Optimizing Bayesian HMM Based x-vector Clustering for the Second DIHARD Speech Diarization Challenge

Mireia Diez[1], Lukáš Burget[1], Federico Landini[1], Shuai Wang[1,2], Honza Černocký[1]

[1]Brno University of Technology, Faculty of Information Technology, IT4I Centre of Excellence, Czechia
[2]Speechlab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China

mireia@fit.vutbr.cz, burget@fit.vutbr.cz

ICASSP 2020

Introduction

This work was developed in the context of **the Second DIHARD Diarization Challenge**[1]

This presentation will cover the core of the system for **Track 1**: single-channel diarization following DIHARD I format

The system consists of an **x-vector extractor**, which provides x-vectors every 0.25s, which are then **clustered by a Bayesian HMM with eigenvoice priors**.

More details on the whole system description for track 1 and on systems for all other tracks can be found in *BUT System for the Second DIHARD Speech Diarization Challenge, F.Landini et.al.*

[1] N. Ryant et al. "The Second DIHARD Diarization Challenge: Dataset, task, and baselines.". 2019.

Bayesian HMM with Eigenvoice priors for Speaker Diarization

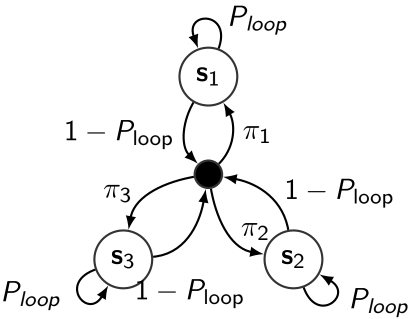
An efficient **Variational Bayes (VB) inference** in **a single probabilistic model addresses the complete Speaker Diarization problem**.

- A single model is used to infer:
  - The assignment of frames to speakers
  - Number of speakers
  - Speaker specific models

Bayesian HMM with Eigenvoice priors for Speaker Diarization II

Our model is a **Bayesian Hidden Markov Model**

- **States** model speaker specific distributions
- **Transitions** between states represent speaker turns



Bayesian HMM with Eigenvoice priors for Speaker Diarization III

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ be the sequence of observed x-vectors.

States modeled by PLDA-like model

$$p(\mathbf{x}_t|\mathbf{y}_s) = \mathcal{N}(\mathbf{x}_t; \mathbf{m}_s, \mathbf{\Sigma}_{wc}), \tag{1}$$

$$\mathbf{m}_s = \mathbf{m} + \mathbf{V}\mathbf{y}_s, \tag{2}$$

$$p(\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s; \mathbf{0}, \mathbf{I}) \tag{3}$$

Same model and inference as our original Bayesian HMM with Eigenvoice priors[2], but with a single Gaussian per state and $\mathbf{V}$, $\mathbf{m}$ and $\mathbf{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^T$ initialized from the PLDA model pretrained on large amount of x-vectors[3]

[2] M. Diez et al. "Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors". 2019.

[3] M. Diez et al. "Bayesian HMM based x-vector clustering for Speaker Diarization". 2019.

Problem & approach

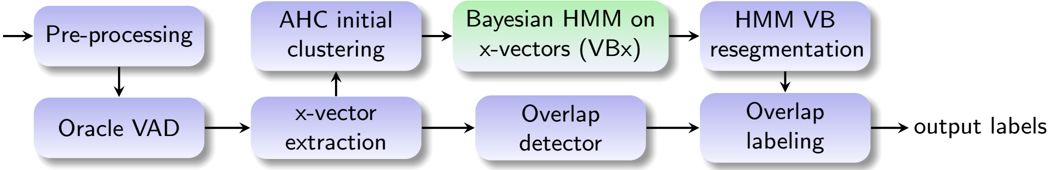
$\mathbf{Z} = \{z_1, z_2, ..., z_T\}$ is the sequence of latent discrete assignments of observations (x-vectors) to HMM states (speakers)

- We seek for the assignment of observations to speakers $p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}, \mathbf{Y}|\mathbf{X})d\mathbf{Y}$
- Variational Bayes with mean-field approximation $p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) \sim q(\mathbf{Z}) \prod_s q(\mathbf{y_s})$[4]

[4] M. Diez et al. "Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors". 2019.

System description

Diarization:



- Weighted Prediction Error (WPE) is used to de-reverberate the speech signal
- x-vectors are extracted from the input conversation using a 1.5s sliding window and a shift of 0.25s
- x-vectors are centered, whitened and length normalized
- x-vectors are pre-clustered using AHC
- x-vectors are clustered using the BHMM model
- A BHMM model is used at frame-level as re-segmentation step
- Overlapped speech is detected and post-processed to get two speaker labels

System description - x-vector extraction

- Time-delay neural network **TDNN**[5]
- Trained for speaker classification on VoxCeleb training and VoxCeleb2 development data with data augmentation: 6 million utterances from 7146 speakers
- **Utterances are cut** into 2s segments for the neural network training
- 64-dimensional **Fbanks** are used as input features, using an energy-based voice activity detector (VAD) to remove silence
- For test, **512 dimensional x-vectors** are extracted from the penultimate layer every 0.25s from (up to) 1.5s segments
- x-vectors are **centered and whitened** using statistics estimated from DIHARD development and evaluation data, and then **length normalized**

[5] D. Snyder et al. "Deep Neural Network Embeddings for Text-Independent Speaker Verification". 2017.

