Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Factor analysis-based approaches applied to the speaker diarization task of meetings: a preliminary study

Pavel Tomasek, Corinne Fredouille, Driss Matrouf

University of Avignon, CERI/LIA, Avignon France
xtomas23@stud.fit.vutbr.cz, (firstname.lastname)@univ-avignon.fr

Abstract

This paper presents a preliminary study on the use of the Factor Analysis (FA) methods in an automatic speaker diarization process, dedicated to the meeting rooms. Indeed, the speaker diarization process, based on the top-down E-HMM approach, integrates a FA-based speaker modeling in an additional resegmentation step, which aims at helping the refinement of the output segmentation. Classically applied in speaker recognition to deal with channel variability issues, two main schemes of the FA application are proposed here: to deal with the (1) inter-speaker variability and with (2) the inter-segment variability. Different kinds of experiments have been conducted on the dataset of the last NIST/RT'09 evaluation campaign, leading to very interesting and promising results. For instance, they show that the couple of schemes proposed in this paper obtained competitive performance, compared to the baseline process, despite the small amount of development data used in this paper for the FA parameter estimation. Unexpectedly, they tend to show that the inter-segment variability component can be helpful for speaker diarization.

1. Introduction

The speaker diarization task, also known as "Who spoke when" consists in determining the speaker turns automatically in an audio document and grouping together the speech segments belonging to the same speaker [1]. This task is required in a more overall process, called Rich Transcription (RT), aiming at providing different kinds of labels (types of acoustic environment, speaker gender, speaker turns, spoken terms, named entities, etc), used afterwards for indexing purposes of audio documents.
For the last ten years, the speaker diarization task has been involved in various domains, such as telephone conversations, broadcast news, and more recently meetings, mainly driven by the international RT evaluation campaigns organized by the National Institute of Standards and Technology (NIST). If each domain has its own characteristics, they can present a large variability over audio documents as well as inside a same document, hardly managed by automatic systems.
In other fields of the automatic speech processing, factor analysis (FA) methods have demonstrated their reliability in dealing with variability issues. They have been firstly applied successfully in speaker recognition, reaching very large performance improvement on telephone data [2, 3, 4]. Then, they have been involved in other fields like for instance language recognition [5, 6] or video gender detection [7]. More recently, some attempts to exploit the effectiveness of the FA approaches in the speaker diarization field have been proposed [8]. Mostly, these studies are focused on the use of speaker factors to enhance the robustness of the speaker models involved in the speaker diarization process. As opposed to the previous studies on FA (especially in speaker recognition), channel factors are not involved to deal with the inter-session variability. Moreover, in [8], FA-based approaches are applied in the specific context of telephone conversations (2 speaker conversations), issued from the NIST/SRE evaluation campaign of the speaker recognition systems for which large corpora are available as recommended for the FA implementation.
In this paper, the authors propose to investigate the use of the FA-based approaches for the speaker diarization task in the meeting context (multiple speaker conversations). Here, a top-down speaker diarization approach, based on the E-HMM scheme [9], is used jointly with FA-based approaches to enhance speaker model quality. Different approaches, involving speaker factors only and both speaker and channel factors are evaluated through meeting data issued from the last NIST/RT evaluation campaign [10]. These approaches aim at modeling two main kinds of variability: (1) the inter-speaker variability inside a meeting show, and (2) the inter-segment variability for a same speaker inside a show.
In this context, the paper is organized as follows: section 2 gives a description of the top-down approach-based system, used in this paper. Section 3 presents the basis of the FA. After detailing the experimental protocol used in this paper, the joint use of the top-down and the FA approaches is described in sections 5 and 6, coupled with the results obtained for each case. Finally, a discussion followed by some perspectives are proposed in section 7.

2. Baseline speaker diarization system

The diarization system employed in this paper is developed using the open source ALIZE speaker recognition toolkit [11]. It involves 3 main steps, in addition to some preprocessing to accommodate multiple channels:
• a speech activity detection (SAD) process, required to remove non-speech segments from the speaker diarization process.
• a speaker segmentation and clustering process to detect speaker turns and group together speaker homogeneous segments, and
• a resegmentation process, to refine the output segmentation.

