Application of Factor Analysis to Speaker Diarization

Presentation of ERASMUS internship project in LIA

September 2009 – January 2010

Pavel Tomášek

xtomas23@stud.fit.vutbr.cz

Supervised by: *Corinne Fredouille and Driss Matrouf*

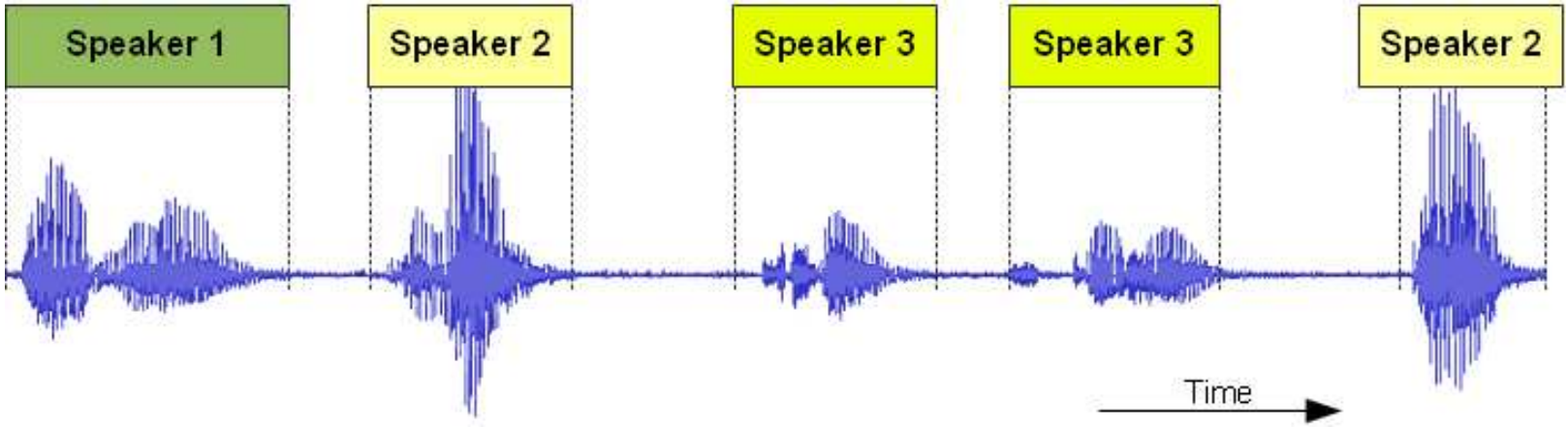
Outline

- Speaker Diarization
- Factor Analysis
- Experiments
- Results
- Summary
- Further work

Speaker Diarization (1/2)

Answer for question: *"Who spoke when?"*

No a priory knowledge about speakers



Speaker Diarization (2/2)

Application:

- Simplification of management of audio databases (audio databases indexation)
- Speaker adaptation for speech recognition

Baseline Diarization System

Components:

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

Factor Analysis (1/4)

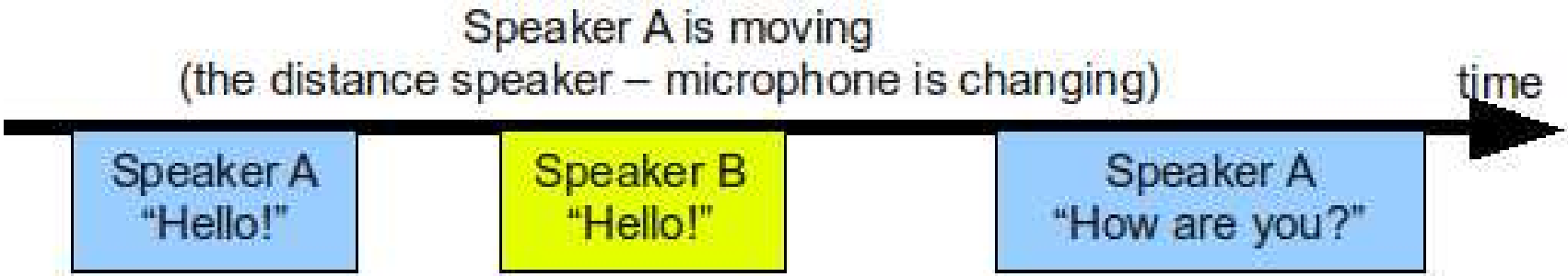
Very useful in modeling session variability in:

- *speaker verification*
- *language identification*

What about applying factor analysis in speaker diarization domain?