System description - PLDA models

- **Out-of-domain** PLDA model is trained using VoxCeleb training set
- **In-domain** PLDA model is trained on the limited DIHARD dev set
- Both models are estimated from **centered, whitened and length-normalized x-vectors** extracted from **3s segments**
- **Domain adaptation** strategy: Interpolation of the two PLDA models

System description - AHC

- **x-vectors** are extracted for 1.5s windows with 0.25s overlap
- **Conversation dependent PCA**, x-vectors (and also PLDA model) projected so as to keep only a **30%** of the total variability
- The projected x-vectors are once more length-normalized
- **PLDA** based pairwise similarity measure
- AHC stopping **threshold** fine-tuned on the development set

System description - BHMM clustering of x-vectors

- Uses the **same PLDA models** as the ones trained for the **AHC**
- **BHMM initialized** from the AHC diarization output (AHC set to undercluster)
- **Input features** are x-vectors extracted every 0.25s
- Parameters analyzed:
  - Acoustic scaling factor F_A, counteracts the assumption of statistical independence between observations by scaling down the log likelihood of the observations
  - Loop probability P_{loop}

System description - Frame-level BHMM re-segmentation

- **19 MFCC + E + Δ** features, extracted from 16kHz speech.
- Neither mean nor variance normalization are applied
- Gender-independent **UBM-GMM** with 1024 diagonal-covariance Gaussian components
- The **dimensionality** of the speaker specific i-vector-like latent variable $\mathbf{y}_s$, is 400
- UBM-GMMs and total variability matrix **trained using VoxCeleb2** dataset
- A **single iteration** of this frame-level BHMM is applied

Evaluation data and metric

The **DIHARD II dataset** is the evaluation set

- Created for the second DIHARD challenge
- Includes utterances coming from several sources (YouTube, court rooms, meetings, etc.)
- The corpus consists of 192 development and 194 evaluation recordings, containing around 18h and 17h of speech,

The system is evaluated in terms of the **Diarization Error Rate (DER)** as defined by NIST, with the format established for track 1 of the second DIHARD challenge

- We use the oracle speech activity labels
- No collar used for the evaluation
- Overlap speech regions are evaluated

Results
AHC optimization

DER results attained with AHC using PLDA models trained on VoxCeleb (out-of-domain), DIHARD dev (in-domain) and when interpolating them

		PLDA trained on		
Set	VB reseg.	VoxCeleb	DIHARD dev	Interp.
Dev	No	20.46	20.55	19.74
	Yes	19.84	20.20	**19.21**
Eval	No	21.12	22.29	**20.96**
	Yes	20.11	21.48	**19.97**

Results

BHMM Optimization

DER for different clustering methods and thresholds

		Threshold	
Set	method	Optimal for AHC	Under-clustered
Dev	AHC	20.46	(33.55)
	BHMM	19.33	**18.34**
Eval	AHC	21.12	(33.31)
	BHMM	19.90	**19.14**

Results

BHMM Optimization

DER for different x-vectors extracting frame rates

			Frame rate	
Set	F_A	P_{loop}	0.75s	0.25s
Dev	1.0	0.0	**19.55**	23.20
	0.4	0.8	20.13	**18.34**
Eval	1.0	0.0	**20.29**	22.89
	0.4	0.8	22.74	**19.14**

Results

BHMM Optimization

DER results attained with BHMM using PLDA models trained on VoxCeleb (out-of-domain), DIHARD dev (in-domain) and when interpolating them

		PLDA trained on		
Set	VB reseg.	Voxceleb	DIHARD dev	Interp.
Dev	No	18.34	17.87	17.90
	Yes	18.35	18.16	18.23
Eval	No	19.14	18.83	**18.39**
	Yes	18.95	18.80	**18.38**

Summary

This x-vector level BHMM is **the core of our winning system** on track 1 of the second DIHARD speech diarization challenge, obtaining 18.42% DER

- Performance gains
  - Improved x-vector extractor
  - increasing the frame-rate for x-vector extraction
  - using x-vector level BHMM diarization with PLDA model interpolation for "domain adaptation"
- Compared to last year's approach, the described system improves performance by close to an absolute 7% DER
- Around half of the remaining error (9% DER) corresponds to overlapped speech error
- **Open source recipe** including feature extraction, initial AHC and VBx `https://github.com/BUTSpeechFIT/VBx`

BRNO UNIVERSITY OF TECHNOLOGY speech@fit