2.1. Multi-channel handling

The speaker diarization task involved in this paper relates to multiple distant microphones located on meeting room tables (MDM task of the NIST/RT evaluation plan [10]). To deal with this task, a single virtual channel is formed using the BeamformIt 2.0 toolkit[1] with a 500 ms analysis window and a 250 ms frame rate.

2.2. Speech Activity Detection

The Speech Activity Detection (SAD) algorithm employs feature vectors composed of 12 un-normalized Linear Frequency Cepstrum Coefficients (LFCCs) plus energy augmented by their first and second derivatives. It utilises an iterative process, coupling both a Viterbi decoding and a model adaptation scheme applied to a two-state HMM. States represent speech and non-speech events and each one is associated with a 32-component Gaussian Mixture Model (GMM), trained on separate data using an EM/ML algorithm. State transition probabilities are fixed to 0.5. Finally, duration based-rules are applied in order to refine the speech/non-speech segmentation yielded by the iterative process.

2.3. Speaker segmentation and clustering

This step is the core of the LIA speaker diarization system. It relies on a one-step segmentation and clustering algorithm, following a top-down scheme in the form of an Evolutive Hidden Markov Model (E-HMM) [9]. Each E-HMM state aims at characterizing a single speaker and the transitions represent the speaker turns. Here the signal is characterized by 20 LFCCs, computed every 10 ms using a 20 ms window. The cepstral features are augmented by energy and no feature normalization is applied.
As detailed in [12], the segmentation process begins by initializing the E-HMM with only one state (denoted $L0$) representing the entire audio show. An iterative process is then started where a new speaker/state is added to the E-HMM at each iteration. Successive Viterbi decoding and speaker model training loops attribute speech segments to the different speakers involved in the E-HMM. This iterative process is performed until a stop criterion is reached, which is based on the ability, or not, for a new speaker to be added to the E-HMM.
At each iteration, a new speaker is added and associated with the longest segment selected among those assigned to $L0$ speaker and considered as unlabelled yet. The selection of this segment is constrained by a minimum 6s duration. If no segment belonging to $L0$ responds to this criterion, the iterative process is stopped.
The speaker modelling involved in the iterative process is based on Gaussian Mixture Models (GMM), estimated through the EM/ML (Expectation - Maximization / Maximum Likelihood) algorithm. GMM models associated with each HMM state are composed of 16 Gaussian components (with diagonal covariance matrix) except for the last speaker added, for which only 8 Gaussian components are estimated. This difference in the number of Gaussian components aims at balancing the assumed small amount of data attributed to the last speaker compared with the others.

2.4. Resegmentation steps

The speaker segmentation and clustering process is followed by a couple of resegmentation steps, used to refine the segmentation outputs. For each of them, an HMM is generated from the previous stage outputs (and set of speakers associated with) and an iterative speaker model training/Viterbi decoding loop is launched. For this iterative process, all the boundaries (except those of speech/non-speech segments) as well as the segment labels are re-examined. Moreover, a speaker/state can be deleted after an iteration if it does not attract enough speech segments (less than 8 seconds). In contrast to the segmentation and clustering stage, here a Maximum A Posteriori (MAP) [13] based-adaptation (coupled with a Universal Background Model (UBM)) replaces the EM/ML algorithm for speaker model estimation since the segmentation step provides an initial distribution of speech segments among the set of speakers detected.
If the first resegmentation step relies on the same features as the segmentation and clustering process (20LFCC plus energy without any feature normalization), the second resegmentation step employs 16LFCCs, energy, and their first derivatives, extracted every 10 ms using a 20ms window, making up a feature vector of 34 coefficients. As opposed to the previous steps, the parameter vectors are normalized, segment-by-segment[2], to fit a zero-mean and unity-variance distribution. This type of parameterization, especially the feature normalization, is rather typical of the speaker recognition domain and may be effective to correct some minor segmentation errors in a later stage.

