine) UBM component. The optimisation of the expanded matrix is also a good option, taken alone after the simple process showed in this paper or applied after the latter proposal.

A complete study on the optimal model dimensionality for the session variability modelling is also a promising investigation. Fixing the core system thanks to the solution proposed in this paper will certainly help, by freezing a part of the free factors, but it seems preferable to estimate directly the variability modelling power of the FA subsystem from the FA parameters and some data, using a cross validation approach. Finally, one of the main advantages of decoupling the two UBMs is to optimise both models using dedicated criteria. It opens the opportunity to add more discriminative aspects in these models, as it is done for example in the language recognition field.

6. Acknowledgements

The work presented in this paper have been done related to the MoBio project. MoBio (Mobile Biometry) is a FP7 european project founded by European Community. `http://www.mobioproject.org/`