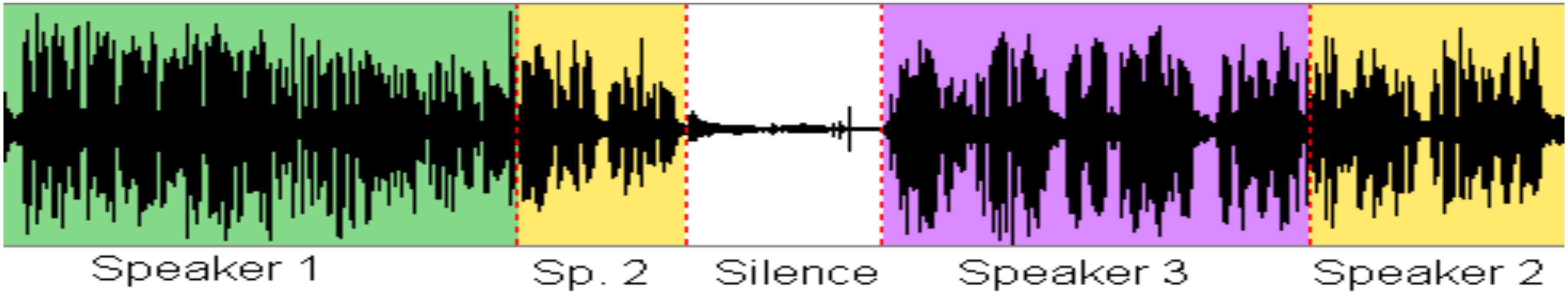
INVESTIGATION OF SEGMENTATION IN I-VECTOR BASED SPEAKER DIARIZATION OF TELEPHONE SPEECH

ZČU, FAV, KKY

Zbyněk Zajíc, Marie Kunešová, Vlasta Radová

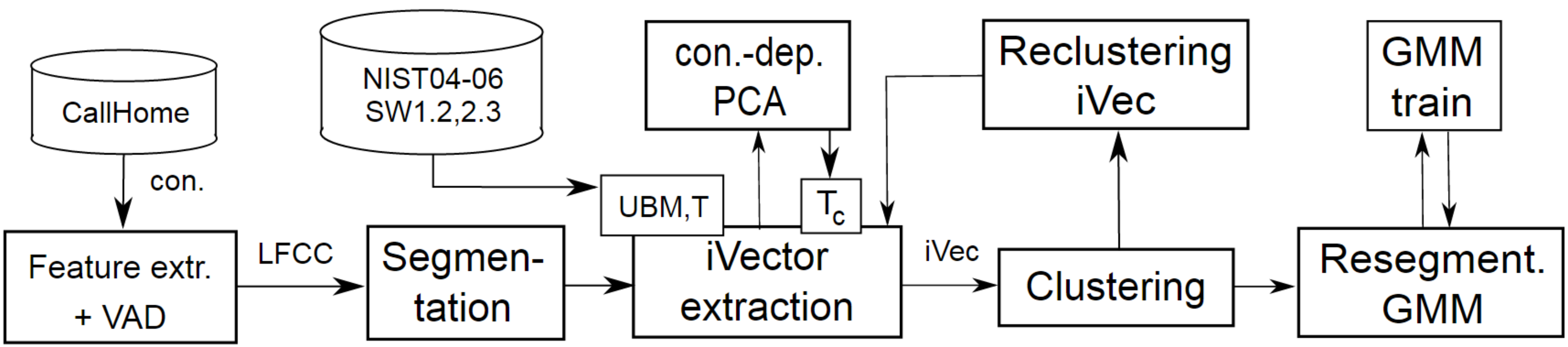
Introduction

- Speaker Diarization = “Who spoke when?”
- No prior knowledge about the speakers
- Most common approach – split the audio stream into short speech segments, then cluster them
- Segments – obtained by splitting speech by constant length (CL) or through speaker change detection (SCD)
- In telephone speech diarization, CL is typically used
- Our goal: to compare the two options



Speaker Diarization System

- Feature extraction – LFCC
- Voice activity detection
- Segmentation – SCD or constant length segments
- i-Vector extraction – i-vectors from segments, PCA
- Clustering - K-means + iterative reclustering
- Resegmentation - GMM-based



