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

Mireia Diez\textsuperscript{1}, Lukáš Burget\textsuperscript{1}, Federico Landini\textsuperscript{1}, Shuai Wang\textsuperscript{1,2}, Honza Černocký\textsuperscript{1}

\textsuperscript{1}Brno University of Technology, Faculty of Information Technology, IT4I Centre of Excellence, Czechia
\textsuperscript{2}Speechlab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China

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

ICASSP 2020
1. Introduction

2. Bayesian HMM with Eigenvoice priors for Speaker Diarization

3. System description

4. Results

5. Summary
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.

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
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 $X = \{x_1, x_2, \ldots, x_T\}$ be the sequence of observed $x$-vectors.

States modeled by PLDA-like model

$$p(x_t|y_s) = \mathcal{N}(x_t; m_s, \Sigma_{wc}),$$

$$m_s = m + V y_s,$$

$$p(y_s) = \mathcal{N}(y_s; 0, I)$$

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


Problem & approach

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

- We seek for the assignment of observations to speakers \( p(Z|X) = \int p(Z, Y|X) dY \)
- Variational Bayes with mean-field approximation \( p(Z, Y|X) \sim q(Z) \prod_s q(y_s) \)

\footnote{M. Diez et al. “Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors” 2019.}
System description

Diarization:

- Pre-processing
- Oracle VAD
- AHC initial clustering
- x-vector extraction
- Bayesian HMM on x-vectors (VBx)
- Overlap detector
- HMM VB resegmentation
- Overlap labeling
- output labels

- 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**

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 $P_{\text{loop}}$
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 $y_s$, 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
DER results attained with AHC using PLDA models trained on VoxCeleb (out-of-domain), DIHARD dev (in-domain) and when interpolating them.

**Results**

**AHC optimization**

<table>
<thead>
<tr>
<th>Set</th>
<th>VB reseg.</th>
<th>VoxCeleb</th>
<th>DIHARD dev</th>
<th>Interp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>No</td>
<td>20.46</td>
<td>20.55</td>
<td>19.74</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>19.84</td>
<td>20.20</td>
<td>19.21</td>
</tr>
<tr>
<td>Eval</td>
<td>No</td>
<td>21.12</td>
<td>22.29</td>
<td>20.96</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>20.11</td>
<td>21.48</td>
<td>19.97</td>
</tr>
</tbody>
</table>
## Results

### BHMM Optimization

**DER for different clustering methods and thresholds**

<table>
<thead>
<tr>
<th>Set</th>
<th>method</th>
<th>Optimal for AHC</th>
<th>Under-clustered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>AHC</td>
<td>20.46</td>
<td>(33.55)</td>
</tr>
<tr>
<td></td>
<td>BHMM</td>
<td>19.33</td>
<td><strong>18.34</strong></td>
</tr>
<tr>
<td>Eval</td>
<td>AHC</td>
<td>21.12</td>
<td>(33.31)</td>
</tr>
<tr>
<td></td>
<td>BHMM</td>
<td>19.90</td>
<td><strong>19.14</strong></td>
</tr>
</tbody>
</table>
**Results**

BHMM Optimization

DER for different x-vectors extracting frame rates

<table>
<thead>
<tr>
<th>Set</th>
<th>$F_A$</th>
<th>$P_{\text{loop}}$</th>
<th>Frame rate 0.75s</th>
<th>Frame rate 0.25s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>1.0</td>
<td>0.0</td>
<td><strong>19.55</strong></td>
<td>23.20</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.8</td>
<td>20.13</td>
<td><strong>18.34</strong></td>
</tr>
<tr>
<td>Eval</td>
<td>1.0</td>
<td>0.0</td>
<td><strong>20.29</strong></td>
<td>22.89</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.8</td>
<td>22.74</td>
<td><strong>19.14</strong></td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th>Set</th>
<th>VB reseg.</th>
<th>PLDA trained on</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Voxceleb</td>
<td>DIHARD dev</td>
<td>Interp.</td>
<td></td>
</tr>
<tr>
<td>Dev</td>
<td>No</td>
<td>18.34</td>
<td>17.87</td>
<td>17.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>18.35</td>
<td>18.16</td>
<td>18.23</td>
<td></td>
</tr>
<tr>
<td>Eval</td>
<td>No</td>
<td>19.14</td>
<td>18.83</td>
<td><strong>18.39</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>18.95</td>
<td>18.80</td>
<td><strong>18.38</strong></td>
<td></td>
</tr>
</tbody>
</table>
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