BUT System for the Second DIHARD Speech Diarization Challenge

Federico Landini\textsuperscript{1}, Shuai Wang\textsuperscript{1,2}, Mireia Diez\textsuperscript{1}, Lukáš Burget\textsuperscript{1}, Pavel Matějka\textsuperscript{1}, Kateřina Žmolková\textsuperscript{1}, Ladislav Mošner\textsuperscript{1}, Anna Silnova\textsuperscript{1}, Oldřich Plchot\textsuperscript{1}, Ondřej Novotný\textsuperscript{1}, Hossein Zeinali\textsuperscript{1}, Johan Rohdin\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

landini@fit.vutbr.cz, mireia@fit.vutbr.cz

ICASSP 2020
| 1  | Challenge and Datasets                          |
| 2  | Systems Overview                                 |
| 3  | Track 1                                          |
| 4  | Track 2                                          |
| 5  | Tracks 3 and 4                                   |
| 6  | Summary                                          |
## Challenge and Datasets

- **Second DIHARD Challenge**: diarization in hard conditions

<table>
<thead>
<tr>
<th>Challenge and Datasets</th>
<th>Systems Overview</th>
<th>Track 1</th>
<th>Track 2</th>
<th>Tracks 3 and 4</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenge and Datasets

- Second DIHARD Challenge: diarization in hard conditions
Challenge and Datasets

- Second DIHARD Challenge: diarization in hard conditions

Datasets
- Track 1: DIHARD II with oracle VAD
- Track 2: DIHARD II with system VAD
- Track 3: CHiME-5 with oracle VAD
- Track 4: CHiME-5 with system VAD
Challenge and Datasets

- Second DIHARD Challenge: diarization in hard conditions

Datasets
- Track 1: DIHARD II with oracle VAD
- Track 2: DIHARD II with system VAD
- Track 3: CHiME-5 with oracle VAD
- Track 4: CHiME-5 with system VAD

Our results allowed us to obtain the first position on all tracks
DIHARD II corpus

- **Single-channel data**
  - Recordings from different sources comprising audiobooks, child language, courtroom, meetings, restaurant conversations, interviews, web videos and more
  - Lasting between 5 to 10 minutes and accounting for around 2 hours per source
  - Amount of speakers per recording ranging from 1 to 10

- Development set with 23:49 hours and evaluation set with 22:29 hours
- Systems evaluated in terms of the Diarization Error Rate (DER)
- No collar used for the evaluation and overlapped speech regions are evaluated
CHiME-5 corpus

- Multi-channel data from the CHiME-5 dinner party corpus
  - conversational speech collected in dinner parties at homes with 4 participants
  - lasting between 2 to 3 hours and held in three locations: kitchen, dining, living
- Each session collected with 6 microphone arrays
- Each array evaluated individually
- Three sets: train, development and evaluation
  - with 16, 2 and 2 sessions respectively
  - with 40:33, 4:27 and 5:12 hours respectively
- Systems evaluated in terms of the Diarization Error Rate (DER)
- No collar used for the evaluation and overlapped speech regions are evaluated
<table>
<thead>
<tr>
<th>Challenge and Datasets</th>
<th>Systems Overview</th>
<th>Track 1</th>
<th>Track 2</th>
<th>Tracks 3 and 4</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Track 1**
- Pre-processing
- Oracle VAD
- x-vectors extraction
- AHC initial clustering
- Bayesian HMM on x-vectors (VBx)
- HMM VB resegmentation
- Overlap detector
- Overlap labeling
- Output labels

**Track 2**
- Pre-processing
- DNN VAD
- x-vectors extraction
- AHC initial clustering
- Bayesian HMM on x-vectors (VBx)
- HMM VB resegmentation
- Output labels
**Track 1**

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

**output labels**
**Track 1**

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

**Track 2**

- Pre-processing
  - DNN VAD
  - AHC initial clustering
  - x-vectors extraction
  - Bayesian HMM on x-vectors (VBx)
  - HMM VB resegmentation
  - Output labels
<table>
<thead>
<tr>
<th>Challenge and Datasets</th>
<th>Systems Overview</th>
<th>Track 1</th>
<th>Track 2</th>
<th>Tracks 3 and 4</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Tracks 3 and 4
Tracks 3 and 4

- Pre-processing
- Oracle/NN-based VAD
- x-vectors extraction
- Multi-channel processing
- AHC clustering
- Output labels
Track 1

