Intra-Session Variability Compensation and a Hypothesis Generation and Selection Technique for Speaker Segmentation

C. Vaquero[1,2], A. Ortega[1], E. Lleida[1]

[1]VIVOLAB, GTC, I3A,
University of Zaragoza, Spain
[2]Agnitio S.L, Spain

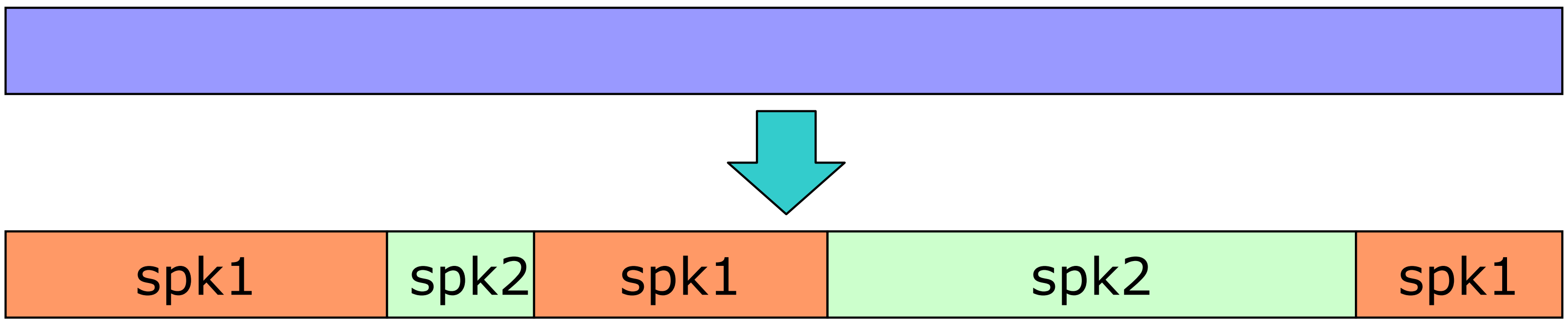
Instituto Universitario de Investigación en Ingeniería de Aragón
Universidad Zaragoza

1542 Universidad Zaragoza

AGNITIO

Segmentation of two-speaker conversations

- Speaker Diarization Problem:
  - Who spoke when?
- Number of speaker is known and limited to two.
  - Easier task
  - Knowing the boundaries solve the problem: segmentation problem



Segmentation system

- New approaches for the segmentation of 2-speaker conversations: Factor Analysis using Eigenvoices[1]

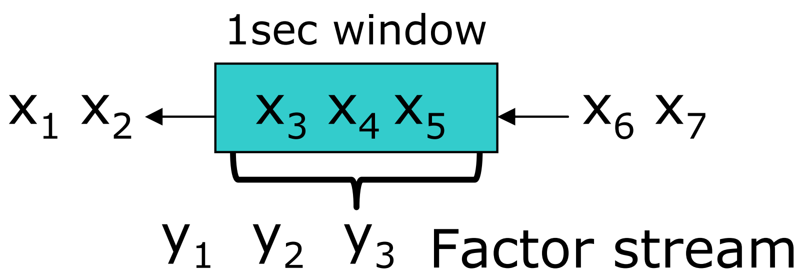
$$M(s) = M_{UBM} + Vy(s)$$

  - $M(s)$ speaker GMM-sv, M_{UBM} UBM GMM-sv $(Dx1)$
  - V models inter-speaker variability (DxR)
  - $y(s)$: speaker factors $(Rx1, R<<D)$
  - Fewer parameters to estimate
  - Need less data to model speaker
  - We can estimate $y(s)$ on small segments