[1] Available at: http://www.icsi.berkeley.edu/xanguera/beamformit

[2] segments are issued from the output segmentation yielded by the previous step

3. Factor analysis-based methodology

3.1. GMM-UBM speaker recognition and FA paradigm

GMM models are linear combinations of Gaussians, generally used for approximating a complex probability density function. A GMM is defined by a set of M Gaussians $\mathcal{N}(.|\mu_g, \Sigma_g)$, along with their associated weights α_g ($g \in 1, ..., M$):

$$\sum_{g=1}^{g=M} \alpha_g \mathcal{N}(.|\mu_g, \Sigma_g). \quad (1)$$

The GMM-UBM framework is a standard in speaker verification [14]. It is also used in other audio classification tasks, such as language recognition. The UBM, also called generic or world model, is a GMM that represents all the possible observations. For each target pattern (language, speaker,...), a specific GMM is obtained by adapting the UBM *via* the MAP criterion [13]. Only GMM means are adapted, the other GMM parameters are taken from the UBM without any modification.
For the FA paradigm, we need to define the GMM mean super-vector concept. A GMM mean super-vector is defined as the concatenation of the GMM component means. Let D be the dimension of the feature space, the dimension of a super-vector mean is $M \cdot D$, where M is the number of Gaussians in the GMM. In order to ease the understanding of the FA development, we introduce the following matrix notation: let $\mathbf{A}$ be a $MD \times K$ matrix formed by concatenating vertically M matrices of dimension $D \times K$. Let us denote by $\{\mathbf{A}\}_{[g]}$ the g^{th} matrix in $\mathbf{A}$ (usually corresponding to the g-th component in the model). Let this GMM be parameterized by $\theta = \{\mathbf{m}_{[g]}, \mathbf{\Sigma}_g, \alpha_g\}_{g=1}^{M}$, where $\mathbf{m}_{[g]}$, $\mathbf{\Sigma}_g$, α_g are the mean, the covariance matrix and the weight of the g-th Gaussian in the GMM; $\mathbf{m}$ denotes the mean GMM super-vector, which is the concatenation of the GMM means $\mathbf{m}_{[g]}$. $\mathbf{\Sigma}$ is the block diagonal matrix where the g-th diagonal block is $\mathbf{\Sigma}_g$.

The term *session variability* encompasses a number of phenomena including transmission channel effects, environment noise (other people, cars, TV, etc.), variable room acoustics (hall, park, etc.), microphone position relative to the mouth, and the variability introduced by the speaker himself/herself. The solutions proposed in the literature involve work at various levels of the AAP (feature space, model space and score space). In spite of the use of sophisticated feature extraction modules, the session variability introduces a bias in estimated model parameters. This bias could dramatically influence the classification performance. This is mainly caused by the fact that the training databases cannot offer an exhaustive coverage of all the potential sources of session variability.

In order to take into account the session variability in the modeling process, the factor analysis model for session h belonging to speaker s can be written as:

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(h,s)}, \quad (2)$$

where $\mathbf{m}_{(h,s)}$ is the session-speaker dependent super-vector mean, $\mathbf{D}$ is a $(MD \times MD)$ diagonal matrix, $\mathbf{y}_s$, the speaker vector (a MD vector), $\mathbf{U}$ is the session variability matrix of lower rank R (a $MD \times R$ matrix) and $\mathbf{x}_{(h,s)}$ are the channel factors, an R vector (theoretically, $\mathbf{x}_{(h,s)}$ is independent of s). Both, $\mathbf{y}_s$ and $\mathbf{x}_{(h,s)}$ are normally distributed among $\mathcal{N}(0, I)$. $\mathbf{D}$ satisfies the equation $\mathbf{I} = \tau \mathbf{D}^t \mathbf{\Sigma}^{-1} \mathbf{D}$, where τ is the *relevance factor* required in the standard MAP adaptation ($\mathbf{D}\mathbf{D}^t$ represents the *a priori* covariance matrix of $\mathbf{y}_s$).