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

Challenge and Datasets
Systems Overview
Track 1
Track 2
Tracks 3 and 4
Summary

Federico Landini, Shuai Wang, Mireia Diez, Lukáš Burget et. al.
BUT System for the Second DIHARD Speech Diarization Challenge
Track 1

We explored four approaches for pre-processing

- Denoising provided by organizers ¹
- Denoising based on Wave-U-Net ²
- Denoising based on neural network autoencoders ³
- Dereverberating with weighted prediction error (WPE) ⁴

The best performing one was WPE

---

¹ https://github.com/staplesinLA/denoising_DIHARD18
² C. Macartney and T. Weyde, Improved speech enhancement with the wave-u-net
³ O. Plchot et al., Audio Enhancing with DNN Autoencoder for Speaker Recognition
⁴ T. Nakatani et al., Speech dereverberation based on variance-normalized delayed linear prediction, and L. Drude et al., NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing
For Track 1 the oracle voice activity detection labels are used
**Track 1**

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

- **x-vectors**: DNN based speaker embeddings\(^5\)

- Extractor trained on VoxCeleb 1 and 2 with augmentations with some tweaks with respect to Kaldi SRE16 recipe\(^6\)

- **x-vectors** extracted on 1.5s windows every 0.25s\(^7\)
  - Instead of standard 1.5s windows every 0.75s

---

\(^5\) D. Snyder et al., *Deep Neural Network Embeddings for Text-Independent Speaker Verification*

\(^6\) More details in *BUT System Description for DIHARD Speech Diarization Challenge 2019*

\(^7\) Comparative analysis in *Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge*
Track 1

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

- Agglomerative hierarchical clustering with similarity matrix
Track 1

- Pre-processing
- AHC initial clustering
- Bayesian HMM on x-vectors (VBx)
- HMM VB resegmentation

Oracle VAD → x-vectors extraction → Overlap detector → Overlap labeling → output labels

- Agglomerative hierarchical clustering with similarity matrix
- Based on the interpolation of two PLDA models:
  1. trained on VoxCeleb segments
  2. trained on DIHARD II development segments
Track 1

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

**Agglomerative hierarchical clustering with similarity matrix**

Based on the interpolation of two PLDA models:
1. trained on VoxCeleb segments
2. trained on DIHARD II development segments

<table>
<thead>
<tr>
<th>PLDA model</th>
<th>DER</th>
<th>VoxCeleb</th>
<th>Interpolated</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>20.46</td>
<td>19.74</td>
<td></td>
</tr>
<tr>
<td>eval</td>
<td>21.12</td>
<td>20.96</td>
<td></td>
</tr>
</tbody>
</table>

- More analysis in *Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge*
Track 1

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

- States represent speaker specific distributions
- Transitions between states represent speaker turns
- Each speaker distribution is modeled by a Gaussian modeled using a PLDA like model
- The model infers the amount of speakers, the speaker models and assignment of frames to speakers
- More details in Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge
Track 1

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

<table>
<thead>
<tr>
<th>DER</th>
<th>PLDA model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VoxCeleb</td>
</tr>
<tr>
<td>AHC</td>
<td>dev</td>
</tr>
<tr>
<td></td>
<td>eval</td>
</tr>
<tr>
<td>VBx</td>
<td>dev</td>
</tr>
<tr>
<td></td>
<td>eval</td>
</tr>
</tbody>
</table>
Track 1

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

<table>
<thead>
<tr>
<th>DER</th>
<th>PLDA model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VoxCeleb</td>
</tr>
<tr>
<td>AHC</td>
<td>dev</td>
</tr>
<tr>
<td></td>
<td>eval</td>
</tr>
<tr>
<td>VBx</td>
<td>dev</td>
</tr>
<tr>
<td></td>
<td>eval</td>
</tr>
</tbody>
</table>
Track 1

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

VBx has a 0.25s resolution so we use VB resegmentation with MFCCs every 10ms
Track 1

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

- VBx has a 0.25s resolution so we use VB resegmentation with MFCCs every 10ms
- Same modeling as before in terms of states and transitions
Track 1

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

- VBx has a 0.25s resolution so we use VB resegmentation with MFCCs every 10ms
- Same modeling as before in terms of states and transitions
- Speaker distributions are modeled by an i-vector extractor like model (i.e GMM with parameters constrained by eigenvoice priors) trained on VoxCeleb
Track 1

- Pre-processing
  - Oracle VAD
  - x-vectors extraction
  - AHC initial clustering