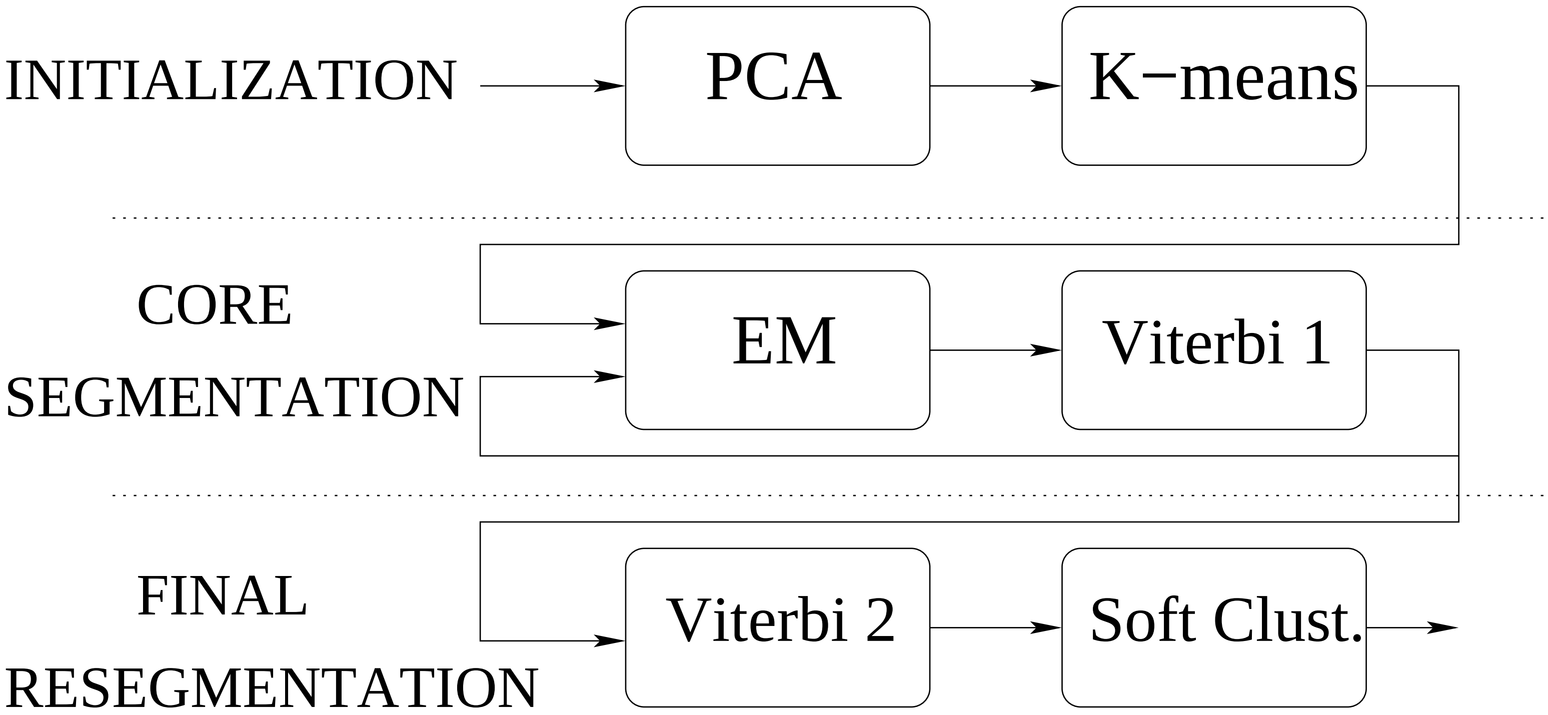
[1]Castaldo, F. et al. "Stream Based Speaker Segmentation Using Speaker Factors and Eigenvoices", ICASSP, Las Vegas, NV, USA, 2008.

