UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Factor Analysis-based Approaches Applied to the Speaker Diarization Task of Meetings: A preliminary study

Pavel Tomasek, *Corinne Fredouille, Driss Matrouf*

University of Avignon, Computer Science lab – CERI/LIA

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

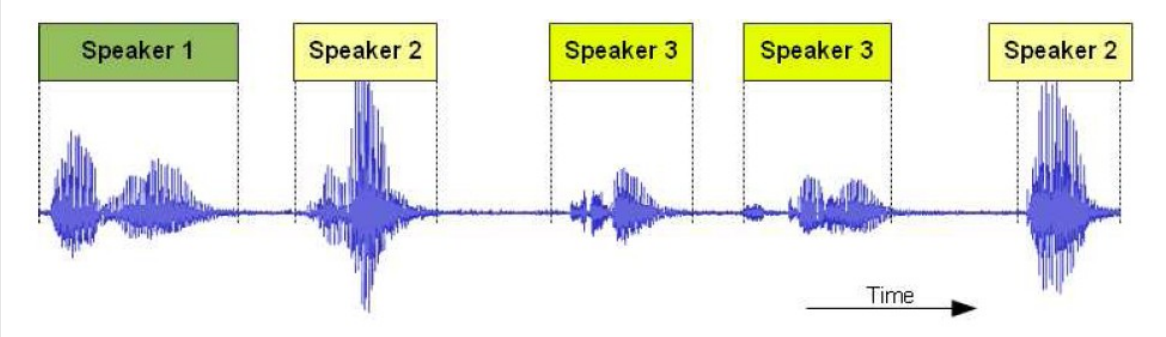
Outline

- Speaker diarization task : goal and system
- Factor analysis paradigm
- Objective of this preliminary study
- Investigation of two FA application strategies in the speaker diarization process
  - Protocols, experiments and results
- Perspectives

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Speaker Diarization 1/2

- « Who spoke when ? » task in the context of an audio document:
  - No a priori knowledge about the number of speakers neither on the speakers' identity



- Participation of the LIA to the NIST Rich Transcription evaluation campaigns since 2003 :
  - Top-down strategy-based speaker diarization system

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Speaker Diarization 2/2

- Top-down-based strategy :

  (1) Speaker Activity Detection : 2 pre-trained GMM models for speech/non speech used in a Viterbi decoding and MAP adaptation iterative process

  (2) Segmentation step based on :
  - an evolutive HMM (E-HMM) in which speakers are added at each iteration
  - Viterbi decoding coupled with the speaker model retrain (EM algorithm)

  (3) First resegmentation step based on :
  - Viterbi decoding coupled with the speaker model adaptation (MAP algorithm)

  (4) Second resegmentation step based on :
  - Similar to the previous one except for the parameterization

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Factor Analysis (FA) 1/2

- To model session variability
- Relevance proved in :
  - Speaker verification
  - Language identification
  - Video genre classification (based on audio uniquely)

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Factor Analysis (FA) 2/2

- Speaker modeling based on the classical MAP adaptation of a GMM/UBM :

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

- Speaker modeling within the FA paradigm :

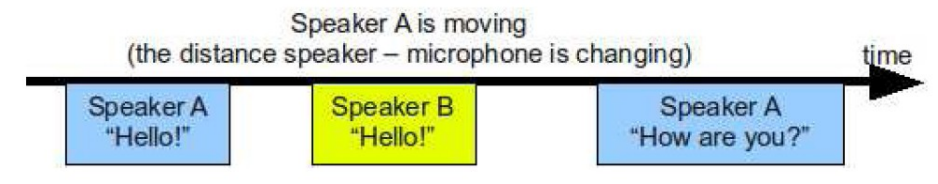
$$m_{(h,s)} = m + Dy_s + Ux_{(h,s)}$$

With $Ux_{(h,s)}$ *modeling the session variability*

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

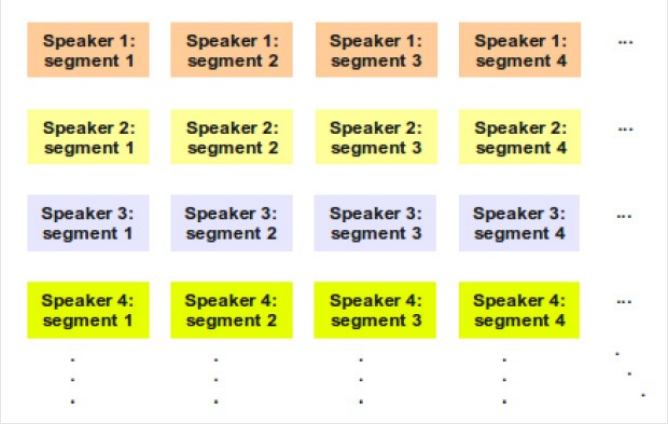
Objectives

- To study how can FA be useful for the speaker diarization task when dealing with a single audio file ?
- FA expectation : to deal with intra-session channel variability



Example of intra-session channel variability

  - By localizing a subspace (*U*) containing :
    - (1) the inter-segment variability
    - (2) the inter-speaker variability



UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Experimental protocol

- Experiments conducted on two datasets issued from meeting recordings :
  - Development dataset : **23 audio files** (NIST RT'04, '05, '06 campaigns), **7 different meeting rooms**, **10 to 18 mn** per record, **4 to 9 participants** per meeting (about 100 speakers)
  - Evaluation dataset : **7 audio files** (NIST RT'09 campaign), **17 to 27mn** per record, **4 to 7 participants** per meeting
- Recording conditions : table mounted multiple distance microphone (MDM conditions)
- Performance measurement : Diarization Error Rate (DER %)
- FA-modeling applied only in the third step of speaker diarization system

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Inter-speaker variability 1/2

- Assumption :
  - Main relevant speaker information located in a low dimension sub-space
  - Rest of the speaker information in the full space
  - FA helpful when very small amounts of training data available for the speakers

- Speaker modeling within the FA paradigm :

$$m_s = m + Dy_{(s,\,full)} + Ux_{(s,\,low)}$$

With *U* matrix common to all speakers

Inter-speaker variability 2/2

Eval. dataset, U matrix (rank 100) estimated on the dev. dataset

	DER % of Speaker Diarization systems		
	Baseline	**with Ux(s,low)**	**w/o Ux(s,low)**
EDI-10	3,2	3,2	39,5
EDI-15	33,8	34,6	37
IDI-16	15	14,9	33,9
IDI-10	14,1	15,7	16,5
NIS-14	47,9	49,4	46,1
NIS-15	20,4	16,6	19,5
NIS-09	18,7	19,3	18,7
Overall	*18,9*	*19,1*	*29,6*

- Large difference between the overall DER for the FA-based systems :
  - Relevant speaker information in the low dimension sub-space, but not enough to outperform the baseline system
  - Rather expected given the limited number of speakers (100) in the dev. dataset regarding the rank of the *U* matrix

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Inter-segment variability 1/4

- Assumption :
  - Inter-segment information in a low dimension sub-space

- Speaker modeling within the FA paradigm :

$$m_s = m + Dy_s + Ux_{(s,seg)}$$

With *U* matrix common to all speakers, estimated on the development dataset

- *Note : if the distinction between segments is made per speaker for the U estimation, it is not applied in this way for the speaker diarization task for which all the segments belonging to the same speaker are merged to estimate the speaker model*

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Inter-segment variability 2/4

Eval. Dataset, U matrix (rank 10) estimated on the dev. dataset, segments longer than 1s

	DER % of Speaker Diarization systems		
	Baseline	**with Ux(s,seg)**	**w/o Ux(s,seg)**
EDI-10	3,2	3,1	3,2
EDI-15	33,8	34,3	47
IDI-16	15	14,6	12,3
IDI-10	14,1	15,6	14
NIS-14	47,9	49,4	42
NIS-15	20,4	13	19
NIS-09	18,7	19,1	18,1
Overall	*18,9*	*18,5*	*19,3*

- Slight difference between the overall DER for the FA-based systems :
  - Some speaker information present in the inter-segment component, but not significant to influence the speaker diarization process

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Inter-segment variability 3/4

Segment filtering for U matrix estimation

Eval. Dataset, with $Ux_{(s,seg)}$ (U matrix rank : 10 estimated on the dev. dataset), variable minimum segment length

	DER % per min. segment duration				
	0s	**1s**	**2s**	**5s**	**10s**
EDI-10	3,2	3,1	3,1	3	3
EDI-15	34,5	34,3	34,3	34,2	34
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	**4,9**	15	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*

- Basic idea : remove very short segments used for the U matrix estimation
  - Quite stable performance according to the different minimum segment durations except for the *NIS-15* file for which an absolute gain of 15,5% is observed with segments longer than 1s

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Inter-segment variability 4/4

FA followed by a new resegmentation step

Eval. Dataset, with $Ux_{(s,seg)}$ (U matrix rank : 10 estimated on the dev. dataset), minimum segment length : 1s

	FA	FA+Reseg.
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***

- Very slight gain, especially for the *NIS-15* file :
  - Not significant, but demonstrates that FA-based speaker modeling brings relevant changes in the segmentation outputs, useful for a classical Viterbi decoding/MAP adaptation iterative process

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Summary

- Two strategies investigated to apply FA-based speaker diarization process :
- To deal with inter-speaker variability :
  - Relevant speaker information demonstrated in the low sub-space despite the small number of speakers used for the U matrix estimation (more than 1000 speakers classically used in the speaker verification task)
- To deal with the inter-segment variability :
  - Poor improvement except for one file
  - Tends to show that some information about the speaker is brought by the inter-segment component : intrinsic characteristics of speakers' voice or speaker's position according to the microphones ?
  - Useful for an addition resegmentation step

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Perspectives

- To investigate larger number of speakers when dealing with the inter-speaker variability
- Regarding the inter-segment variability :
  - To investigate a **more « natural » application of the FA approach** in the speaker modeling (similar to the U matrix estimation) : considering **segment-dependent speaker models** rather than a single speaker model
  - To use the FA when dealing with the **multiple distant microphones** (as opposed to the beamformed single virtual channel) in order to emphasize information brought by each individual microphone
- To investigate the application of the FA-based speaker modeling in the first step of the speaker diarization system (**segmentation step**)

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Thank you for your attention
Any questions ???