- Bayesian HMM on x-vectors (VBx)
  - Overlap detector

- HMM VB resegmentation
  - Overlap labeling

- VBx has a 0.25s resolution so we use VB resegmentation with MFCCs every 10ms
- Same modeling as before in terms of states and transitions
- Speaker distributions are modeled by an i-vector extractor like model (i.e GMM with parameters constrained by eigenvoice priors) trained on VoxCeleb

<table>
<thead>
<tr>
<th>DER</th>
<th>VBx</th>
<th>+ resegmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>17.90</td>
<td>18.23</td>
</tr>
<tr>
<td>eval</td>
<td>18.39</td>
<td>18.38</td>
</tr>
</tbody>
</table>
Previous steps output one speaker per frame but there could be overlapped speech
Previous steps output one speaker per frame but there could be overlapped speech.

We used a logistic regression classifier to determine if x-vectors correspond to overlapped speech or not.
Previous steps output one speaker per frame but there could be overlapped speech.

We used a logistic regression classifier to determine if x-vectors correspond to overlapped speech or not.

Then, a heuristic assigns each frame in an overlapped speech segment to the two closest speakers (in time) according to the diarization labels from the previous step.
Previous steps output one speaker per frame but there could be overlapped speech.

We used a logistic regression classifier to determine if x-vectors correspond to overlapped speech or not.

Then, a heuristic assigns each frame in an overlapped speech segment to the two closest speakers (in time) according to the diarization labels from the previous step.

<table>
<thead>
<tr>
<th>DER</th>
<th>No ov. proc.</th>
<th>With ov. proc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>18.23</td>
<td>18.02</td>
</tr>
<tr>
<td>eval</td>
<td>18.38</td>
<td>18.21</td>
</tr>
<tr>
<td>Challenge and Datasets</td>
<td>Systems Overview</td>
<td>Track 1</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------</td>
<td>---------</td>
</tr>
</tbody>
</table>

## Track 1 recipe

- [https://github.com/BUTSpeechFIT/VBx](https://github.com/BUTSpeechFIT/VBx)
Track 1 recipe

- [https://github.com/BUTSpeechFIT/VBx](https://github.com/BUTSpeechFIT/VBx)

- Pre-processing
  - Oracle VAD
  - AHC initial clustering
  - x-vectors extraction

- Bayesian HMM on x-vectors (VBx) → output labels

- Only the most relevant modules are included
- Simplification in PLDA interpolation which improves results
Track 1 recipe

- https://github.com/BUTSpeechFIT/VBx

- Pre-processing
- Oracle VAD
- AHC initial clustering
- x-vectors extraction
- Bayesian HMM on x-vectors (VBx)
- output labels

- Only the most relevant modules are included
- Simplification in PLDA interpolation which improves results

<table>
<thead>
<tr>
<th>DER</th>
<th>No WPE</th>
<th>With WPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>17.87</td>
<td>17.64</td>
</tr>
<tr>
<td>eval</td>
<td>18.31</td>
<td>18.09</td>
</tr>
</tbody>
</table>
Track 2

- Pre-processing
- AHC initial clustering
- Bayesian HMM on x-vectors (VBx)
- HMM VB resegmentation
- Output labels

Pre-processing:
- DNN VAD

AHC initial clustering:
- x-vectors extraction

Bayesian HMM on x-vectors (VBx):
- HMM VB resegmentation

DER:
- Track 1: dev 18.23, eval 17 / 19
- Track 2: dev 23.81, eval 19 / 21

Federico Landini, Shuai Wang, Mireia Diez, Lukáš Burget et.al.
BUT System for the Second DIHARD Speech Diarization Challenge
Track 2

- Pre-processing
  - DNN VAD
- AHC initial clustering
  - x-vectors extraction
- Bayesian HMM on x-vectors (VBx)
- HMM VB resegmentation
  - output labels

- DNN-based VAD instead of oracle:
  - trained for binary, speech/non-speech, classification of 10ms speech frames
  - trained on the development set
- Slightly simpler pipeline: no overlap detection and PLDA trained on VoxCeleb
Track 2

- DNN-based VAD instead of oracle:
  - trained for binary, speech/non-speech, classification of 10ms speech frames
  - trained on the development set

- Slightly simpler pipeline: no overlap detection and PLDA trained on VoxCeleb

<table>
<thead>
<tr>
<th>DER</th>
<th>Track 1</th>
<th>Track 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>18.23</td>
<td>23.81</td>
</tr>
<tr>
<td>eval</td>
<td>18.38</td>
<td>27.11</td>
</tr>
</tbody>
</table>
DNN-based VAD instead of oracle:
- trained for binary, speech/non-speech, classification of 10ms speech frames
- trained on the development set

