Application of Factor Analysis to Speaker Diarization

Bc. Pavel Tomášek,
xtomas23@stud.fit.vutbr.cz

January 24, 2010

Abstract

This is a report of my Socrates ERASMUS internship
in *Laboratoire Informatique d'Avignon*
(September 2009 – January 2010, home university:
Brno University of Technology – Faculty of Information Technology).
This work aims at experimenting with application of factor analysis to speaker
diarization. Both speech processing techniques are also briefly described.

Contents

Chapter 1

Introduction

We live in time when information may have very high value but also the quantity of data everywhere (broadcast news, internet) keeps growing. To be able to quickly extract some specific and needful information we need systems which can properly manage the data. Such a system can also use speaker diarization for audio data. Speaker diarization can be helpful in indexation of audio databases or speaker adaptation for speech recognition. In case that we have to work with very long recordings the distance of a person and a microphone (and other acoustic characteristics like noise or background environment) can change. This is the reason for experimenting with factor analysis, which can remove such a disturbing variability during the time of recording and improve the diarization process.

This work is aimed at application of factor analysis to speaker diarization. In the following chapters readers can find a description of used diarization architecture in detail with basis of applied factor analysis.

This is a report of my Socrates ERASMUS internship (from September 2009 to January 2010) in *Laboratoire Informatique d'Avignon* (LIA). My supervisors were Corinne Fredouille and Driss Matrouf.

1.1 Context and Motivation

Speaker diarization give us an answer on question "Who spoke when?" In which situations, in which fields can be such an answer appreciated? It can be useful for instance in speech recognition systems where it can be one way of speaker adaptation to improve the recognition results. Another useful application can be in audio indexing. If there is huge amount of audio data (an audio library) we are able to find all the utterances of a certain person. It can be also useful for people who work with huge databases of audio recordings (a TV or radio companies and also military and security services).

Previous paragraph was only about the speaker diarization. And what about the factor analysis? Factor analysis is a very helpful instrument in speaker verification and also in language identification. The experiments of this report try to show if factor analysis can be also suitable in speaker diarization domain. That means if factor analysis helps speaker diarization produce better scores.

1.2 Contents

The main body of this report is divided as follows: next chapter, 2, includes an introduction of used speaker diarization system. Chapter 3 describes the basis of factor analysis in general and includes a few ways of application of joint factor analysis with their results. The last chapter, 4, reviews what has been done in this work and mentions the most important questions and possible future work. Then comes the Bibliography, listing all sources of material.

Chapter 2

Speaker Diarization

In this chapter the diarization system used for experiments is described. Following paragraphs also contain information about the way how to obtain such a system and how to make it run.

For obtaining the base knowledge of diarization process it is good to read the following paper: [7]. Thesis of Xavier Anguera [1] is also suitable to get more detail information about all the particular parts of such a diarization system.

Simply, speaker diarization tries to find answer for question: *"Who spoke when?"* The output of a speaker diarization system specifies when which speaker was speaking (see figure 2.1).

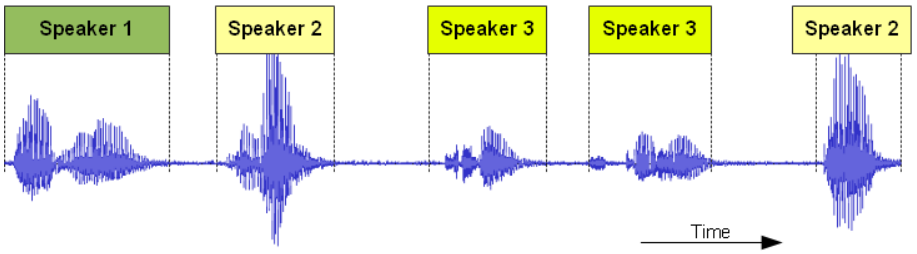

Figure 2.1: Speaker diarization system tries to find answer for *"Who spoke when?"*

2.1 Selected Diarization System

For purposes of this work I selected a local system developed in LIA. It is based on *ALIZE* toolkit. Basic information about this toolkit can be found there:

- official website: `mistral.univ-avignon.fr`
- wiki page: `mistral.univ-avignon.fr/wiki`

The toolkit is an open source code written in C++ and can be downloaded from its website: `mistral.univ-avignon.fr/en/download.html`.

To start with experiments, we need both Alize Library and Mistral RAL package. The subversion repository is also available and is there:

- `svn://mistral.univ-avignon.fr/svn/ALIZE/branches/ALIZE`
- `svn://mistral.univ-avignon.fr/svn/LIA_RAL/branches/LIA_RAL`

LIA_SpkSeg is the dedicated tool for speaker diarization. At the moment, the *LIA_SpkSeg* is not yet included in the stable Mistral RAL package because of its state of development. Contact Corinne Fredouille (`corinne.fredouille@univ-avignon.fr`) to get this non-published part of Mistral RAL package. Then, it can be found in *PACK_LIA_RAL/LIA_Seg/* directory.

Preparations: To run the diarization, download the ALIZE library and Mistral RAL package (as described above 2.1). My works took place in a Linux operating system (Ubuntu 9.04), Microsoft operating systems were not tested.

For compilation typical instruments (automake, autoconf, autogen, gcc 4, g++ 4, libsvm, libtool) are essential, a README file is also included in the package.

How to Run the Diarization: One of possible ways of execution of speaker diarization process:

```
LIA_SpkSeg -config cfg/config_STEP.cfg -processType STEP
-listFileToSegment lst/list
```

where one can specify the configuration and select the subprocess of the diarization (which can be also specified in the configuration file; the command line parameters overwrite configuration in the config file). The STEP can be one of these:

- `ReSegAcoustic` ... speech/silence segmentation
- `Seg` ... speaker segmentation
- `ReSeg` ... speaker re-segmentation, realignment

The steps are described in the section 2.2. Each step has its own configuration. In such a configuration types of mixtures, file extensions, file formats, paths, processing constraints *etc*are defined.

2.2 System Description Step-by-Step

In the following paragraphs there is a detailed view of each step of the speaker diarization process. Most of the information come from [7] and form the source code. From an abstract point of view diarization system components are:

- Speech activity detection
- Speaker change detection
- Speaker clustering
- Viterbi algorithm

Figure 2.2 shows a data flow diagram of the Speaker Diarization System.

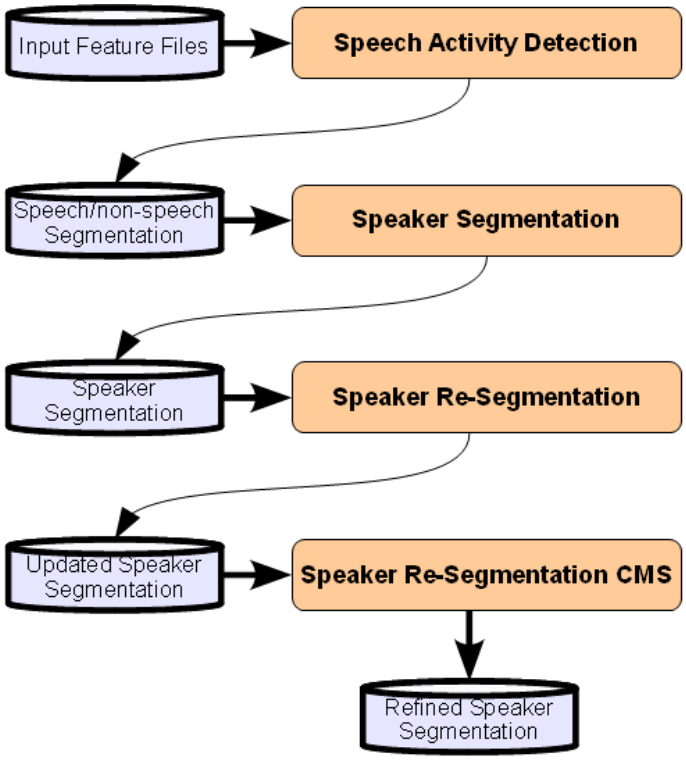

Figure 2.2: Scheme of the Speaker Diarization System

2.2.1 Input Audio Transformation

Input recordings should be transformed to *sph* file type (example: "sox file.wav file.sph"; SOX is a Sound eXchange program).

SPH [4], SPHERE file type (SPeech HEader Resources) is a file format defined by National Institute of Standards and Technology (NIST) and is used with speech audio.

2.2.2 Parametrization, Feature Extraction

- 20 ms frame length
- 10 ms frame rate
- 39 acoustic features used by acoustic re-segmentation
- 20 Mel frequency cepstral coefficients (MFCC) augmented by the normalized log-energy used by segmentation and re-segmentation step
- 34 coefficients used by re-segmentation CMS
- no coefficient normalization is applied
- 128 Gaussians

More about parametrization is in [3] and [7].

sfbcep, a tool from an open source speech signal processing SPRO toolkit (used version is 4.0), is used for extracting acoustic features. It creates a filterbank derived cepstral features from an input waveform. Description of DCT with other info is placed on project's web page [2].