Segmentation

- **A) Constant Length Segments**
  - Speech regions are split in fixed length intervals
  - Boundaries do not correspond to speaker change points – segments may contain more than 1 speaker
    - Because of this, shorter segments are preferable
    - But: at least 1-2 seconds needed for i-vector extraction, we use 2s with 1s of overlap
- **B) Speaker Change Detection**
  - Audio is split in likely speaker change points
  - Uses a pair of sliding windows, computing the distance between their contents
  - Peaks in the distance signify a likely speaker change

Segmentation – SCD (1/2)

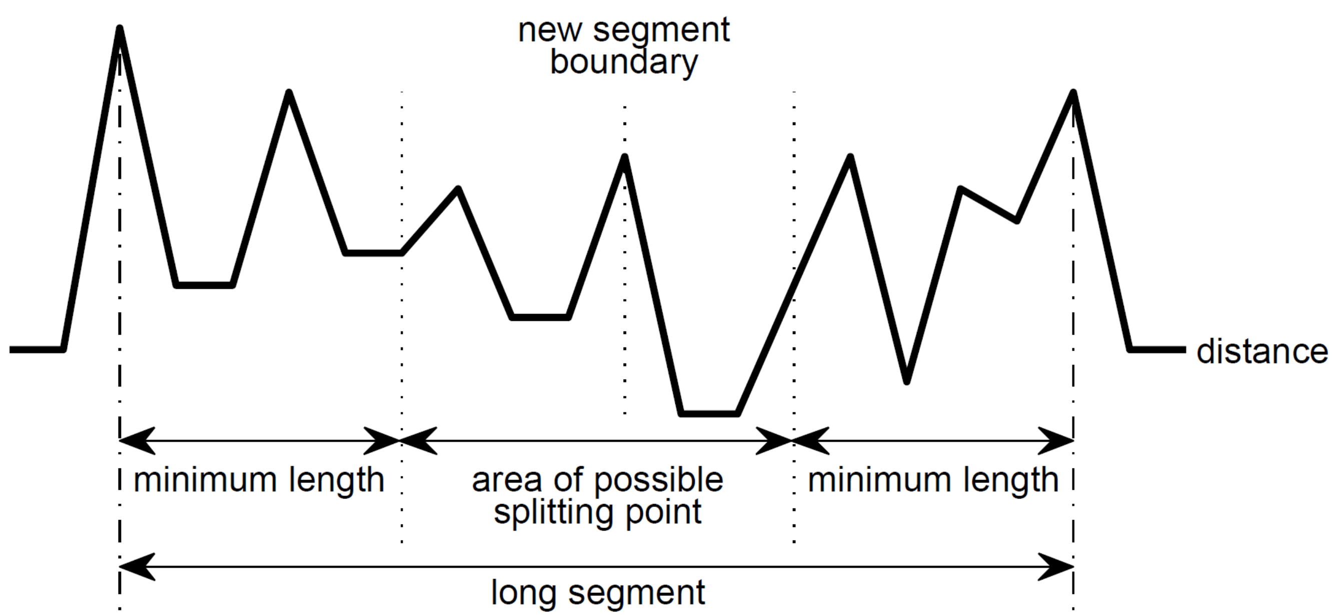
- To calculate distance, we use the Generalized Likelihood Ratio (GLR) - distance between windows is defined as

$$d(X_i, X_j) = \log \frac{L(X_i \quad X_j \mid M)}{L(X_i \mid M_i) \cdot L(X_j \mid M_j)}$$

  - M, M_i and M_j are single Gaussians estimated from the data
- Peaks in the distance are measured by topographic prominence, i.e. how much they stand out within the signal
- For consistency with constant length segmentation, we define a minimum and maximum segment length and use a two-step algorithm

Segmentation – SCD (2/2)

- Step 1: find the most likely speaker changes – peaks with a prominence higher than a threshold
- Step 2: further split long segments, so that all have length within the target range
  - Segments are split either at the most prominent peak within target area or at the point where the distance is highest



i-Vector Extraction

Each segment is represented by a single i-vector:

1. A supervector of GMM based statistics is accumulated
   - Supervector contains the first and zeroth statistical moments of the acoustic features, related to a Universal Background Model (UBM)
   - The UBM: a GMM trained on a large amount of data
2. Dimensionality reduction of the supervector via Factor Analysis → i-vector

$$= m_0 + Tw +$$

   - $\boldsymbol{\phi}$ - supervector, $\boldsymbol{m_0}$ - mean vector of ϕ or UBM's mean supervec., $\boldsymbol{T}$ - total variability space matrix, $\boldsymbol{w}$ – i-vector of one segment
3. Further dimensionality reduction: Conversation-dependent Principal Component Analysis (PCA)

