Investigation of Speaker Embeddings for Cross-show Speaker Diarization

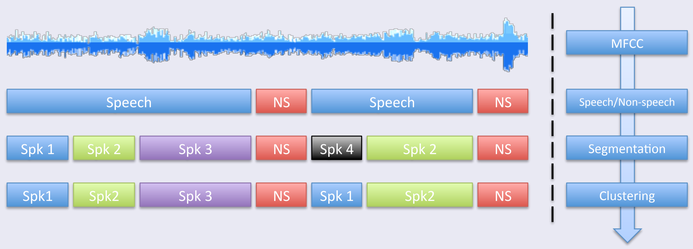
Mickael Rouvier, Benoit Favre

Aix-Marseille Université, Laboratoire d'Informatique Fondamentale de Marseille, CNRS, France
{firstname.lastname}@lif.univ-mrs.fr

Summary: Diarization with speaker embeddings

- **Goal:** Extract compact features that characterize speakers.
- **Method:** Learn a set of high-level feature representations through deep learning.
- **Application:** Speaker embedding is applied to Single and Cross-show Speaker Diarization.
- **Results:** The new representation brings an improvement over i-vectors
  - **Single-show condition:** Shallow hidden layers give best results (0.19 points)
  - **Cross-show condition:** Deeper hidden layers yield better performance (0.82 points)
- **Conclusion:** Deep representations model higher level features which help generalizing to different acoustic conditions.

Speaker Diarization

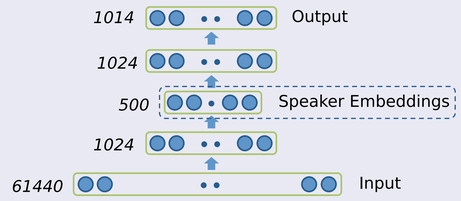


- **Task:** The goal of speaker diarization is to annotate temporal regions of audio recordings with speaker labels, in order to answer the question "who spoke and when".
- **Single-show condition:** Each show in the collection is processed independently.
- **Cross-show condition:** The same speaker in multiple shows has to be labeled with the same identity.
- **Steps:**
  - Speech/non-speech segmentation: HMM
  - Segmentation: Gaussian Likelihood Ratio (GLR)
  - Local Clustering: ILP Clustering (process individually each show in the collection)
  - Global Clustering: ILP Clustering (process globally the collection, only for cross-show condition)

Acknowledgments: this work has been carried out with the support of the A*MIDEX project (n° ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French Government program, managed by the French National Research Agency (ANR). The Tesla K40 used for this research was donated by the NVIDIA Corporation.

Speaker Embeddings

- **Problem:** i-vector/PLDA pipeline not efficient on short segments
  - i-vector extracted on total variability space
  - PLDA has difficulty to disentangle useful information from background noise
- **Goal:** Extract new features on the **speaker space**.
- **Method:** Train a DNN to perform the speaker identification task
  - Input: First-order Baulm-Welch statistics (61,440 dimension)
  - Output: Speaker identification (1,014 dimension)
  - Speaker embeddings: extract one of the hidden layers as the new feature representation
- **Observation:** Although learned through the identification task, speaker embeddings are shown to be effective for speaker verification



Experiments

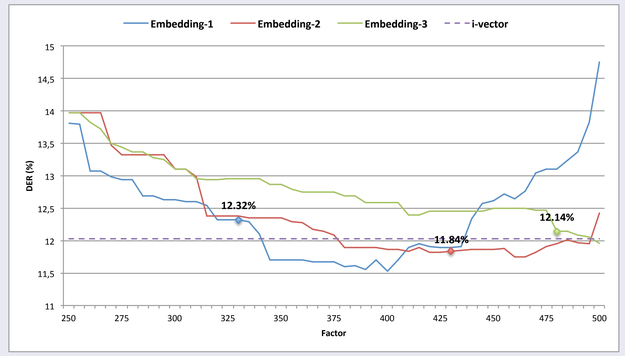
- **Speaker diarization:** based on the LIUM Speaker diarization toolkit
- **Corpus:**
  - **Train:** 300h of French broadcast news (ESTER, ETAPE, EPAC, REPERE)
  - **Dev/Test:** REPERE 2013 French evaluation campaign (3h/10h)
- **Speaker embeddings:** (all params tuned on dev)
  - DNN: 3 hidden layers
  - Function activation: ReLU
  - Speaker embedding layer: 500 dimension
  - Other hidden layers: 1024 dimension
- **i-vectors:** dimension 150, from 1024 UBM
- **Normalization:** Whitening followed by Length-normalization
- **PLDA:**
  - On i-vectors: 25 dimensions
  - On Speaker-embeddings: 200 dimensions
- **Metric:** Diarization Error Rate (DER)

$$DER = \frac{\#Spk + \#Miss + \#FA}{\#Total} \quad (1)$$

Experiments

- **Single-show condition**

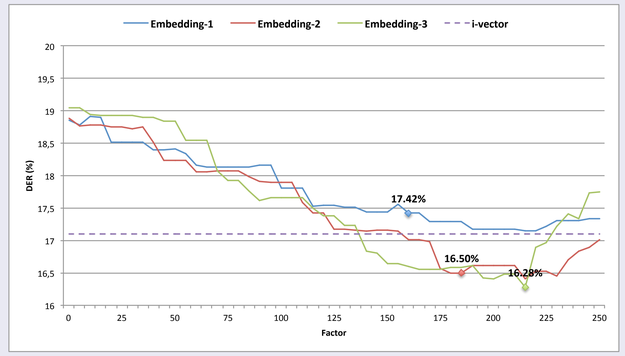
  Results obtained on single-show speaker diarization. We observe that shallow hidden layers give best performance:



  Results in DER obtained by using the representation extracted from the different hidden layers in single-show diarization.

- **Cross-show condition**

  Results obtained on cross-show speaker diarization. We observe that deeper hidden layers yield better performance:



  Results in DER obtained by using the representation extracted from the different hidden layers in cross-show diarization.

Conclusion

- Deep representations model higher level features which help generalizing to different acoustic conditions.
- We plan to explore different input spaces for training representations.
- We plan to test embeddings on different tasks.