2.2.3 Speech Activity Detection

In the system this process is called *resegAcoustic*. It is a real acoustic *resegmentation*. Therefore, system supposes initial label files and feature files from audio signal as an input. Example of such a label file is here (filename.lbl; format – in seconds: start_time end_time label): "`0 1279 speech`".

The system uses two state HMM which are representing speech and nonspeech. In the output there is a label file containing signal divided into "speech" and "nonspeech" segments.

Approximate speed of this part is at about 1/40 real time (tested on LIA server, this estimation is very inexact).

2.2.4 Speaker Segmentation

This part of the diarization process tries to find all possible speaker turns. This is done only on the speech segments detected in previous step.

There is used EM training algorithm and Viterbi decoding on an *evolutive Hidden Markov Model* (E-HMM, [6]), where each state represents a speaker.

The process: First speaker, "L_0", represents whole speech – HMM with only one state. Then the system searches for the best segment which extracts from speaker 0. Now, the system searches for all segments of the new speaker (training a model and Viterbi algorithm; a new HMM state is added and new transition probabilities are computed ... this represents the evolution in E-HMM).

This process continues until there is no more speech left for another new speaker or if there is no gain in terms of likelihood.

Approximate speed of this part is at about 1/2 real time (tested on LIA server, this estimation is very inexact).

2.2.5 Speaker Re-Segmentation

This part represents an iterative process of examination and a realignment of results gained in previous step. This step includes a possibility of removing of relevant speakers.

A Maximum a Posteriori (MAP [8, p. 143]) adaptation is used. All the GMM models of speakers of an E-HMM depending are adapted by the current segmentation. Afterwards, Viterby decoding is used.

Approximate speed of this part is at about 1/10 real time (tested on LIA server, this estimation is very inexact).

2.2.6 Speaker Re-Segmentation CMS

This process behaves almost like the previous one with a few differences. It aims at refining the segmentation obtained in the previous step (speaker re-segmentation 2.2.5).

Speech/non-speech segmentation is updated by the output of previous step. Then the features are normalized (variance and mean normalization) and configuration is set to use these normalized features and Cepstral Mean Subtraction (CMS, a compensation technique for convolutive distortions).

Approximate speed of this part is at about 1/3 real time (tested on LIA server, this estimation is very inexact).

2.3 Scores

This section contains a description of how the scores are computed. Information about evaluation data and also the scores of *LIA_SpkSeg* follows. Tested software was built in 2004 and then modified in 2006, 2007.

2.3.1 Diarization Error Rate

A NIST tool *"md-eval-v21.pl"* was used for computation of the error rate produced by the diarization system, it is available here:
`www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl`.
It is an important script which goes frame by frame and counts all missed speech (MISS; speech is labeled as a non-speech), false alarm speech (FA; opposite of MISS) and speaker errors (SPKER; speaker A is labeled as another speaker). Afterwards it prints all the error rates also with an overall diarization error rate (simply DER = MISS + FA + SPKER) [1, p. 142].

2.3.2 Evaluation Set

Tests were made on *NIST evaluation campaign RT 2009 MDM* set (16 kHz, mono, 16 bit linear, Little Endian). List of selected files follows with used duration of each file. There are not used whole recordings but only parts of them, strictly limited by Unpartitioned Evaluation Map (UEM).

- EDI_20071128-1000 (29:24 seconds)
- EDI_20071128-1500 (38:48 seconds)
- IDI_20090128-1600 (30:06 seconds)
- IDI_20090129-1000 (30:05 seconds)
- NIST_20080201-1405 (20:20 seconds)
- NIST_20080227-1501 (18:55 seconds)
- NIST_20080307-0955 (21:19 seconds)

Statistics of Evaluation data

Table 2.1 shows number of segments with an average duration of segment and number of speakers for each file in evaluation set. These numbers are based on output of re-segmentation CMS process of LIA speaker diarization system (it is a final system output, not real official data).

Table 2.2 is below to compare the system final output with official reference. The numbers show that the reference data contain much more shorter segments In general the reference data also contain more speakers.

	Number of segments	**Average duration**	**Number of speakers**
EDI_20071128-1000	360	4.09s	4
EDI_20071128-1500	478	2.95s	4
IDI_20090128-1600	272	6.28s	4
IDI_20090129-1000	457	3.23s	5
NIST_20080201-1405	187	6.13s	5
NIST_20080227-1501	155	7.04s	6
NIST_20080307-0955	139	8.80s	7

Table 2.1: Evaluation set – diarization system output: number of segments with average duration of segment per file, number of speakers

Figures 2.3 and 2.4 show relative representation of number of segment duration for output of speaker diarization system and reference data.

2.3.3 Results

In table 2.3 there are shown scores of the LIA speaker diarization system. The LIA speaker diarization system includes three main processing steps. They

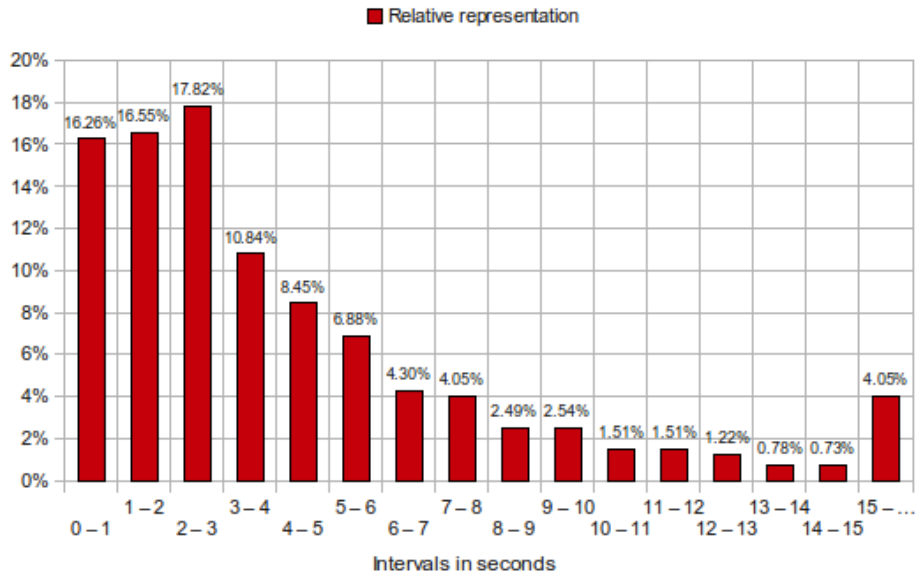

Figure 2.3: Evaluation set – diarization system output: relative representation of number of segments in duration intervals (number of segments with duration from zero to one second, number of segments with duration from one second to two seconds, *etc*)

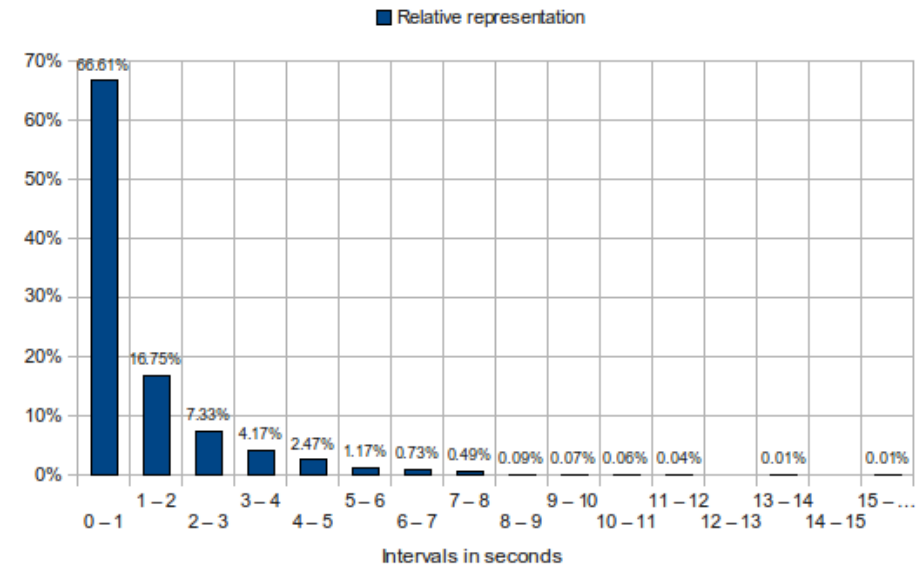

Figure 2.4: Evaluation set – official reference data: relative representation of number of segments in duration intervals (number of segments with duration from zero to one second, number of segments with duration from one second to two seconds, *etc*)

	Number of segments	Average duration	Number of speakers
EDI_20071128-1000	1183	1.15s	5
EDI_20071128-1500	1422	0.89s	5
IDI_20090128-1600	1124	1.45s	5
IDI_20090129-1000	1067	1.28s	5
NIST_20080201-1405	1535	0.73s	6
NIST_20080227-1501	968	1.08s	7
NIST_20080307-0955	833	1.38s	12

Table 2.2: Evaluation set – official reference data: number of segments with average duration of segment per file, number of speakers

are segmentation, re-segmentation, re-segmentation using CMS In the first column there are names of tested files, in the second column (Seg) there are diarization error rates (DER). In the third column (ReSeg) there are DER of re-segmentation process, and in the fourth column (ReSegCMS) there are DER of re-segmentation process of segmentation process of LIA speaker diarization system using CMS.