Factor Analysis (2/4)

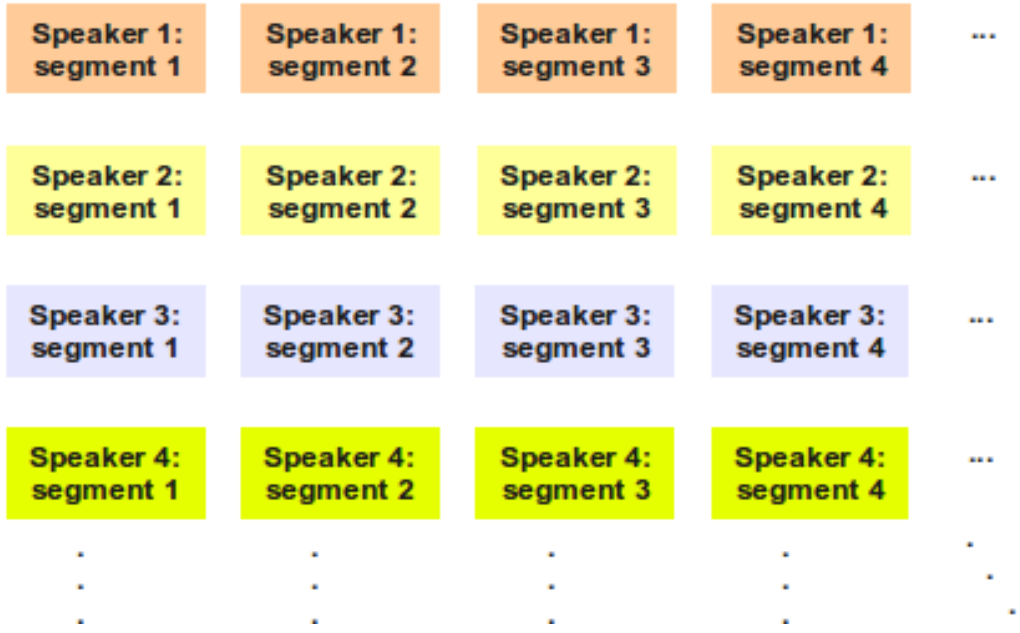
How can factor analysis be useful in speaker diarization domain?



Factor Analysis (3/4)

Objectives:

- Localize a subspace (U) containing the channel variability



Factor Analysis (4/4)

Original modeling:

- A standard GMM-UBM using MAP adaptation
- $m_{(h,s)} = m + Dy_s$
- Nothing which represents the channel variability

New modeling using factor analysis:

- $m_{(h,s)} = m + Dy_s + Ux_{(h,s)}$
- $Ux_{(h,s)}$ represents the channel variability, D is a diagonal matrix $MFxMF$, F is a dimension of a feature space, 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

Experimental Protocol

- Development set: *NIST campaign 2008 (RT)*
- Evaluation set: *NIST evaluation campaign 2009 (RT)*
- Toolkit: *ALIZE*, *LIA RAL*
- Parametrization:
  - experiments with 128 and 256 Gaussians
  - experiments with 21 and 34 coefficients

Experiments

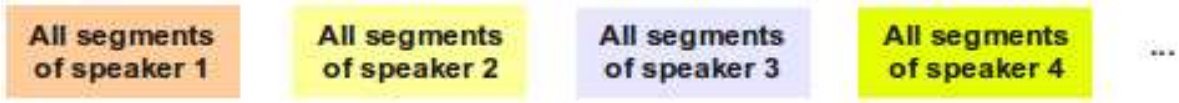
Various approaches for channel matrix estimation:

- Various index structures (variability between speaker clusters or between segments of each speaker)
- Various minimal segment durations (no filtration, 1s, 2s, 5s, ...)
- Concatenation (sub-clustering) of short segments (no concatenation, 1s, 2s, ...)
- Various ranks (5, 10, ..., 100, 120)

