Bayesian Models in Machine Learning

Bayesian GMM for Speaker Diarization (VBx-GMM)

Lukáš Burget, Mireia Diez

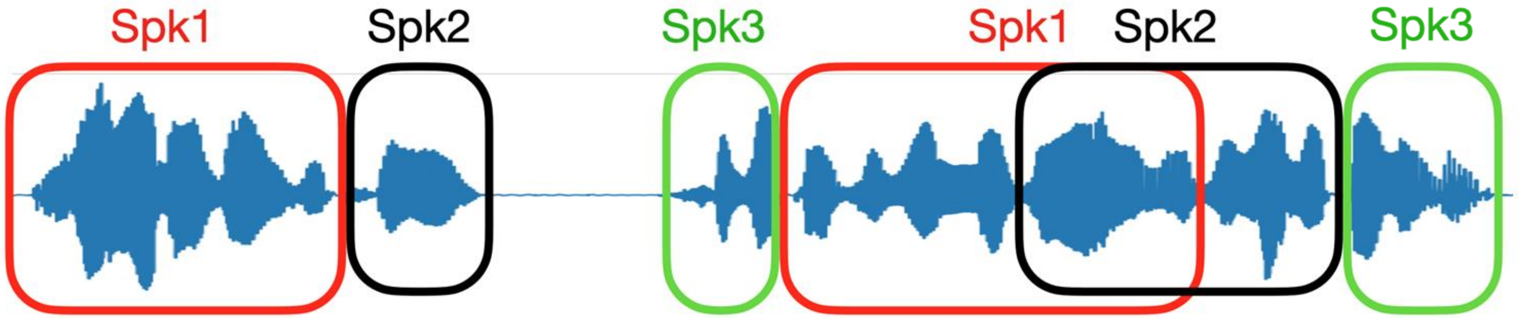
BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

BAYa lectures, December 2023

Speaker Diarization

What is it?

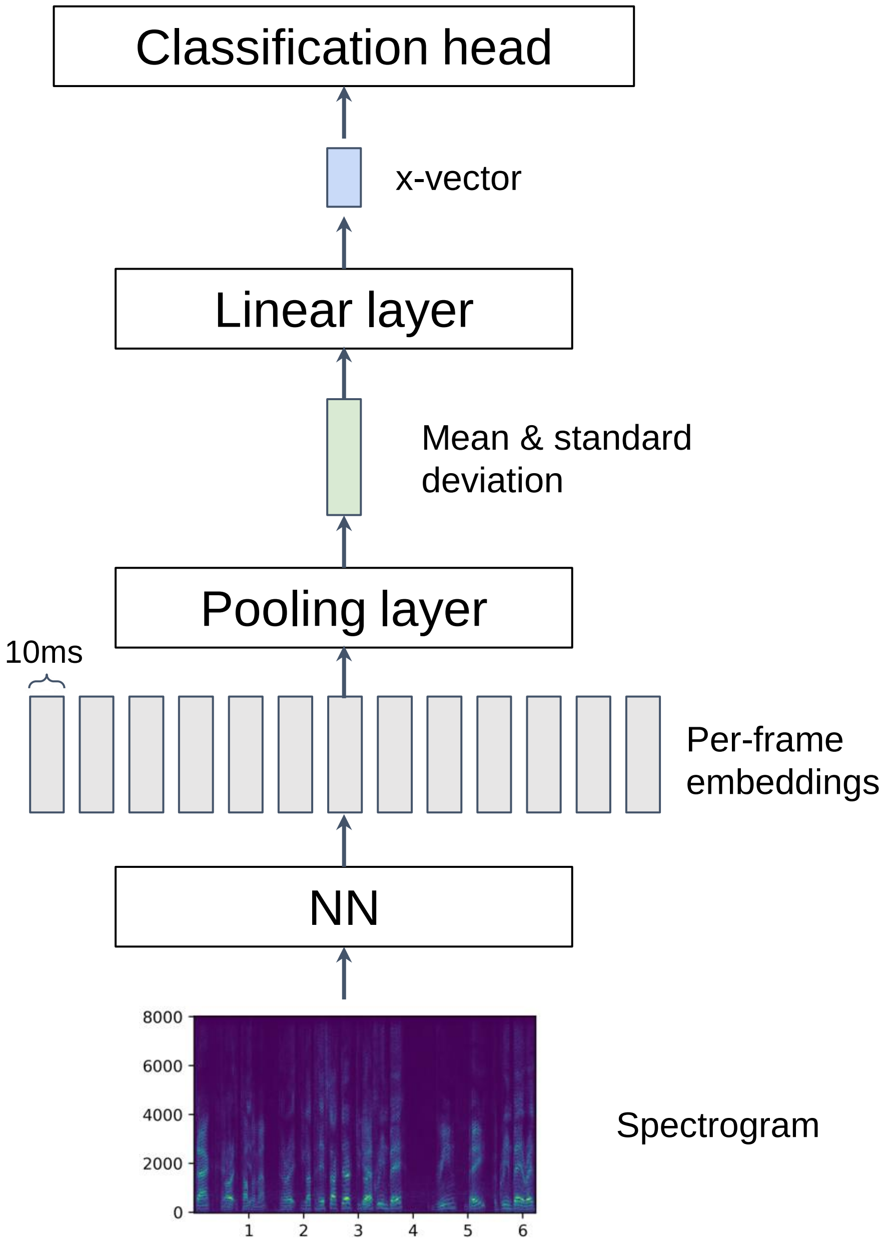
- The task of automatically determining the speaker turns in a recording of a conversation or finding "who spoke when"