As written below, two recordings (EDI_20071128-1500 and NIST_20080201-1405) have higher DER than the others. I listened to parts of all the recordings and the main problem of these tow is lots of crosstalk (overlap) which causes higher diarization error rate.

	Diarization Error Rate (%)		
	Seg	ReSeg	ReSegCMS
EDI_20071128-1000	09.39	03.44	03.21
EDI_20071128-1500	46.05	33.38	33.82
IDI_20090128-1600	33.98	15.66	14.95
IDI_20090129-1000	15.85	14.38	14.08
NIST_20080201-1405	55.69	55.73	47.93
NIST_20080227-1501	25.24	20.34	20.41
NIST_20080307-0955	23.42	19.29	18.67
Overall	**27.63**	**19.80**	**18.93**

Table 2.3: System output: Diarization Error Rates of files in NIST evaluation campaign RT 2009 MDM set (October 2009)

Speed of the system: Approximate speed of this speaker diarization system (without initial feature extraction) is at about (1/40 + 1/2 + 1/10 + 1/3) 1/1 real time (tested on LIA server, this estimation is very inexact).

2.4 Summary

After description of speaker diarization system, its parts and scores, the next topic is about factor analysis.

Chapter 3

Factor Analysis

In this chapter I try to explain basis of factor analysis. This topic is rather recently important in speaker verification [5] and language identification domain because of the big potential of improvement, possible decrease of EER (Equal Error Rate).

The mentioned potential for speaker diarization is in revealing the channel variability of audio recordings to reinforce the comparison of two segments (if the segments are from the same speaker or not). The goal is in localization a subspace containing the channel variability.

3.1 Original Modeling a Speaker

Original approach – a standard Gaussian Mixture Models (GMM) creates a speaker model from UBM (Universal Background Model) by MAP (Maximum a Posteriori [8, p. 143]) adaptation technique. Such a modeling can be mathematically described:

$$m_{(h,s)} = m + Dy_s$$

where

- $m_{(h,s)}$ is speaker session dependent supervector mean
- D is a diagonal matrix $MFxMF$
- F is a dimension of a feature space
- M is a number of Gaussians in the GMM
- y_s is a speaker vector

As we can see, there is nothing which represents the variability of channel. So, such a model is session-dependent. The next section is about modeling the session variability.

3.2 Modeling a Session Variability

A session variability in speaker verification means *variability in sessions* (it means in different recordings). What changes is for example position of the microphone and other acoustic characteristics (the environment may differ).

But, in this domain – speaker diarization, the situation is a little bit different. There is usually one long file containing several speakers. During time the position of the microphone and other acoustic characteristics may differ. And this is the variability, the *variability in channel*, we will try to model to improve the overall diarization error rate.

Example of a situation where one speaker is moving (the distance from speaker to microphone is changing) is shown in figure 3.1. Diarization system which does not contain a factor analysis process may encounter difficulties recognizing the first speech segment (by the figure: "Hello!") and the third speech segment (by the figure: "How are you?") as utterances of the same speaker (by the figure: "Speaker A"). When using factor analysis in diarization system we can model speakers without such a variability (in this case the variability is caused by moving of the speaker A).

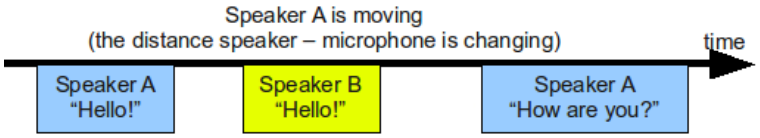

Figure 3.1: Two speakers are talking, one of them is moving, this is a typical case where application of factor analysis can help

Considering session variability a speaker model consists of three parts (by [5]):

- a speaker-session independent component
- a speaker dependent component
- a session dependent component

Considering session variability the new speaker modeling can be mathematical described:

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

where

- D is a dimension of a feature space $MDxMD$
- M is a number of Gaussians in the GMM
- (h, s) is a session h of a speaker s
- $m_{(h,s)}$ is a speaker session dependent supervector mean
- y_s is a speaker vector
- U is a session variability matrix
- $x_{(h,s)}$ are the channel factors

3.3 A Tool for Modeling the Variability

In LIA_RAL/LIA_SpkTools there is a module called *FactorAnalysis*. This module is responsible for computing Factor Analysis statistics and is also capable of estimating (among others):

- speaker model ... $m + Dy + Ux$
- true speaker model ... $m + Dy$
- session model ... $m + Ux$

This module can also compute log likelihoods of the Factor Analysis model.

EigenChannel program located in LIA_RAL/LIA_SpkDet uses this *FactorAnalysis* module for modeling the session variability (the U matrix).

3.4 Application of Factor Analysis to Speaker Diarization

This section is about joint of the factor analysis and speaker diarization system. The basic idea of application of factor analysis is to make a speech processing system independent on channel variability (such as a change of distance of microphone or environment where the recording takes place). Approaches have been studied as reported in the next sections.

3.4.1 New Module Using Factor Analysis in LIA_SpkSeg

I have made a new module in LIA Speaker Diarization System. It is a re-segmentation process, *ReSegFA*, using Factor Analysis. I use there output from diarization system (modified to index structure: speaker per line, cluster per label file, symbolic link to feature file) as an input for factor analysis. It works like re-segmentation of LIA speaker diarization system (2.2.5) but uses speaker models modeled by Factor Analysis. It works in the following cycle:

1. Indexing speakers
2. Estimation of factor analysis statistics of speakers
3. Speakers model acquisition
4. Viterbi decoding

Re-segmentation step of speaker diarization system using factor analysis (indexing, estimation of statistics, modeling, Viterbi decoding) is repeated until stop criterion (small difference between last and last-but-one segmentation, can be set in configuration). The cycle is also shown in the figure 3.2.

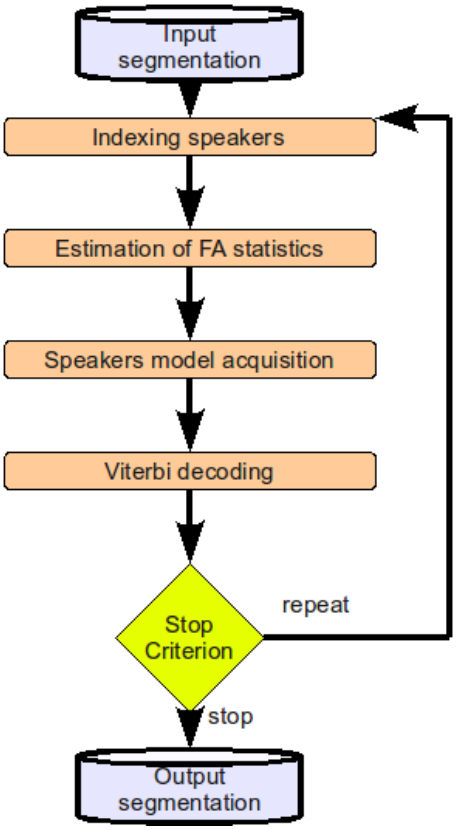

Figure 3.2: Scheme of the re-segmentation step using Factor Analysis

3.5 Protocol of the First Set of Experiments

The idea applied in this set of experiments is in modeling the channel variability between the speakers (for illustration see figure 3.4). This idea is very controversial and the following experiments shows if such a modeling the channel variability can suit to our combination of development and evaluation data. The modeling may cause reduction of differences between the speakers thus may also cause possible increase of speaker error rate (part of DER described here: 2.3.1).

Re-segmentation process using factor analysis is used as a last step of LIA speaker diarization system (after re-segmentation CMS). It means that an output of re-segmentation using CMS is used as an input for re-segmentation using factor analysis.

This section is divided into two parts. The first contains a description of used database, toolkits and system settings. The second part includes results with conclusions.

3.5.1 Database, toolkits, settings

The experiments tested in this section are based on evaluation and development set of files described below in separated sub-sections.

Evaluation Set

There is used *NIST evaluation campaign 2009 (RT)* as an evaluation set (models with 21 coefficients, 128 Gaussians). List of files is already presented in chapter about speaker diarization (2.3.2).

If we look at the scores of speaker diarization system presented here: 2.3, we can see quite good average of DERs. But in two cases we have DER too high (it is the case of EDI_20071128-1500 and NIST_20080201-1405 – mainly due to crosstalk). It is difficult to evaluate tests on these two files containing such a big initial error.