Segmentation System

- Extract Speaker factors



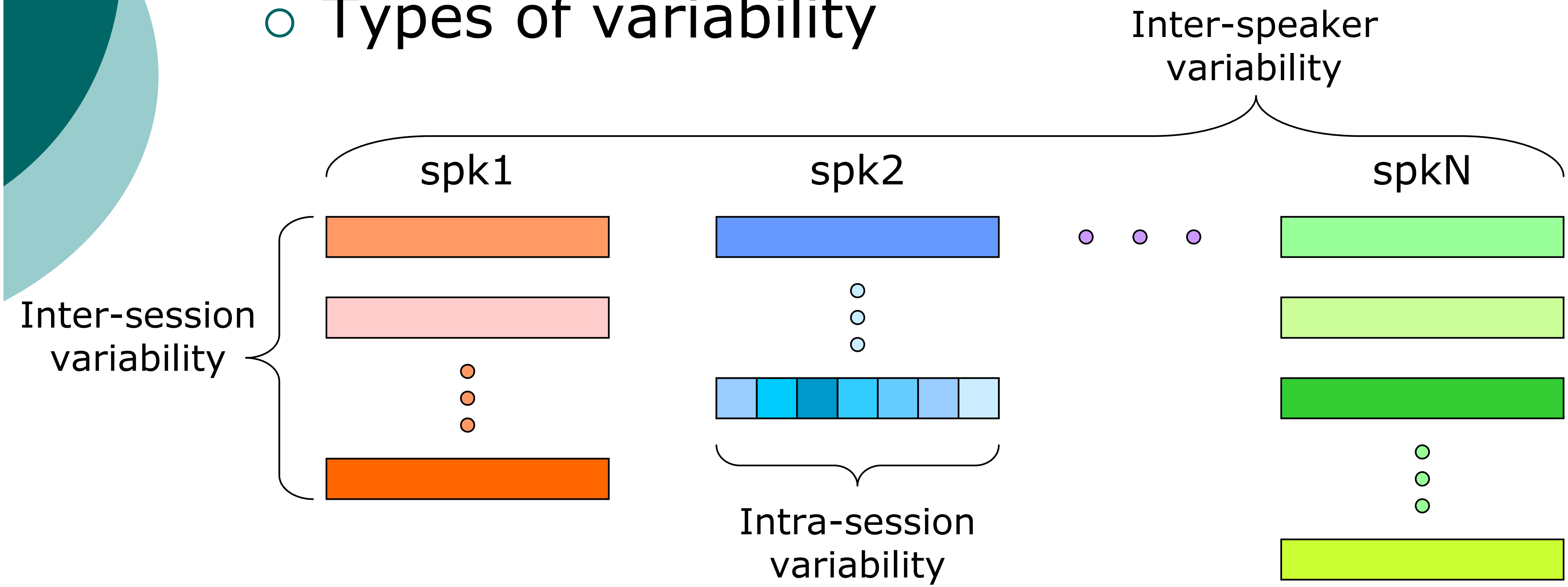
- Block Diagram[2]



[2]Vaquero, C. et al. "Confidence measures for speaker segmentation and their relation to speaker verification", INTERSPEECH, Makuhari, Chiba, Japan, September 2010

Variability Compensation

- Types of variability



- Only inter-speaker variability is modeled
- Are other types of variability degrading speaker segmentation performance?