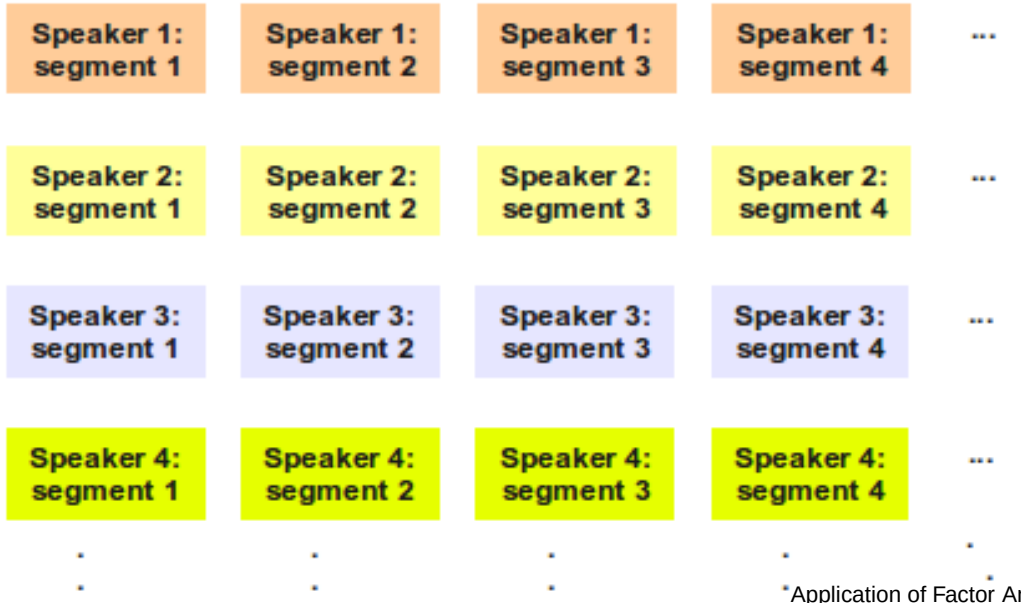
Speaker modeling with or without Ux

Experiments – Index Structures

Channel matrix estimated by variability between speaker clusters:



Channel matrix estimated by variability between segments of each speaker:



Experiment 1/3

Various speaker modeling:

(Specification: 128 Gaussians, 34 coefficients; U matrix details: rank 100, speakers on one line; one training iteration of speaker model)

	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*

Experiment 2/3

Various speaker modeling:

(Specification: 128 Gaussians, 34 coefficients; U matrix details: rank 10, speaker per line, segments shorter than 1s eliminated; one training iteration of speaker model)

	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*

Experiment 3/3

Various minimal segment durations (filtration of segments for channel matrix estimation):

(Specification: $m + Dy + Ux$, 128 Gaussians, 34 coefficients; U matrix details: rank 10, speaker per line; four training iteration of speaker model)

	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*

The Best Results

Overall DER without and with FA: (Specification: $m + Dy + Ux$, 128 Gaussians, 34 coefficients; U matrix details: rank 10, speaker per line, segments shorter than 1s eliminated; four training iteration of speaker model)

	Diarization Error Rate (%)	
	Original	After FA
EDI_20071128-1000	3.21	3.11
EDI_20071128-1500	33.82	34.27
IDI_20090128-1600	14.95	14.58
IDI_20090129-1000	14.08	15.61
NIST_20080201-1405	47.93	49.17
NIST_20080227-1501	20.41	4.90
NIST_20080307-0955	18.67	19.13
Overall	*18.93*	*17.66*

Progress of NIST_20080227-1501

Why is NIST_20080227-1501 so well performing?

	Original error rate	Error rate after FA	Gain
NIST_20080227-1501	20.41%	4.90%	**15.51%**

One possible explanation:

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

Progress of NIST_20080227-1501

What about another step of standard re-segmentation without factor analysis?