But one way is possible. We can select recordings which have better DER than the average (IDI_20090128-1600 and IDI_20090129-1000). Than we can make tests on the system with Factor Analysis as a part of the LIA speaker diarization system.

The other way is to run tests on all the files in evaluation set and observe DER of all the files (including files with the worst DER).

Development Set

NIST campaign RT 2008 MDM is used there as a development set.

List of files follows, augmented with used duration of each file. There are not used whole recordings but only parts of them, strictly limited by Unpartitioned Evaluation Map (UEM):

- AMI_20041210-1052 (12:10 seconds)
- AMI_20050204-1206 (11:54 seconds)
- CMU_20050228-1615 (12:01 seconds)

- CMU_20050301-1415 (11:58 seconds)
- CMU_20050912-0900 (17:51 seconds)
- CMU_20050914-0900 (17:58 seconds)
- EDI_20050216-1051 (18:00 seconds)
- EDI_20050218-0900 (18:10 seconds)
- ICSI_20000807-1000 (11:22 seconds)
- ICSI_20010208-1430 (9:59 seconds)
- ICSI_20010531-1030 (12:11 seconds)
- ICSI_20011113-1100 (11:59 seconds)
- LDC_20011116-1400 (10:01 seconds)
- LDC_20011116-1500 (10:01 seconds)
- NIST_20030623-1409 (11:13 seconds)
- NIST_20030925-1517 (11:02 seconds)
- NIST_20050427-0939 (11:55 seconds)
- NIST_20051024-0930 (10:15 seconds)
- NIST_20051102-1323 (18:06 seconds)
- VT_20050304-1300 (11:58 seconds)
- VT_20050318-1430 (12:04 seconds)
- VT_20050623-1400 (18:02 seconds)
- VT_20051027-1400 (9:38 seconds)

Figure 3.3 show relative representation of segment duration of reference data.

Training a World Model

A world model is trained on *LIA RT 2008 MDM* development set, which is described above (3.5.1).

Details of the world model:

- 34 coefficients
- 128 Gaussians
- 26 training iterations

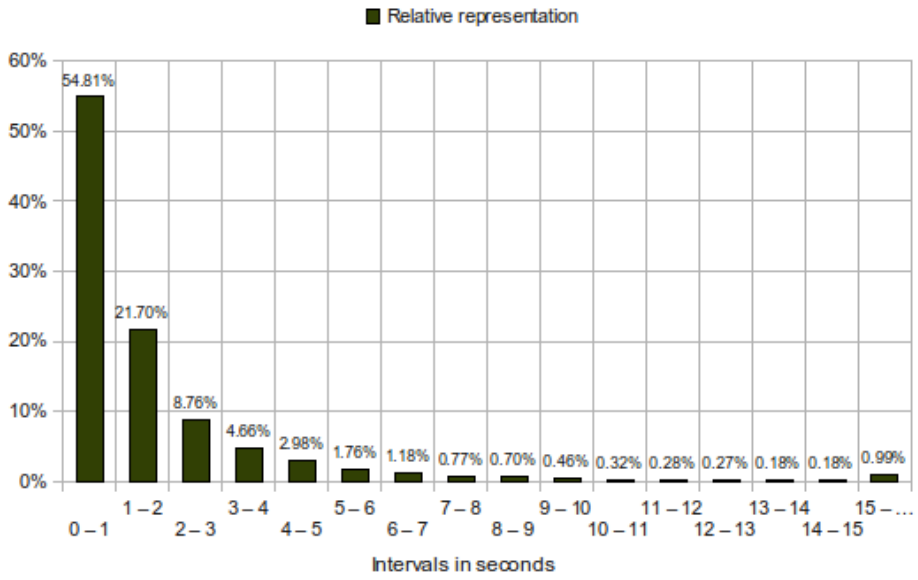

Figure 3.3: Development set – official reference data: relative representation of number of segments in duration intervals (number of segments with duration from zero to one second, number of segments with duration from one second to two seconds, *etc*)

Estimation of the Channel Variability Matrix

To estimate a channel matrix there is a need to prepare an index structure for this purpose. This structure will serve as an input for *EigenChannel* program (LIA_RAL/LIA_SpkDet/EigenChannel) using the world model from previous step (3.5.1).

Figure 3.4 illustrates this simple index structure of speaker cluster per column.

Details of the channel variability matrix:

- 34 coefficients
- 128 Gaussians
- 100 channel matrix rank
- 6 training iterations

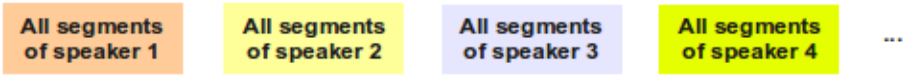

Figure 3.4: Index structure – the first approach: speaker cluster per column, all speakers on one line

3.5.2 Test 1.1 – various number of coefficients

This test is aimed at comparing between factor analysis working with feature files containing 21 coefficients and normalized (variance and mean normalization, like in re-segmentation part using CMS in speaker diarization system) feature files containing 34 coefficients.

Hypothesis: feature files containing more coefficients will be more suitable and the diarization error rate will be lower.

Specifications

- Speaker modeling: $m + Dy + Ux$
- Number of FA training iterations: *1*
- Number of Gaussians: *128*
- Number of coefficients: *34*, and then *21*
- Channel matrix rank: *100*

Results

Results of this experiment are presented in tables 3.1, 3.2 and 3.3.

	Diarization Error Rate (%)		
	Original	**FA, 34 coeff**	**FA, 21 coeff**
EDI_20071128-1000	03.21	*03.16*	*05.49*
EDI_20071128-1500	33.82	*34.63*	*37.66*
IDI_20090128-1600	14.95	*14.91*	*20.01*
IDI_20090129-1000	14.08	*15.71*	*17.85*
NIST_20080201-1405	47.93	*49.40*	*59.92*
NIST_20080227-1501	20.41	*16.62*	*20.46*
NIST_20080307-0955	18.67	*19.34*	*22.42*
Overall	**18.93**	***19.09***	***22.97***

Table 3.1: Test 1.1 – Overall DER with FA using 34 and 21 coefficients

Conclusion

The results presented in table 3.1 show scores of baseline system without factor analysis (column "Original") of system using factor analysis with feature files containing 21 coefficients (column "FA, 21 coeff") and normalized (variance and mean normalization) feature files containing 34 coefficients (column "FA, 34 coeff"). Using normalized feature files containing 34 coefficients is better (overall DER is 3.88% lower than with 21 coefficients). This confirms our hypothesis.

Re-segmentation step of speaker diarization system using factor analysis (indexing, estimation of statistics, modeling, Viterbi decoding) is repeated until stop criterion (small difference between last and last-but-one segmentation, can be set in configuration). In tables 3.2 and 3.3 there are scores of particular

	DER (%) of particular iterations	**Gain** (%)
EDI_20071128-1000	03.10; 03.16	*-0.06*
EDI_20071128-1500	32.99; 33.94; 34.63	*-1.64*
IDI_20090128-1600	14.87; 14.79; 14.91	*-0.04*
IDI_20090129-1000	15.50; 15.69; 15.71	*-0.21*
NIST_20080201-1405	47.46; 48.02; 49.17; 49.08; 49.40	*-1.94*
NIST_20080227-1501	19.57; 19.25; 18.95; 18.60; 17.72; 17.01; 16.62	*+2.95*
NIST_20080307-0955	18.65; 19.25; 19.34; 19.34	*-0.69*
Simple average gain		**-0.23**

Table 3.2: Test 1.1 – Detailed view of DER of each iteration (indexing, FA statistics, modeling, Viterbi; scores of each iteration delimited by semicolon) using 34 coefficients

	DER (%) of particular iterations	**Gain** (%)
EDI_20071128-1000	4.84; 5.21; 5.49;	*-0.65*
EDI_20071128-1500	33.85; 36.40; 37.66;	*-3.81*
IDI_20090128-1600	17.77; 19.50; 20.01;	*-2.24*
IDI_20090129-1000	17.47; 17.85;	*-0.38*
NIST_20080201-1405	53.27; 56.82; 57.32; 59.92;	*-6.65*
NIST_20080227-1501	25.68; 22.76; 21.25; 20.46;	*+5.22*
NIST_20080307-0955	21.33; 22.10; 22.42;	*-1.09*
Simple average gain		**-1.37**

Table 3.3: Test 1.1 – Detailed view of DER of each iteration (indexing, FA statistics, modeling, Viterbi; scores of each iteration delimited by semicolon) using 21 coefficients

iterations of factor analysis. As we can see, the results are in almost all cases coming worse and worse. The stop criterion of re-segmentation is not suitable.

3.5.3 Test 1.2 – various speaker modeling

This test is aimed at comparing between factor analysis modeling speaker containing all the variability $m + Dy + Ux$ and factor analysis modeling speaker without the channel variability $m + Dy$. The results will show if there is any information in the channel matrix.