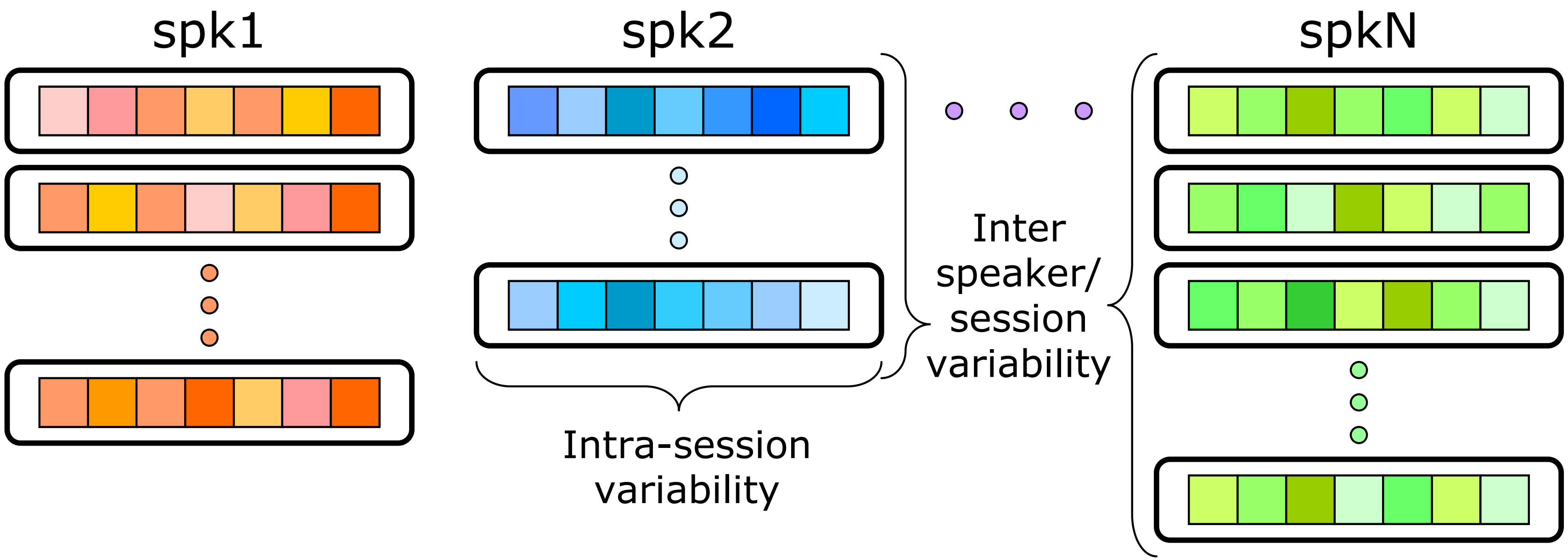
Variability Compensation

- Inter-session variability compensation
  - Very important for speaker recognition
  - Not for speaker segmentation/diarization
    - Unsupervised task: we do not have prior information of the speakers in a session
    - Inter-session variability may help to separate speakers during a session
      - They may use different communication channels

Variability Compensation

- Intra-session variability compensation
  - State of the art speaker recognition systems do not compensate for it
  - Important for speaker segmentation/diarization
    - Most systems are based on clustering of very small pure segments
    - Compensating the variability among small segments for a single speaker may improve the clustering performance

Intra-Session Variability Compensation



- Obtain a stream of speaker factors from each recording
- Consider every session as a different class
- Model inter-speaker/inter-session variability as between-class variance
- Model intra-session as within-class variance

Intra-Session Variability Compensation

- Linear Discriminant Analysis (LDA)
  - Technique for dimensionality reduction
  - Maximize between-class variance
  - Minimize within-class variance
- Within Class Covariance Normalization (WCCN)
  - Normalize within-class covariance to be the identity matrix for all classes
- Both have been successful for inter-session compensation in speaker recognition[3]

[3]Dehak, N. et al. "Front-end factor analysis for speaker verification", IEEE Transactions on Audio, Speech, and Language Processing, August 2010

Evaluation

- Experimental setup
  - NIST SRE 2008 summed channel condition
  - 2213 five minute telephone conversations
  - Speech/non speech labels are given
  - Performance in terms of speaker segmentation error
    - Not taking into account overlapped speech
    - 0.25 sec forgiveness collar

Evaluation: Intra-session variability compensation, small UBM (256g)

Segmentation system (spk factors)	Seg Err (%)
Baseline (20) no reseg	3.0
WCCN (20) no reseg	**2.5**
Baseline (50) no reseg	2.8
LDA (50 to 20) no reseg	2.6
WCCN (50) no reseg	**2.0**
LDA (50 to 20) + WCCN no reseg	2.4

Features: 12 MFCC

Evaluation: Intra-session variability compensation + resegmentation

Segmentation system (factors)	Seg Err (%)
Baseline (20) + reseg	2.1
WCCN (20) + reseg	**1.7**
Baseline (50) + reseg	2.1
WCCN (50) + reseg	**1.7**