3.2. Application of FA in speaker diarization

Given the success met by the use of the FA in the speaker recognition domain, we were tempted to use it in the speaker diarization domain. However, the nature of the session variability problem is not the same in the two domains. In the speaker diarization problem, since we use a single virtual channel (typically, to deal with the multimicrophone context, see section 2.1), we can consider that the same microphone is used for all speakers and for all speech segments. In this paper, we will experiment two main hypotheses:

- **Inter-speaker variability:** here we assume that the main speaker information can be located in a low dimension sub-space, and the rest of speaker information in the full space. We think that this speaker modeling manner can be helpful in the case of small amounts of training data. Indeed, the number of parameters representing the speaker in the low dimension sub-space is very small with respect to the dimension of the full space, which makes them estimated robustly from small amounts of data. In this context, the model equation for a given speaker can be represented as follows:

  $$\mathbf{m}_s = \mathbf{m} + \mathbf{D}\mathbf{y}_{(s,full)} + \mathbf{U}\mathbf{x}_{(s,low)}, \quad (3)$$

  $\mathbf{D}\mathbf{y}_{(s,full)}$ is the speaker part in the full space, and $\mathbf{U}\mathbf{x}_{(s,low)}$ is the speaker part related to the low dimension space. The $\mathbf{U}$ matrix is common to all speakers and is estimated from a development dataset.

- **Inter-segment variability:** Here, we assume that the inter-segment information can be located in a low dimension sub-space. The FA-based speaker model can therefore be written as follows:

  $$\mathbf{m}_s = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(s,seg)}, \quad (4)$$

$\mathbf{Dy}_s$ is the speaker part in the full space, and $\mathbf{Ux}_{(s,seg)}$ is the speaker-segment component. The **U** matrix is common to all speakers and is estimated from a development dataset.
Regarding the classical FA paradigm, the **U** matrix estimation (using a development dataset) and its application during testing phase (using an evaluation dataset) have to be performed in the same experimental conditions to be efficient. Here, since the **U** matrix is estimated by using the inter-segment information, it is difficult to respect this constraint in this preliminary study since this implies that a speaker model were estimated per segment. Consequently, it important to note that, in this paper, all the segments belonging to the same speaker are merged and considered as a unique segment utilized for the speaker modeling in the speaker diarization process (testing phase). The use of segment-speaker-dependent models (several models for each speaker) will be studied in future work.

For both hypotheses, it is worth noting that the FA-based speaker modeling approaches presented above are integrated within a third resegmentation process of the baseline speaker diarization system. This additional process follows the same features as the second resegmentation step described in section 2.4.

4. Experimental protocol

Experiments reported in this paper have been conducted according to the evaluation plan of the NIST/RT'09 campaign [10], focusing on the Multiple Distant Microphone (MDM) condition. Seven meeting files are available for this evaluation, described in table 1 through their orginal file name, the short name used later in the paper, their length (in seconds), and the number of speakers involved in the meeting. Performance of the baseline speaker diarization system (described in section 2) is also provided here. Performance is given in terms of Diarization Error Rate (DER). These files are used in next sections as the evaluation dataset permitting comparisons between the different approaches proposed in this paper.
A second dataset (named development dataset), composed of 23 meeting files issued from the previous RT evaluation campaigns, is used to estimate the set of parameters involved in the FA-based system (notably **U** matrix). The reference segmentations have been used for this estimation, for which about 100 speakers are present and about 86% of speech segments are less than 3 second long (55% between 0 and 1s, 22% between 1 and 2s, and 9% between 2 and 3s).
In this preliminary study, all the parameters of FA-based systems - rank of the **U** matrix, number of iterations for speaker modeling - have been empirically tuned a posteriori on the evaluation dataset.

Original name	*Short name*	*Length (in s.)*	*Speaker Nb*	*%DER*
EDI-2007 1128-1000	EDI-10	1472	4	3,2
EDI-2007 1128-1500	EDI-15	1410	4	33,8
IDI-2009 0128-1600	IDI-16	1708	4	15,0
IDI-2009 0129-1000	IDI-10	1476	5	14,1
NIST-2008 0201-1405	NIS-14	1146	5	47,9
NIST-2008 0227-1501	NIS-15	1091	6	20,4
NIST-2008 0307-0955	NIS-09	1223	7	18,7
Overall	*Overall*	*1360*	*5*	*18,9*

Table 1: Evaluation dataset: Each meeting, used as evaluation dataset in the experiments, is characterized by their original name, a short name used in the experimental section, their length (in seconds), the number of speakers present in the meeting, as well as the performance in terms of %DER obtained by the baseline speaker diarization system.