Hypothesis: modeling speaker without disturbing channel variability should normally be better. But in this set of experiments it might not be truth. The reason is in channel variability which is estimated between speaker clusters.

Specifications

- Various speaker modeling: $m + Dy + Ux$ in comparison with $m + Dy$
- Number of FA training iterations: *1*
- Number of Gaussians: *128*

- Number of coefficients: *34*
- Channel matrix rank: *100*

Results

Results of this experiment are presented in tables 3.4 and 3.5.

	Diarization Error Rate (%)		
	Original	**FA,** $m + Dy + Ux$	**FA,** $m + Dy$
EDI_20071128-1000	03.21	03.16	39.54
EDI_20071128-1500	33.82	34.63	37.02
IDI_20090128-1600	14.95	14.91	33.93
IDI_20090129-1000	14.08	15.71	16.54
NIST_20080201-1405	47.93	49.40	46.10
NIST_20080227-1501	20.41	16.62	19.48
NIST_20080307-0955	18.67	19.34	18.67
Overall	**18.93**	**19.09**	**29.56**

Table 3.4: Test 1.2 – Overall DER with FA using different speaker modeling

	DER (%) of particular iterations	**Gain** (%)
EDI_20071128-1000	3.48; 4.68; 7.28; 17.04; 29.05; 34.75; 39.54	*-36.06*
EDI_20071128-1500	36.20; 37.28; 36.61; 37.02	*+0.26*
IDI_20090128-1600	15.77; 17.46; 18.02; 19.88; 25.07; 31.14; 34.83; 34.88; 34.31; 34.01; 33.97; 33.93	*-16.47*
IDI_20090129-1000	16.07; 16.54	*-0.47*
NIST_20080201-1405	46.38; 46.48; 46.10; 46.10	*+0.38*
NIST_20080227-1501	19.64; 19.48; 19.48	*+0.00*
NIST_20080307-0955	18.67; 18.67; 18.67	*+0.00*
Simple average gain		**-7.48**

Table 3.5: Test 1.2 – Detailed view of DER of each iteration (indexing, FA statistics, modeling, Viterbi; scores of each iteration delimited by semicolon) using $m + Dy$ models

Conclusion

The results presented in table 3.4 show scores of baseline system without factor analysis (column "Original") of system using factor analysis modeling speaker containing all the variability $m+Dy+Ux$ (column "FA, $m+Dy+Ux$") and factor analysis modeling speaker without the channel variability $m+Dy$ (column "FA, $m + Dy$"). There is a big difference between these two overall averages. This means that the channel matrix contains some needful information. By removing channel variability with this information it influences the results negatively.

Re-segmentation step of speaker diarization system using factor analysis (indexing, estimation of statistics, modeling, Viterbi decoding) is repeated until

stop criterion (small difference between last and last-but-one segmentation, can be set in configuration). In table 3.5 there are scores of particular iterations of factor analysis. As we can see, the results are in almost all cases coming worse and worse. The stop criterion of re-segmentation is not suitable, like in the previous experiment (3.5.2).

3.5.4 Test 1.3 – various number of Gaussians

This test is aimed at comparing between factor analysis working with feature files containing 128 and 256 Gaussians. It also uses different modeling – session modeling. This test will also show, if this modeling can be more useful with data used in these experiments.

Hypothesis: feature files containing more Gaussians will be more effective and the diarization error rate will be lower.

Specifications

- Speaker modeling: $m + Ux$
- Number of FA training iterations: *1*
- Number of Gaussians: *128* and then *256*
- Number of coefficients: *34*
- Channel matrix rank: *100*

Results

Results of this experiment are presented in tables 3.6, 3.7 and 3.8.

	Diarization Error Rate (%)		
	Original	**FA, 128 Gauss.**	**FA, 256 Gauss.**
EDI_20071128-1000	03.21	*04.45*	*04.82*
EDI_20071128-1500	33.82	*37.16*	*39.95*
IDI_20090128-1600	14.95	*20.13*	*19.73*
IDI_20090129-1000	14.08	*20.09*	*19.95*
NIST_20080201-1405	47.93	*55.67*	*54.99*
NIST_20080227-1501	20.41	*09.50*	*09.03*
NIST_20080307-0955	18.67	*23.36*	*23.32*
Overall	**18.93**	***21.78***	***21.98***

Table 3.6: Test 1.3 – Overall DER with FA using different number of Gaussians

Conclusion

The results presented in table 3.6 show scores of baseline system without factor analysis (column "Original") and system using factor analysis with feature files containing 128 Gaussians (column "FA, 128 Gauss.") and 256 Gaussians

	DER (%) of particular iterations	**Gain** (%)
EDI_20071128-1000	3.61; 4.45	*-0.84*
EDI_20071128-1500	34.50; 36.00; 37.16	*-2.66*
IDI_20090128-1600	18.29; 19.62; 20.13	*-1.84*
IDI_20090129-1000	19.75; 20.07; 20.09	*-0.34*
NIST_20080201-1405	51.82; 53.94; 55.67	*-3.85*
NIST_20080227-1501	21.50; 19.72; 14.79; 10.89; 9.57; 9.50	*+12.00*
NIST_20080307-0955	22.33; 23.19; 23.36	*-1.03*
Simple average gain		**0.33**

Table 3.7: Test 1.3 – Detailed view of DER of each iteration (indexing, FA statistics, modeling, Viterbi; scores of each iteration delimited by semicolon) using 128 Gaussians

	DER (%) of particular iterations	**Gain** (%)
EDI_20071128-1000	3.99; 4.82	*-0.83*
EDI_20071128-1500	34.16; 36.94; 39.95;	*-5.79*
IDI_20090128-1600	17.70; 19.25; 19.73;	*-2.03*
IDI_20090129-1000	19.80; 19.95;	*-0.15*
NIST_20080201-1405	50.86; 52.61; 54.99;	*-4.13*
NIST_20080227-1501	22.72; 21.91; 18.97; 15.81; 10.63; 10.06; 9.34; 9.03;	*+13.69*
NIST_20080307-0955	22.20; 23.03; 23.32;	*-1.12*
Average gain		**-0.05**

Table 3.8: Test 1.3 – Detailed view of DER of each iteration (indexing, FA statistics, modeling, Viterbi; scores of each iteration delimited by semicolon) using 256 Gaussians

(column "FA, 256 Gauss."). Hypothesis is not confirmed. The scores are similar even the overall average DER is 0.20% worse for factor analysis working with feature files containing 256 Gaussians.

Re-segmentation step of speaker diarization system using factor analysis (indexing, estimation of statistics, modeling, Viterbi decoding) is repeated until stop criterion (small difference between last and last-but-one segmentation, can be set in configuration). In tables 3.7 and 3.8 there are scores of particular iterations of factor analysis.

3.5.5 Conclusion of First Set of Experiments

- In comparison with original DER (without FA, written in section 2.3), the errors are generally worse. The best scores (but not better than the original DER) are in test 1.1 (3.5.2). In this experiment $m + Dy$ function, channel matrix rank 100, 128 Gaussians, 34 coefficients were used. These results show that modeling of the variability used in this set of experiments does not suit to our combination of development and evaluation data.

- As we can see, the results in each iteration are in almost all cases coming

worse and worse. The stop criterion of re-segmentation is not suitable. The iterating works fine almost only for file NIST_20080227-1501.

3.6 Protocol of the Second Set of Experiments

The idea applied in this set of experiments is in modeling the channel variability between speaker segments (for illustration see figure 3.5). This idea is much more convenient than the first one (3.5). The following experiments show if such a modeling of the channel variability can be more suitable to our combination of development and evaluation data.

Re-segmentation process using factor analysis is used as a last step of LIA speaker diarization system (after re-segmentation CMS). It means that an output of re-segmentation using CMS is used as an input for re-segmentation using factor analysis.

This section is divided into two parts. The first contains a description of used database, toolkits and system settings. The second part includes results with conclusions.

3.6.1 Database, toolkits, settings

The experiments tested in this section are based on evaluation and development set of files described below in separated sub-sections.

Evaluation Set

The evaluation set is the same as in the first set of experiments (3.5.1).

Development Set

The development set is the same as in the first set of experiments (3.5.1).

Training a World Model

The world model is the same as in the first set of experiments (3.5.1).

Estimation of the Channel Variability Matrix

To estimate a channel matrix there is a need to prepare an index structure for this purpose. This structure will serve as an input for *EigenChannel* program (LIA_RAL/LIA_SpkDet/EigenChannel) using the world model from previous step.

There is used another indexation to estimate the channel variability differently. There is a speaker per line structure (lines contain separated segments, rows contain speakers).

Figure 3.5 illustrates this simple structure of speaker per line, segment per column.