Features: 12 MFCC

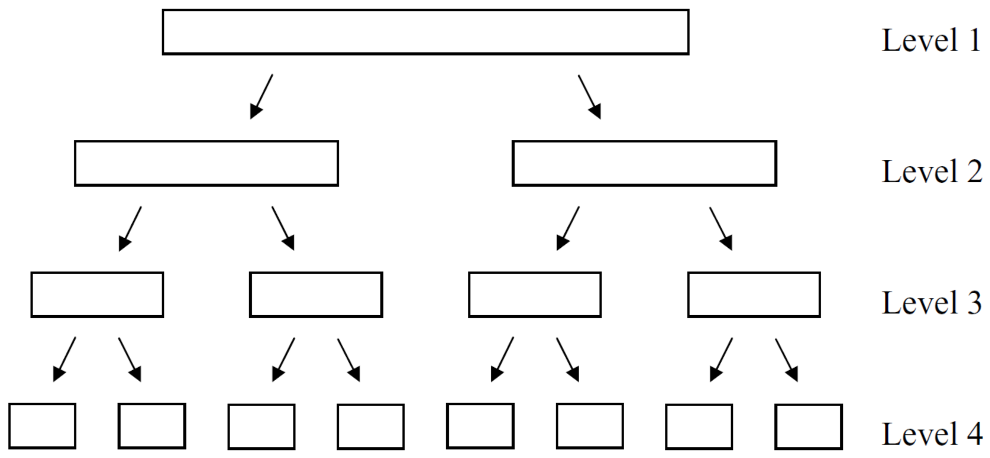
Evaluation, Intra-session variability compensation: New results with a larger UBM (1024g)

Segmentation system (factors)	Seg Err (%)
Baseline (50)+reseg	1.8
WCCN (50)+reseg	**1.4**
Baseline (100)+reseg	1.9
LDA (100 to 50)+reseg	1.5
WCCN (100)+reseg	1.4
LDA (100 to 50)+WCCN+reseg	**1.3**

Features: 19 MFCC + delta

Segmentation Hypothesis Generation and Selection

- Iteratively split the conversation into two halves
  - Obtain 4 levels
- Segment every slice separately
- Select best segmented slices (confidence measures)
- Agglomerate best slices until we have 2 spks
- Run Viterbi reseg
  - For every level
  - MFCC
  - 32 Gaussians
- Select best level
  - Confidence measures
  - Majority voting



Segmentation Hypotheses Generation and Selection: Confidence Measures[2]

- BIC
  - MFCC space
  - 32 comp. GMM speaker models, 64 comp GMM Null hyp
  - BIC_{2spks}-BIC_{Null}
- KL
  - Speaker factor space
  - Gaussian speaker models
- Fusion of both measures
  - Using FoCal toolkit[4]
  - Optimized to segregate those files having less than 1% segmentation error

[2]Vaquero, C. et al "Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification", in Proc Interspeech, Makuhari, Japan, 2010.

[4]Brümmer, N. online: Online: http://sites.google.com/site/nikobrummer/focal

Evaluation: Hypothesis Generation and Selection, small UBM (256g)

Segmentation system	Seg Err (%), no comp	Seg Err (%), WCCN
Level 1	2.1	1.7
Level 2	2.1	1.8
Level 3	2.3	2.1
Level 4	2.5	2.0
Hypothesis Selection	**1.9**	**1.7 (1.5*)**
Best Selection	1.1	0.9

Features: 12 MFCC

***Major Voting**

Evaluation, Hypothesis Generation and Selection: New results with a larger UBM (1024g)

Segmentation system	Seg Err (%), LDA (100-50)+WCCN
Level 1	1.3
Level 2	1.2
Level 3	1.5
Level 4	1.7
Hypothesis Selection	**1.0**
Best Selection	0.7

Features: 19 MFCC + delta

Conclusions

- Intra-session variability compensation
  - It helps for speaker segmentation
  - WCCN obtains better performance than LDA and similar to LDA+WCCN
  - # spk factors ↑ ⇒ computational cost ↑↑
    - WCCN is better for low computational cost applications
    - LDA (100 – 50)+WCCN is the best configuration.
  - WCCN (20) reduces seg error: **2.1%** to **1.7%**
  - LDA (100–50)+WCCN: **1.9%** to **1.3%** (large UBM)
  - WCCN helps the PCA+K-means initialization
- Hypothesis Generation and Selection
  - No compensation: **2.1%** to **1.9%**, up to 1.1%
  - WCCN: **1.7%** to **1.5%** (major voting) up to 0.9%
  - LDA (100–50)+WCCN: **1.3%** to **1.0%** (large UBM) up to 0.7%

Thank you!