5. Inter-speaker variability

This section compares performance obtained by the baseline system with and without the implication of the FA based-speaker modeling related to the inter-speaker variability configuration. The performance comparison, given in terms of %DER for each meeting files of the evaluation dataset, is presented in table 2. The second and the third columns report DER scores of FA-based systems with and without the implication of the low dimension speaker information ($\mathbf{Ux}_{(s,low)}$) in the speaker modeling process respectively.

*Note: The **U** matrix rank was set to 100. The number of training iterations for speaker modeling was set to 1.* There is a large difference between overall averages of FA using speaker modeling with and without $\mathbf{Ux}_{(s,low)}$ component. The comparison between results of tables 1, providing baseline system performance and 2 shows that the low dimension sub-space contains relevant information about the speaker, but not enough to increase performance with respect to the baseline system. Nevertheless, this result is rather expected, because the number of speakers used to train the **U** matrix is quite small regarding the rank of this matrix (100). Indeed, only about 100 speakers are present in the development dataset used in this paper. Note that the system corresponding to the third column is like MAP adaptation but applied on data from which the term **Ux** is subtracted (at frame level).

	DER (%) of FA-based systems	
	$\boldsymbol{m} + \boldsymbol{Dy}_{(s,full)} + \boldsymbol{Ux}_{(s,low)}$	$\boldsymbol{m} + \boldsymbol{Dy}_{(s,full)}$
EDI-10	3.2	39.5
EDI-15	34.6	37.0
IDI-16	14.9	33.9
IDI-10	15.7	16.5
NIS-14	49.4	46.1
NIS-15	16.6	19.5
NIS-09	19.3	18.7
Overall	*19.1*	*29.6*

Table 2: Inter-speaker variability: %DER obtained on the evaluation dataset by FA-based systems using different speaker modeling schemes (involving or not $\mathbf{Ux}_{(s,low)}$ component).

	DER (%) of FA-based systems	
	$\boldsymbol{m} + \boldsymbol{Dy}_s + \boldsymbol{Ux}_{(s,seg)}$	$\boldsymbol{m} + \boldsymbol{Dy}_s$
EDI-10	3.1	3.2
EDI-15	34.3	47.0
IDI-16	14.6	12.3
IDI-10	15.6	14.0
NIS-14	49.4	42.0
NIS-15	13.0	19.0
NIS-09	19.1	18.1
Overall	*18.5*	*19.3*

Table 3: Inter-segment variability: %DER obtained on the evaluation dataset by FA-based systems using different speaker modeling schemes (involving or not $\mathbf{Ux}_{(s,seg)}$ component).

6. Inter-segment variability

Three experiments are presented in this section. The first one aims at measuring the impact of the FA-based speaker modeling related to the inter-segment variability. Two kinds of results are presented with ($\mathbf{m} + \mathbf{Dy}_s + \mathbf{Ux}_{(s,seg)}$) and without ($\mathbf{m} + \mathbf{Dy}_s$) using the inter-segment component. In the later case, the inter-segment component ($\mathbf{Ux}_{(s,seg)}$) is firstly estimated and discarded.

The second experiment presents results of FA-based systems using some constraints in channel matrix estimation (short segment filtering). The last experiment shows results of FA-based systems, when a standard (no FA) resegmentation step is applied after the FA based-process.

6.1. Implication of the inter-segment component

Results (in terms of %DER) are shown in table 3. The second and the third columns contain scores of FA-based systems involving or not the inter-segment component ($\mathbf{Ux}_{(s,seg)}$).