Diarization system

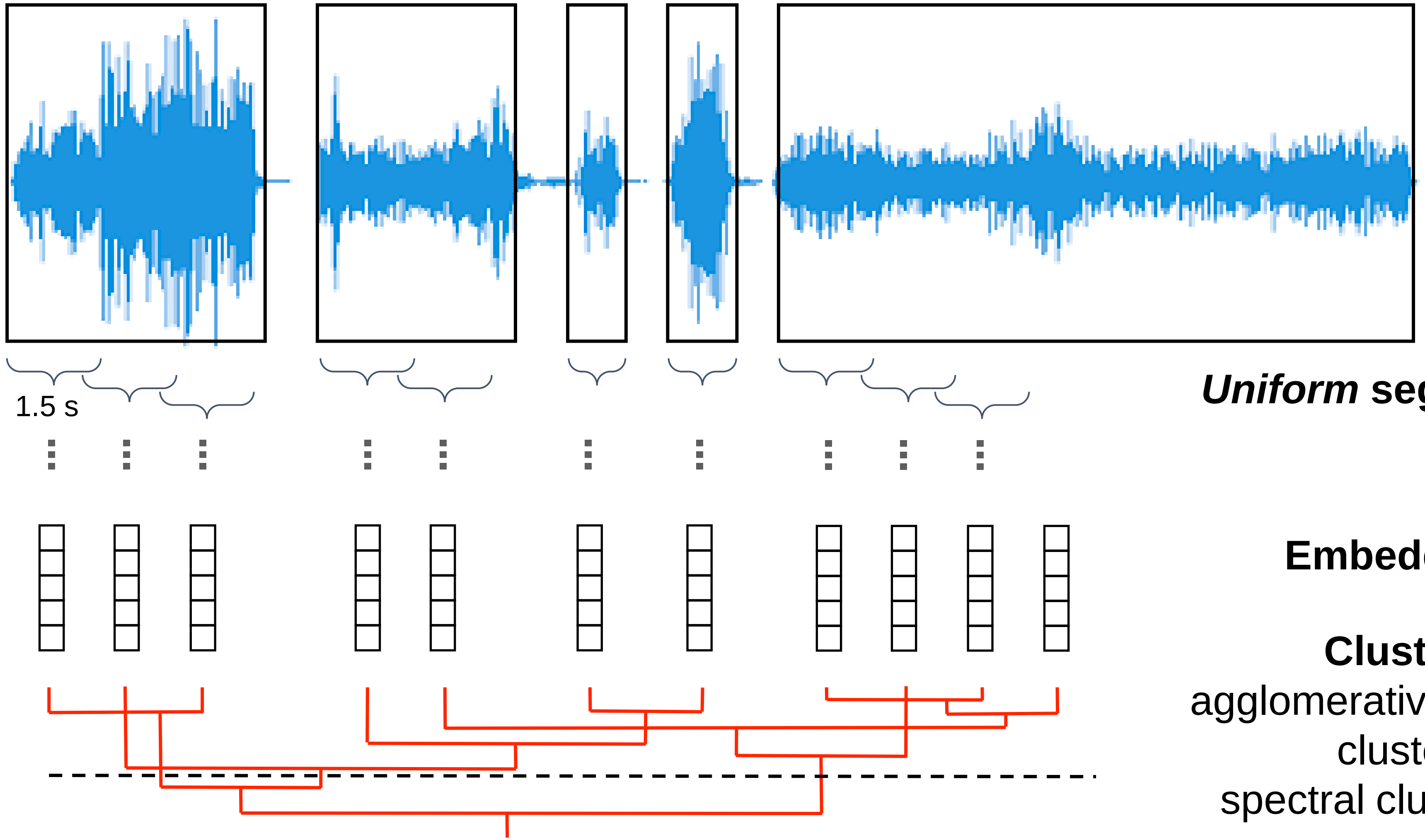
Embedding extraction

- Trained for **speaker recognition** task
    - On single speaker utterances
- **x-vectors** are the low-dimensionality embeddings representing the speaker characteristics of the input utterance
- Different **architectures** for the extraction of per-frame embedding extraction: TDNN, ResNet, etc.
- Several **objectives** can be used AAM loss, Softmax loss, etc.