	Original error rate	Error rate after FA	ReSeg error rate
NIST_20080227-1501	20.41%	4.90%	**3.46%**

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

Summary

- Data-dependence
- Average gain is positive but not very high, *1.27%*
- Experiments show that application of factor analysis to speaker diarization can be useful

Further work

- Future work is advised to be based on this work
- Another approaches of estimation of channel variability (another index structures or segmentation constraints)
- Speaker modeling using also probability (a probability that a segment belongs to a speaker)
- Utilization of more training data
- Investigation of NIST_20080227-1501, why it is so well performing

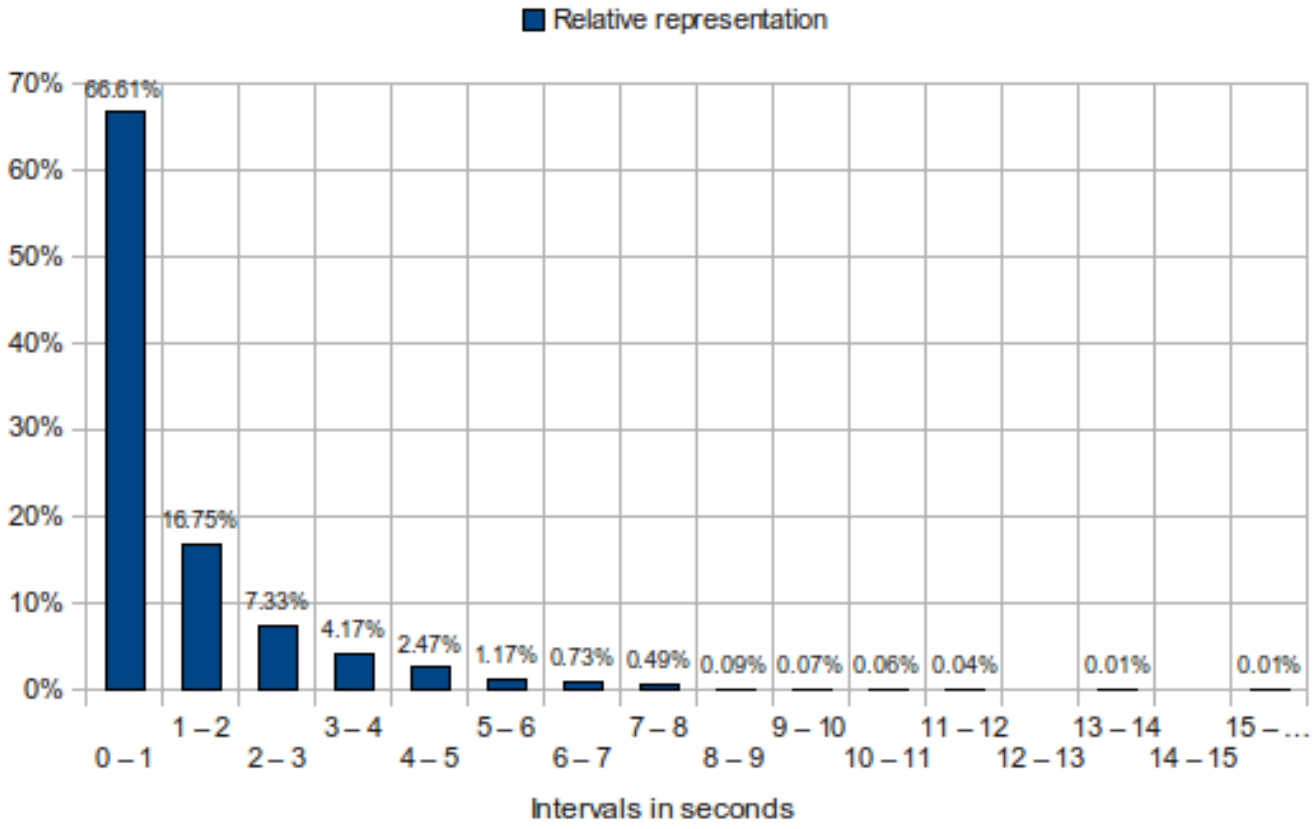
Finally

Thank you for your attention.

Any questions?

Statistics of Evaluation Data (1/2)

Relative representation of number of segments in duration intervals for *reference data:*



Statistics of Evaluation Data (2/2)

Relative representation of number of segments in duration intervals for *LIA speaker diarization system output:*