*Note: The **U** matrix was estimated only on segments equal or longer than 1 second and the matrix rank was set to 10. The number of training iterations for speaker modeling was set to 1.*

As opposed to the previous section, there is not a large difference between overall averages. The inter-segment component seems to contain some information, but its influence in the speaker diarization process is not as large as the one observed in the approach tested in the previous section (Inter-speaker variability). Indeed, %DER are rather variable between meeting files, regarding the different speaker modeling schemes. Especially, the slight increase of the overall %DER in the third column is mainly due to the *EDI-20071128-1500* meeting file for which the %DER augments drastically (from 33.8% for the baseline to 47.0% for the FA: $\mathbf{m} + \mathbf{Dy}_s$ configuration) when the inter-segment component is not involved while stable or decreasing %DER can be observed for the other meeting files. Conversely, regarding now the use of the inter-segment component (second column), the decrease of the overall %DER is mainly due to the significant gain observed on the *NIST-20080227-1501* meeting file.

6.2. Experiment with segment filtering for U matrix estimation

The basic idea of this experiment is to remove very short segments (hundreds of milliseconds) which could be disturbing for the **U** matrix estimation involved in the inter-segment component. In other words, the **U** matrices, involved in this section, differ in the minimal segment duration used in their estimation process. Table 4 reports DER scores of FA-based systems, for which **U** matrices have been estimated either on all the segments available, or on segments equal to or longer than 1, 2, 5, or 10 seconds.

*Note: Speaker modeling is based on the $\boldsymbol{m} + \boldsymbol{Dy}_s + \boldsymbol{Ux}_{(s,seg)}$ scheme. The **U** matrix rank was set to 10. The number of training iterations for speaker modeling was set to 4*[3]. In these results, we can observed an interesting difference between the overall DER average of the FA-based system combined with **U** matrix estimated on all the segments available and the system using **U** matrix estimated on segments equal to or longer than 1 second. Indeed, the later outperforms the first system and exhibits the lower DER compared with the other systems (2, 5, and 10 second long segments). This behaviour is mainly due to the significant gain observed on the *NIST-20080227-1501* meeting file where the DER declines from 16.7% (with all the segments available) to 4.9%, which represents also a 15.5% absolute gain compared to the baseline system (without FA). Regarding the other meeting files, quite stable performance is observed in this configuration.

[3]This change in the number of iterations, compared to the previous sections, is induced by some tuning experiments, not reported here, showing best a posteriori performance on the evaluation dataset for this figure.

	DER (%) per min. segment duration				
	0 s	*1 s*	*2 s*	*5 s*	*10 s*
EDI-10	3.2	3.1	3.1	3.0	3.0
EDI-15	34.5	34.3	34.3	34.2	34.0
IDI-16	14.8	14.6	14.6	14.7	14.6
IDI-10	15.8	15.6	15.7	15.6	15.8
NIS-14	49.4	49.2	49.4	49.3	48.9
NIS-15	16.7	04.9	15.0	18.6	19.2
NIS-09	19.2	19.1	19.2	19.2	18.9
Overall	*19.1*	*17.7*	*18.8*	*19.1*	*19.1*

Table 4: Inter-segment variability: %DER obtained by FA-based systems using different **U** matrices. Difference appears in the minimum duration of segments used for the matrix estimation process.

6.3. Combination of FA with resegmentation step

The FA-based system providing new segmentation outputs, it would be interesting to observe the behaviour of an additional resegmentation step as used in the baseline system. The hypothesis is that the segmentation output altered by the FA-based resegmentation process can be potentially enhanced by a new step of resegmentation without FA.

Note: Speaker modeling is based on the $\boldsymbol{m} + \boldsymbol{D}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{x}_{(s,seg)}$ *scheme. The* **U** *matrix rank was set to 10. The number of training iterations for speaker modeling was set to 4. Segments equal or longer than 1s are used only for the* **U** *matrix estimate.* Results, presented in table 5,

	DER (%)	
	FA	*FA + Resegmentation*
EDI-10	3.1	3.3
EDI-15	34.3	34.8
IDI-16	14.6	14.6
IDI-10	15.6	15.1
NIS-14	49.2	47.3
NIS-15	4.9	3.5
NIS-09	19.1	18.8
Overall	*17.7*	*17.3*

Table 5: Inter-segment variability: %DER without and with the use of a basic resegmentation (without FA) step after applying FA-based process.

shows, as expected, a slight improvement of the overall DER scores after applying a successive resegmentation step. Notably, a further gain can be observed on the *NIST-20080227-1501* meeting file for instance.