Details of the channel variability matrix:

- 34 coefficients
- 128 Gaussians
- various channel matrix rank in the following experiments (5, 10, 20, 40, 60, 80, 100, 120)
- 20 training iterations

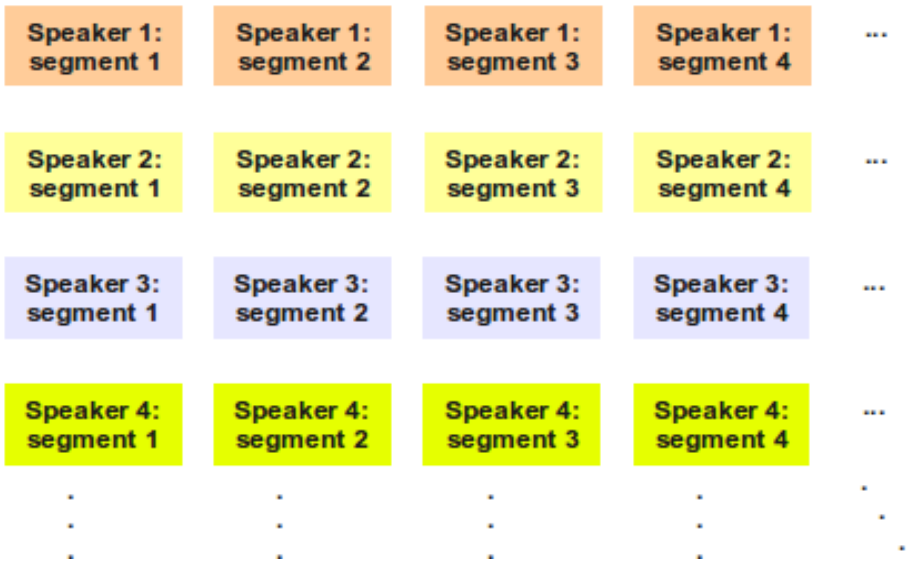

Figure 3.5: Index structure – the second approach: speaker per line, segment per column

Two constraints are applied there:

- The minimal segment duration used in U matrix estimation is set to *1 second* (filtration of very short segments)
- The minimal number of segments on one line is set to *2*

3.6.2 Test 2.1 – various speaker modeling

This test is aimed at comparing between factor analysis modeling speaker containing all the variability $m + Dy + Ux$ and factor analysis modeling speaker without the channel variability $m + Dy$. The results will show if there is any information in the channel matrix.

Hypothesis: modeling speaker without disturbing channel variability should normally be better. The difference is in another channel matrix used in the factor analysis process. Speaker modeling without channel variability should be more suitable in this test than in the previous similar test 1.2 (3.5.3). The channel variability which is estimated between speakers segments (speakers are separated) might serve a better job than channel variability estimated between speakers clusters (test 1.2 3.5.3).

Specifications

- Various speaker modeling: $m + Dy + Ux$ in comparison with $m + Dy$
- Number of FA training iterations: *1*
- Number of Gaussians: *128*

- Number of coefficients: *34*
- Channel matrix rank: *10*

Results

Results of this experiment are presented in table 3.9.

	Diarization Error Rate (%)		
	Original	**FA,** $m + Dy + Ux$	**FA,** $m + Dy$
EDI_20071128-1000	03.21	*03.11*	*03.18*
EDI_20071128-1500	33.82	*34.27*	*46.95*
IDI_20090128-1600	14.95	*14.58*	*12.34*
IDI_20090129-1000	14.08	*15.61*	*13.97*
NIST_20080201-1405	47.93	*49.40*	*42.00*
NIST_20080227-1501	20.41	*13.02*	*18.97*
NIST_20080307-0955	18.67	*19.13*	*18.12*
Overall	**18.93**	***18.54***	***19.32***

Table 3.9: Test 2.1 – Overall DER with FA using different speaker modeling

Conclusion

The results presented in table 3.9 show scores of baseline system without factor analysis (column "Original") of system using factor analysis modeling speaker containing all the variability $m+Dy+Ux$ (column "FA, $m+Dy+Ux$") and factor analysis modeling speaker without the channel variability $m+Dy$ (column "FA, $m + Dy$"). There is not a big difference between these two overall averages. The channel matrix contains some information, but the influence is not so big.

By removing channel variability with this information it influences the results negatively. That also means that hypothesis mentioned in this experiment (3.6.2) is not confirmed.

3.6.3 Test 2.2 – various number of training iterations of speaker model

This test is very data-dependent. This test is aimed at comparing between system using different numbers of factor analysis training iterations (1, 2, 3, 4 or 5).

Hypothesis: the more training iterations the better scores (and the more processing time needed).

Specifications

- Speaker modeling: $m + Dy + Ux$
- Various number of FA training iterations: 1, 2, 3, 4, 5
- Number of Gaussians: *128*

- Number of coefficients: *34*
- Channel matrix rank: *10*

Results

Results of this experiment are presented in table 3.10.

	DER per number of train iterations (%)					
	original	**1**	**2**	**3**	**4**	**5**
EDI_20071128-1000	03.21	03.11	03.11	03.11	03.11	03.11
EDI_20071128-1500	33.82	34.27	34.27	34.27	34.27	34.27
IDI_20090128-1600	14.95	14.58	14.58	14.58	14.58	14.58
IDI_20090129-1000	14.08	15.61	15.61	15.61	15.61	15.61
NIST_20080201-1405	47.93	49.40	49.40	49.17	49.17	49.24
NIST_20080227-1501	20.41	13.02	13.02	13.09	04.90	04.96
NIST_20080307-0955	18.67	19.13	19.13	19.13	19.13	19.13
Overall	**18.93**	**18.54**	**18.54**	**18.53**	**17.66**	**17.67**

Table 3.10: Test 2.2 – Overall DER with FA using different number of training iterations of speaker model

Conclusion

The results presented in table 3.10 show scores of baseline system without factor analysis (column "Original") and system using factor analysis with 1, 2, 3, 4 and 5 training iterations.

Hypothesis is confirmed. The scores are coming better with more training iterations. The best scores are produced by factor analysis using 4 training iterations.

3.6.4 Test 2.3 – various channel matrix rank

This test is also very data-dependent. This test is aimed at comparing between system using different numbers of channel matrix rank (5, 10, 20, 40, 60, 80, 100 or 120) and finding the best rank which suits to our data.

Hypothesis: the channel matrix rank might be similar to the number of speakers in development data (3.5.1, at about 110 speakers) used by estimation of channel matrix.

Specifications

- Speaker modeling: $m + Dy + Ux$
- Number of FA training iterations: *4*
- Number of Gaussians: *128*
- Number of coefficients: *34*
- Channel matrix rank: various (5, 10, 20, 40, 60, 80, 100, 120)

Results

Results of this experiment are presented in table 3.11.

	DER per channel matrix rank (%)							
	5	**10**	**20**	**40**	**60**	**80**	**100**	**120**
EDI_20071128-1000	03.06	03.11	03.08	03.21	03.21	03.21	03.21	03.21
EDI_20071128-1500	33.41	34.27	34.36	34.55	34.48	34.76	34.89	34.66
IDI_20090128-1600	14.58	14.58	14.82	14.92	14.92	15.03	14.99	15.05
IDI_20090129-1000	15.68	15.61	15.82	16.00	15.97	15.95	16.08	16.01
NIST_20080201-1405	48.44	49.17	49.14	49.57	49.82	49.30	49.36	49.46
NIST_20080227-1501	18.40	04.90	16.12	16.21	16.70	16.71	16.93	16.75
NIST_20080307-0955	18.93	19.13	19.24	19.27	19.25	19.14	19.10	19.14
Overall	**18.90**	**17.66**	**18.96**	**19.10**	**19.16**	**19.16**	**19.21**	**19.17**

Table 3.11: Test 2.3 – Overall DER with FA using different channel matrix rank

Conclusion

The results presented in table 3.11 show scores of system using factor analysis with various channel matrix (rank: 5, 10, 20, 40, 60, 80, 100 and 120).

Hypothesis is not confirmed. The best scores are produced by factor analysis using channel matrix rank set to 10.

3.6.5 Test 2.4 – channel matrix per file

This is another test which is also very data-dependent. This test uses channel matrix per file. It means that for each file is estimated special channel matrix which is based on the output segmentation of the last step of LIA speaker diarization system (re-segmentation CMS). We also have to know that the data used there for the estimation are not totally correct (the original scores are here: 2.3.1). In few files there is a big error – such a segmentation is used for estimation of the channel matrix and in these cases such an approach may not be helpful.

Hypothesis: the channel matrix is unique for each file. The results of this experiment might be better than the original results without factor analysis.

Specifications

- In this experiment there are used specific U matrices estimated per file by segmentation obtained by the last step of speaker diarization system (re-segmentation CMS)
- Speaker modeling: $m + Dy + Ux$
- Number of FA training iterations: *4*
- Number of Gaussians: *128*
- Number of coefficients: *34*
- Channel matrix rank: *10*

Results

Results of this experiment are presented in table 3.12.