Clustering

- Extracted i-vectors are clustered to determine which segments were produced by the same speaker
- Clustering was based on cosine similarity between individual i-vectors:

$$sim(w_1,\ w_2) = \frac{w_1^T\ w_2}{\|w_1\|\cdot\|w_2\|}$$

- Test data consisted only of conversations with 2 speakers → we used k-means algorithm with 2 target clusters
- Reclustering:
  - After clustering, we compute one i-vector for each cluster and reclassify the individual segments
  - The process is repeated until convergence

Resegmentation

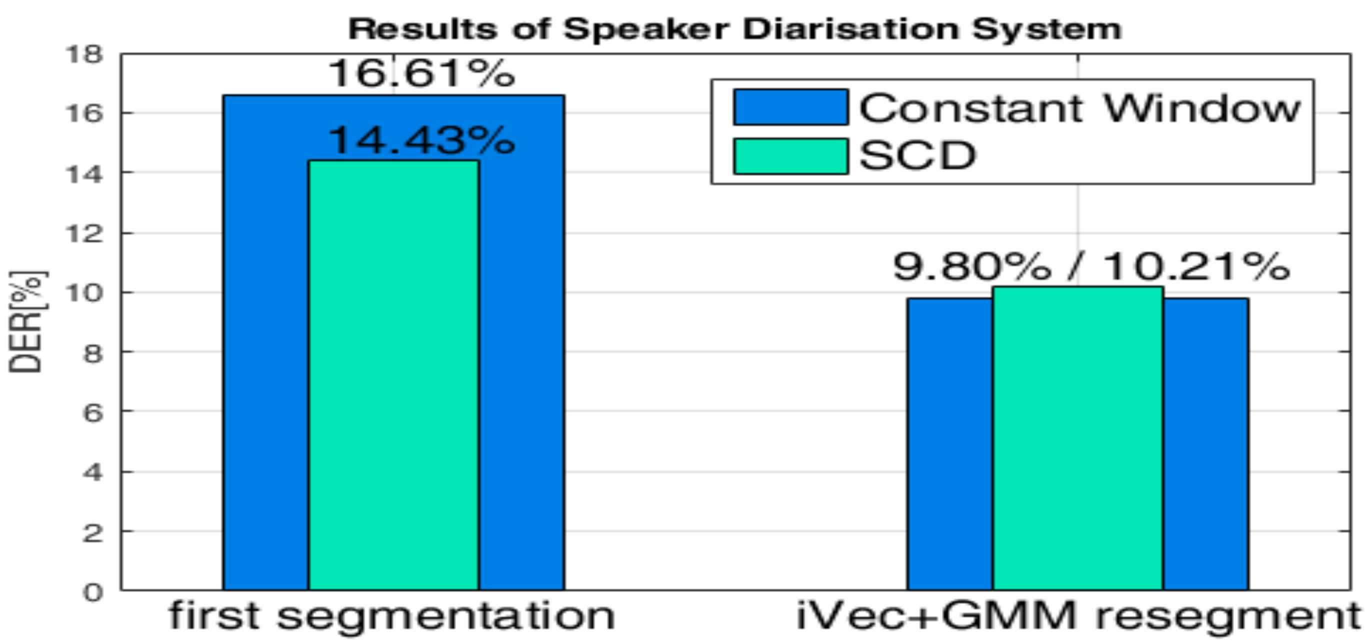
- Segment boundaries are not completely accurate
  - particularly with CL segmentation
- Reclustering works with original segments – boundaries between them are unchanged
- → Resegmentation - used to refine imprecise segment boundaries
  - We train a GMM for each cluster, using the original acoustic features
  - Individual speech frames are classified based on the likelihood of each GMM, with Gaussian smoothing
  - Results in more accurate speaker boundaries

Experiments

- Experiments compared the two segmentation approaches
- Test Data:
  - CallHome corpus of telephone speech
  - Only English conversations with 2 speakers were used
  - ~100 spontaneous conversations, around 5-10 min each
- Training Data for i-vector extraction:
  - NIST SRE (04, 05, 06) and Switchboard corpora
- Performance measured as Diarization Error Rate (DER)
  - In the final results, extra silences were added based on the reference transcripts – error values represent speaker error only

Results

- SCD gave better results at the first clustering stage
- But: after resegmentation, differences were minimal
- SCD is more computationally demanding
- Conclusion: segmentation by constant length is sufficient for the target system



THANK YOU FOR YOUR ATTENTION