7. Discussion

The Factor analysis is an interesting approach allowing robust estimation of certain parameters for which a small amount of data is available. The basic idea is to decompose such parameters into 2 parts: the first one (**U**) containing a large amount of parameters and can be estimated on a large amount of development data, and the second part (**x**) contains small number of parameters and can be estimated on test data itself. In this paper we have tried to apply this paradigm to the following types of information: speaker and segment. For the speaker component, we need to use a large amount of speakers in the development dataset. Regrettably, it was not the case currently, since only 100 speakers are available in the development dataset used, while we need more than a thousand. In spite of that, we have shown that the low dimension sub-space contains a lot of information about the speaker.

For the segment information, we have shown that this component contains some information about the speaker. Perhaps, some of this information does not concern the intrinsic characteristics of the voice speaker, but concerns his or her mouth position with respect to the microphones. For most of the meeting files used in the evaluation dataset, the gain obtained by modeling the segment component was small, except for the meeting file named *NIST-20080227-1501*, for which the gain was very impressive: 76% relative gain. Obviously, our future work will be focused now on finding some explanation for this phenomenon: why the modeling of the inter-segment variability with FA has shown very good performance for only one file among the 7 files used ? Moreover, as said in section 6, we will also study the possibility of using segment-speaker-dependent models. Indeed, the speaker model containing two components - one is common to all speaker segments and the other is specific to each segment - it will be more natural (regarding the FA paradigm) to handle a set of segment-speaker-dependent models rather than a single speaker model (as done in this paper). Obviously, it will be considered the possibility that a set of close segments can share the same segment-speaker-dependent component.

In the case of segment component modeling with FA, the additional application of the re-segmentation using the original strategy, i.e. the standard MAP adaptation, gives a supplementary gain. Before the segment component modeling, the Viterbi segmentation based on MAP modeling met a local maximum which can be left to go toward a better solution thanks to FA-based speaker modeling.

It is important to note that the FA analysis have met large success when coping with channel variability. We know also that, in speaker diarization tasks applied to meeting rooms, several microphones may be generally available, leading to several speaker recordings for the

same speech production. Consequently, instead of using a single virtual channel (formed using the BeamformIt), the FA could be used to model the differences between these recordings, hence all microphones can be used efficiently for training the speaker models. This investigation will be one of our further work.

8. References

[1] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[2] L. Burget, P. Matejka, O. Glembek, P. Schwarz, and J. Cernocky, "Analysis of feature extraction and channel compensation in gmm speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15(7), 2007.

[3] 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*, vol. 16(5), 2008.

[4] R. J. Vogt and S. Sridharan, "Explicit modeling of session variability for speaker verification," *Computer Speech and Language*, vol. 22(1), 2008.

[5] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15(7), 2007.

[6] F. Verdet, D. Matrouf, J.-F. Bonastre, and J. Hennebert, "Factor analysis and svm for language recognition," in *Proceedings of Interspeech'09*, September 2009.

[7] M. Rouvier, D. Matrouf, and G. Linares, "Factor analysis for audio-based video genre classification," in *Proceedings of Interspeech'09*, September 2009.

[8] R. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in *Proceedings of Interspeech'09*, September 2009.

[9] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *2001 : a Speaker Odyssey. The Speaker Recognition Workshop*, Chania, Creete, June 2001, pp. 175–180.

[10] NIST, "The NIST Rich Transcription 2009 (RT'09) evaluation," http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, 2009.

[11] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005)*, Philadelphia, USA, March 2005.

[12] C. Fredouille and N. W. D. Evans, "New implementations of the e-hmm-based system for speaker diarization in meeting rooms," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2008)*, April 2008.

[13] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," in *IEEE Transactions on Speech and Audio Processing*, Avril 1994, vol. 22, pp. 291–298.

[14] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 1rst April 2004.