	Diarization Error Rate (%)	
	Original	**FA**
EDI_20071128-1000	03.21	*03.08*
EDI_20071128-1500	33.82	*34.33*
IDI_20090128-1600	14.95	*14.76*
IDI_20090129-1000	14.08	*15.75*
NIST_20080201-1405	47.93	*48.87*
NIST_20080227-1501	20.41	*14.28*
NIST_20080307-0955	18.67	*19.17*
Overall	**18.93**	***18.70***

Table 3.12: Test 2.4 – Overall DER with FA using channel matrix per file

Conclusion

The results presented in table 3.12 show scores of baseline system without factor analysis (column "Original") and system using factor analysis (column "FA").

As reader can see, the difference is not very big. If we compare only the overall averages, the hypothesis is confirmed, but the scores produced by factor analysis are better only in 3 of 7 cases. Such a technique might be helpful only for files with very low DER.

3.6.6 Test 2.5 – various channel matrix constraints

This test is aimed at comparing between specific U matrices estimated by approach described above in this section (3.6.1) but using different constraints: the minimal segment duration used in U matrix estimation is set to 0 (no constraint, use all segments), 1, 5 and 10 seconds (filtration of shorter segments).

Hypothesis: removing very short (hundreds of milliseconds) segments might be useful. But by elimination of all segments shorter than 10 seconds we can lose a very big significant part of segmentation. Figure 3.3 shows how many segments of development set represent segments with duration shorter than one second. It is *54.81%*. This number represents half of all segments in development set. But, this number does not inform about total duration, of course. Segments shorter than five seconds represent *92.91%* of total number of segments what might be too much to simply eliminate and expect good results with channel matrix estimated on the rest of long segments.

Specifications

- In this experiment there are used specific U matrices estimated by approach described above in this section (3.6.1) but using different constraints: the minimal segment duration used in U matrix estimation is gradually set to *0*, *1*, *5* and *10* seconds

- Speaker modeling: $m + Dy + Ux$
- Number of FA training iterations: *4*
- Number of Gaussians: *128*
- Number of coefficients: *34*
- Channel matrix rank: *10*

Results

Results of this experiment are presented in table 3.13.

	DER per segment duration minimum (%)				
	0 s	**1 s**	**2 s**	**5 s**	**10 s**
EDI_20071128-1000	03.21	03.11	03.10	03.06	03.08
EDI_20071128-1500	34.52	34.27	34.25	34.20	33.95
IDI_20090128-1600	14.84	14.58	14.59	14.66	14.61
IDI_20090129-1000	15.83	15.61	15.68	15.56	15.84
NIST_20080201-1405	49.39	49.17	49.39	49.27	48.90
NIST_20080227-1501	16.66	04.90	15.02	18.60	19.21
NIST_20080307-0955	19.24	19.13	19.21	19.17	18.92
Overall	**19.08**	**17.66**	**18.77**	**19.12**	**19.13**

Table 3.13: Test 2.5 – Overall DER with FA using different channel matrices (U matrices differ in the minimal segment duration used in estimation process)

Conclusion

The results presented in table 3.13 show scores of system using factor analysis working with channel matrix estimated on all segments, segments longer then 1, 2, 5 and 10 seconds.

As reader can see, there is an interesting difference between the overall average of system using U matrix estimated on all segments and system using U matrix estimated on segments longer than 1 second. The hypothesis is confirmed. System using U matrix estimated on segments longer than 1 second is the best performing one. It is much more better than system using U matrix estimated on all segments and is also better than system using U matrix estimated on segments longer than 10 seconds.

3.7 Protocol of the Third Set of Experiments

The idea applied in this set of experiments is in modeling the channel variability between speaker segments (as in the previous set of experiments 3.6). In comparison with the previous set of experiments, the channel matrix is estimated differently – in this set of experiments there is applied something like "sub-clustering". Segments of a speaker are grouped to have a specified minimal total duration.

Factor analysis re-segmentation process is used as a last step of LIA speaker diarization system (after re-segmentation CMS). It means that an output of re-segmentation using CMS is used as an input for re-segmentation using factor analysis.

This section is divided into two parts. The first contains a description of used database, toolkits and system settings. The second part includes results with conclusions.

3.7.1 Database, toolkits, settings

The experiments tested in this section are based on evaluation and development set of files described below in separated sub-sections.

Evaluation Set

The evaluation set is the same as in the first set of experiments (3.5.1).

Development Set

The development set is the same as in the first set of experiments (3.5.1).

Training a World Model

The world model is the same as in the first set of experiments (3.5.1).

Estimation of the Channel Variability Matrix

To estimate a channel matrix there is a need to prepare an index structure for this purpose. This structure will serve as an input for *EigenChannel* program (LIA_RAL/LIA_SpkDet/EigenChannel) using the world model from previous step .

There is used another indexation to estimate the channel variability differently. There is a speaker per line and a group of segments per label file structure (lines contain split speaker cluster, rows separate speakers). There is also defined the minimal duration of a speech segment and the minimal total duration of a speaker sub-cluster.

Details of the channel variability matrix:

- 34 coefficients
- 128 Gaussians
- 10 channel matrix rank
- 20 training iterations

Three constraints are applied there:

- The minimal segment duration used in U matrix estimation is set to *1 second* (filtration of very short segments)
- The minimal duration of a sub-cluster is gradually set to *0*, *20*, *40* and *60* seconds
- The minimal number of sub-clusters on one line (a speaker cluster) is set to *2*

3.7.2 Test 3.1 – sub-clustering

This test is aimed at comparing between specific U matrices estimated by approach described above in this section (3.7.1) but using different constraints: the minimal sub-cluster duration used in U matrix estimation is set to 0 (no sub-clustering), 20, 40, and 60 seconds.

Hypothesis: such a sub-clustering might be useful when estimating channel matrix from a huge amount of segments. This approach might be useful in saving processing time, but this technique is not capable of making the scores significantly improved.

Specifications

- Factor Analysis re-segmentation: use as a last step of LIA speaker diarization system (after re-segmentation CMS)
- Speaker modeling: $m + Dy + Ux$
- Number of FA training iterations: *4*
- Number of Gaussians: *128*
- Number of coefficients: *34*
- U matrix estimation details:
  - Channel matrix rank: *10*
  - The minimal duration of a speech segment: *1 second*
  - The minimal duration of a speaker sub-cluster – gradually: *without* (0 seconds), *20*, *40* and *60* seconds
  - The minimal number of sub-clusters on one line: *2*

Results

Results of this experiment are presented in table 3.14.

	DER per sub-cluster duration minimum (%)			
	0 s	**20 s**	**40 s**	**60 s**
EDI_20071128-1000	03.11	03.11	03.10	03.10
EDI_20071128-1500	34.27	34.37	33.87	33.87
IDI_20090128-1600	14.58	14.58	14.64	14.58
IDI_20090129-1000	15.61	15.67	15.57	15.69
NIST_20080201-1405	49.17	49.07	49.45	48.90
NIST_20080227-1501	04.90	17.49	17.92	18.33
NIST_20080307-0955	19.13	19.19	19.03	19.05
Overall	**17.66**	**19.02**	**19.00**	**19.02**

Table 3.14: Test 3.1 – Overall DER with FA using a sub-clustering technique (U matrices differ in the minimal sub-cluster duration used in estimation process)

Conclusion

The results presented in table 3.14 show scores of system using factor analysis working with channel matrix estimated on grouped segments with minimal sub-cluster duration set to 0 (no sub-clustering), 20, 40, and 60 seconds.

The hypothesis is confirmed. The estimation of channel matrix is faster, but the overall average error rate is worse. The diarization system using U matrix estimated on segments without sub-clustering has the best results.

3.7.3 Test 3.2 – sub-clustering

This is a continuation of the precedent test. In this test U matrices are estimated using different constraints: the minimal sub-cluster duration used in U matrix estimation is set to 1, 2, and 5 seconds including all segments (without filtration of segments shorter than one second).

Hypothesis: if the shortest segments are "concatenated" with their neighbors (to have a specified minimum duration at least) such a sub-clustering might be beneficial.

Specifications

- Factor Analysis re-segmentation: use as a last step of LIA speaker diarization system (after re-segmentation CMS)
- Speaker modeling: $m + Dy + Ux$
- Number of FA training iterations: *4*
- Number of Gaussians: *128*
- Number of coefficients: *34*
- U matrix estimation details:
  - Channel matrix rank: *10*
  - The minimal duration of a speech segment: *0 second*

– The minimal duration of a speaker sub-cluster – gradually: *1*, *2* and *3* seconds

– The minimal number of sub-clusters on one line: *2*

Results

Results of this experiment are presented in table 3.15.