Diarization system

Standard / Traditional / Cascade / Module-based



***Uniform* segmentation**

Embeddings

Clustering:
agglomerative hierarchical clustering,
spectral clustering, etc.

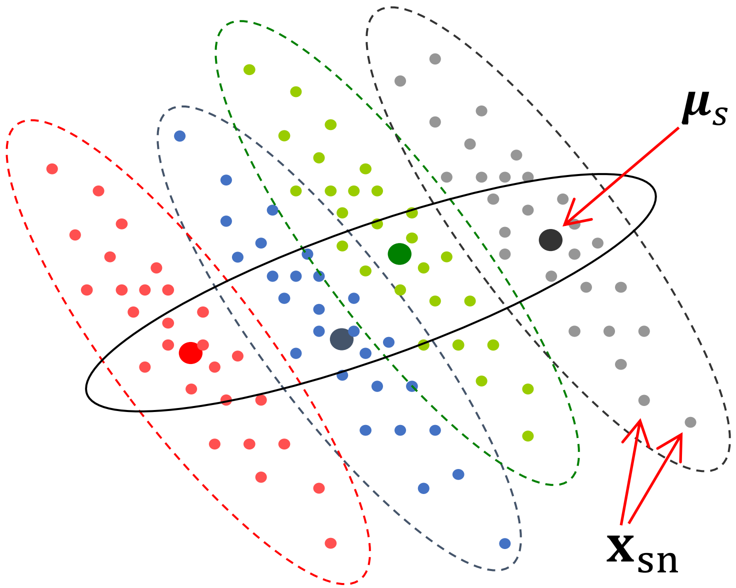
VBx Diarization

- Clustering embeddings using VB inference for Bayesian HMM
  - Federico Landini, Ján Profant, Mireia Diez, Lukáš Burget, “**Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks**”, Computer Speech & Language, Volume 71, 2022, 101254, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2021.101254.
  - Here we consider simplified version using only GMM, which is anyway used in practice
- Uses PLDA trained on x-vectors to model:

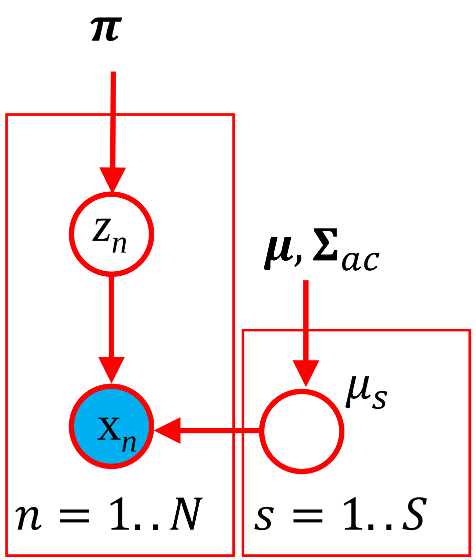
$p(\boldsymbol{\mu}_s) = \mathcal{N}(\boldsymbol{\mu}_s | \boldsymbol{\mu}, \boldsymbol{\Sigma}_{ac})$ - distribution of speaker means

$p(\mathbf{x} | \boldsymbol{\mu}_s) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_{wc})$ - within speaker (channel) variability

- If the x-vectors from a single conversation of several speakers follows the PLDA model, they can be assumed to be distributed according to Bayesian GMM model where:
  - Means of the components follows prior distribution $p(\boldsymbol{\mu}_s)$
  - Speaker specific distributions are $(\mathbf{x} | \boldsymbol{\mu}_s)$
  - Weights $\boldsymbol{\pi}$ determine how much is the speaker speaking



VBx Diarization



- We assume that the observed x-vectors in each conversation were generated as follows:
  - For each speaker $s = 1 \dots S$, mean of the speaker specific distribution was generated as
    - $\boldsymbol{\mu}_s \sim \mathcal{N}(\boldsymbol{\mu}_s | \boldsymbol{\mu}, \boldsymbol{\Sigma}_{ac})$
  - For each x-vector $n = 1 \dots N$
    - $z_n \sim \mathrm{P}(z_n | \boldsymbol{\pi}) = \mathrm{Cat}(z_n | \boldsymbol{\pi})$
    - $\mathbf{x}_n \sim p(\mathbf{x}_n | z_n, \{\boldsymbol{\mu}_s\}) = \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{wc}\right)$
- Given the "observed" x-vector sequence $\mathbf{x} = [x_1, x_2, \dots, x_N]$, the task is to infer (the distribution over) $\mathbf{z} = [z_1, z_2, \dots, z_N]$, which defines assignment of x-vectors (speech frames) to Gaussian components (speaker clusters).
- Variational Bayes inference is used for this purpose as shown before for BHM
- Component weights $\boldsymbol{\pi}$ are not treated as latent variables but learned using as point estimates to maximize ELBO.
  - The weights of the "redundant" components converge to 0 ⇒ It automatically determines the number of speaker in the conversation