	DER per sub-cluster duration minimum (%)		
	1 s	**2 s**	**5 s**
EDI_20071128-1000	03.10	03.10	03.11
EDI_20071128-1500	34.27	34.28	34.29
IDI_20090128-1600	14.52	14.59	14.55
IDI_20090129-1000	15.62	15.58	15.56
NIST_20080201-1405	49.22	48.79	48.86
NIST_20080227-1501	15.16	16.39	17.23
NIST_20080307-0955	19.11	19.25	19.11
Overall	**18.74**	**18.86**	**18.93**

Table 3.15: Test 3.2 – Overall DER with FA using a sub-clustering technique (U matrices differ in the minimal sub-cluster duration used in estimation process)

Conclusion

The results presented in table 3.15 show scores of system using factor analysis working with channel matrix estimated on grouped (concatenated) segments with minimal sub-cluster duration set to 1, 2, and 5 seconds.

The hypothesis is confirmed. The diarization system using U matrix estimated on concatenated segments with minimal duration set to one second has the best results.

3.7.4 Test 3.3 – repeat re-segmentation CMS after FA

The segmentation after application of re-segmentation using factor analysis is a little bit changed. This is the motif of this experiment. Let us see if another step of re-segmentation using CMS can somehow improve segmentation obtained from re-segmentation using factor analysis.

Re-segmentation process using factor analysis is in previous experiments used as a last step of LIA speaker diarization system (after re-segmentation CMS). It means that an output of re-segmentation using CMS is used as an input for re-segmentation using factor analysis. In this another step is added. The last step of this system is not re-segmentation using factor analysis, but re-segmentation using CMS once more (steps of the speaker diarization system: segmentation, re-segmentation, re-segmentation using CMS, re-segmentation using factor analysis, re-segmentation using CMS).

Hypothesis: as data from re-segmentation process using factor analysis are improved the re-segmentation using CMS (without factor analysis) obtain better

segmentation for processing. We can expect, that another step of re-segmentation using CMS might improve the segmentation.

Specifications

- Factor Analysis process: not exactly at the end of LIA speaker diarization system, after FA step there is used re-segmentation CMS again (segmentation, re-segmentation, re-segmentation CMS, FA, re-segmentation CMS)
- Speaker modeling: $m + Dy + Ux$
- Number of FA training iterations: *4*
- Number of Gaussians: *128*
- Number of coefficients: *34*
- U matrix estimation details:
  - Channel matrix rank: *10*
  - The minimal duration of a speech segment: *1 second*
  - The minimal duration of a speaker sub-cluster: *0 seconds* (no sub-clustering)
  - The minimal number of segments per line: *2*

Results

Results of this experiment are presented in table 3.16.

	Diarization Error Rate (%)					
	Original		1 FA iter	+resegCMS	full FA	+resegCMS
EDI_20071128-1000	03.21		03.03	03.23	03.11	03.33
EDI_20071128-1500	33.82		33.00	33.79	34.27	34.82
IDI_20090128-1600	14.95		14.56	14.76	14.58	14.58
IDI_20090129-1000	14.08		15.42	14.89	15.61	15.06
NIST_20080201-1405	47.93		48.86	47.75	49.17	47.32
NIST_20080227-1501	20.41		19.59	19.77	04.90	03.46
NIST_20080307-0955	18.67		18.84	18.84	19.13	18.77
Overall	**18.93**		**18.94**	**18.96**	**17.66**	**17.32**

Table 3.16: Test 3.3 – Overall DER with FA as a last step (segmentation, re-segmentation, re-segmentation CMS, FA) also with another step of re-segmentation CMS in columns "+resegCMS"

Conclusion

The results presented in table 3.16 show scores of baseline system without factor analysis (column "Original"), scores of system using 1 iteration of factor analysis re-segmentation (column "1 FA iter") and then applying another step: re-segmentation CMS (column "+resegCMS"). The next column contains scores of system using factor analysis re-segmentation (not only one iteration, but until

stop criterion, as normally, column "full FA") and then, the next column contains scores of system after application of another step of re-segmentation CMS (column "+resegCMS").

If we compare only the overall DER averages, the hypothesis is confirmed. But the scores produced by application of another step of re-segmentation CMS ("full FA" → "+resegCMS") are better only in 4 of 7 cases.

Repetition of re-segmentation step of diarization system means (in this case) another improvement of segmentation for NIST_20080227-1501, gain *+1.44%*.

3.7.5 About the Progress of DER

Comparing the original scores and the scores of the experiment 3.3 (3.7.4) the overall average diarization error rate is lower, from original *18.93%* to *17.32%*.

The overall average is coming better, but mainly due to one file with enormous progress. The file is *NIST_20080227-1501*. Only by factor analysis the score is improved from *20.41%* to *4.90%* which means *15.51%* gain (34 coefficients, 128 Gaussians, speaker per line, $m+Dy+Ux$ with 4 training iterations, U estimated only by segments of development data with duration at least one second). The gain can be then extra improved by repetition of re-segmentation CMS to *3.46%* (see 3.7.4).

There is no such a big gain in the other evaluation files, even the scores in 4 of 7 cases are worse than the original error rates.

One possible explanation of such a great improvement by application of factor analysis:

- This recording contains a lot of women
- And LIA speaker diarization system uses UBM trained on more men data (amount of data men – women is not balanced)

Statistics of NIST_20080227-1501

As written above the progress of NIST_20080227-1501 is very interesting (from 20.41% to 4.90% without repetition of re-segmentation CMS). The following table 3.17 shows more details about this file ("After FA" means there the output of re-segmentation process using factor analysis with the best scores).

	# segments	Average duration	# speakers
System: Before FA	155	7.04s	6
System: After FA	219	4.98s	6
Reference	968	1.08s	7

Table 3.17: Statistics of NIST_20080227-1501

Relation between Likelihood and Gain

By data obtained by each iteration of factor analysis for each file of evaluation set, there is *no correspondence* between gain (progress in DER) and likelihood (total likelihood for each model to its cluster) nor Viterbi probability (total Viterbi probability for all clusters).

Otherwise such a correspondence might be useful for implementation of a better performing stop criterion. Better stop criterion is needful in factor analysis re-segmentation process because the actual stop criterion let the segmentations become worse (problem of a negative gain in iterations of factor analysis re-segmentation process).

3.8 Summary

After many ideas and experiments including their negative and positive results it is time to make a conclusion of all the work and see what can be done in the future.

Chapter 4

Conclusion

In this report the speaker diarization system with factor analysis technique were briefly described. And after description of basis of used techniques many ideas of application of factor analysis were mentioned with their particular specifications and results.

Experiments show that application of factor analysis to speaker diarization can be useful and experiments included in this report serve as an evidence. The improvement of segmentation (related to reduction of diarization error rate) made by application of factor analysis to the speaker diarization is positive but not very big. The highest gain (against the original diarization system without factor analysis) is 1.27% in average (from 18.93% to 17.66%).

The consequences of improvement of speaker diarization system can influence also possibilities of control of huge amounts of audio data. With better performing speaker diarization system we can be more successful in searching indexed audio databases.

It is also necessary to mention the fact that application of factor analysis is very data-dependent where one of the most important parts is the estimation of channel variability.

4.1 Furter Work

What about the future of factor analysis and speaker diarization? Further experiments will follow and the next work might be aimed at different approaches of estimation of the channel variability (for instance, by experimenting with index structures and segmentation constraints) and/or using more training data. There is also a posibility to model speakers with a probabilities (a probability that a segment belongs to a speaker). It would be also interesting to investigate NIST_20080227-1501, why it is so well performing. Future work can be (and is advised to be) based on experiments written in this report.

As anytime in the past, every question is trying to find its answer and every problem is trying to find its solution. It will not be different with utilization of factor analysis in speaker diarization.

4.2 Acknowledgements

My thanks belongs to all the people in LIA, especially to Corinne Fredouille and Driss Matrouf for their advices and help.

List of Tables

Bibliography

[1] X. Anguera. *Robust Speaker Diarization for Meetings.* PhD thesis, Universitat Politècnica de Catalunya, Barcelona, 2006.

[2] G. Gravier. Spro home page – sfbcep. `http://www.irisa.fr/metiss/guig/spro/spro-3.3.1/node12.html`. last accessed 20th October 2009.

[3] T. Hain and P. Woodland. Segmentation and classification of broadcast news audio. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 98)*, 1998.

[4] ISCI. Icsi speech faq: 3.2 what are the wavfile data formats, and how can i manipulate wavfiles? `http://www.icsi.berkeley.edu/speech/faq/wavfile-fmts.html`. last accessed 20th October 2009.

[5] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre. A straightforward and effcient implementation of the factor analysis model for speaker verification. In *In Proceedings Interspeech*, page 4, 2007.

[6] S. Meignier, J.-F. Bonastre, and S. Igounet. E-hmm approach for learning and adapting sound models for speaker indexing. In *a Speaker Odyssey. The Speaker Recognition Workshop*, pages 175–180, 2001.

[7] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech & Language, Volume 20*, Issues 2–3:303–330, 2004.

[8] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book (for HTK Version 3.3).* Entropics Cambridge Research Lab., 2005.