Slightly simpler pipeline: no overlap detection and PLDA trained on VoxCeleb

<table>
<thead>
<tr>
<th>DER</th>
<th>Track 1</th>
<th>Track 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>18.23</td>
<td>23.81</td>
</tr>
<tr>
<td>eval</td>
<td>18.38</td>
<td>27.11</td>
</tr>
</tbody>
</table>
Tracks 3 and 4

- Pre-processing
- Oracle/NN-based VAD
- x-vectors extraction
- Multi-channel processing
- AHC clustering

WPE method applied on recordings from all channels
NN-based VAD trained on Fisher English data for Track 4
Features: x-vectors computed on 1.5s windows every 0.75s
Average the similarity score matrices of all channels

Results:

<table>
<thead>
<tr>
<th>Track 3 Fusion</th>
<th>CH1</th>
<th>CH2</th>
<th>CH3</th>
<th>CH4</th>
<th>Fusion</th>
<th>dev+train</th>
<th>eval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>55.43</td>
<td>55.34</td>
<td>55.78</td>
<td>54.95</td>
<td>53.58</td>
<td>48.55</td>
<td>48.37</td>
</tr>
<tr>
<td>Fusion Track 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>eval</td>
<td>45.65</td>
<td>58.92</td>
</tr>
</tbody>
</table>
Tracks 3 and 4

- Pre-processing
- Oracle/NN-based VAD
- x-vectors extraction
- Multi-channel processing
- AHC clustering

- WPE method applied on recordings from all channels
Tracks 3 and 4

- WPE method applied on recordings from all channels
- NN-based VAD trained on Fisher English data for Track 4

**Detailed System Overview:**

1. **Pre-processing**
2. Oracle/NN-based VAD
3. x-vectors extraction
4. Multi-channel processing
5. AHC clustering

**Results:**

<table>
<thead>
<tr>
<th>Track 3</th>
<th>CH1</th>
<th>CH2</th>
<th>CH3</th>
<th>CH4</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev+train</td>
<td>55.43</td>
<td>55.34</td>
<td>55.78</td>
<td>54.95</td>
<td>53.58</td>
</tr>
<tr>
<td>eval</td>
<td>48.55</td>
<td>48.37</td>
<td>48.19</td>
<td>48.3</td>
<td>47.93</td>
</tr>
</tbody>
</table>

**Fusion Track 3 and Track 4**

| eval | 45.65 | 58.92 |
Tracks 3 and 4

- WPE method applied on recordings from all channels
- NN-based VAD trained on Fisher English data for Track 4
- Features: x-vectors computed on 1.5s windows every 0.75s
Tracks 3 and 4

- WPE method applied on recordings from all channels
- NN-based VAD trained on Fisher English data for Track 4
- Features: x-vectors computed on 1.5s windows every 0.75s
- Average the similarity score matrices of all channels
Tracks 3 and 4

- WPE method applied on recordings from all channels
- NN-based VAD trained on Fisher English data for Track 4
- Features: x-vectors computed on 1.5s windows every 0.75s
- Average the similarity score matrices of all channels
- Results:

<table>
<thead>
<tr>
<th>DER Track 3</th>
<th>CH1</th>
<th>CH2</th>
<th>CH3</th>
<th>CH4</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev+train</td>
<td>55.43</td>
<td>55.34</td>
<td>55.78</td>
<td>54.95</td>
<td>53.58</td>
</tr>
<tr>
<td>eval</td>
<td>48.55</td>
<td>48.37</td>
<td>48.19</td>
<td>48.3</td>
<td>47.93</td>
</tr>
</tbody>
</table>
Tracks 3 and 4

- WPE method applied on recordings from all channels
- NN-based VAD trained on Fisher English data for Track 4
- Features: x-vectors computed on 1.5s windows every 0.75s
- Average the similarity score matrices of all channels
- Results:

<table>
<thead>
<tr>
<th></th>
<th>CH1</th>
<th>CH2</th>
<th>CH3</th>
<th>CH4</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DER Track 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dev+train</td>
<td>55.43</td>
<td>55.34</td>
<td>55.78</td>
<td>54.95</td>
<td>53.58</td>
</tr>
<tr>
<td>eval</td>
<td>48.55</td>
<td>48.37</td>
<td>48.19</td>
<td>48.3</td>
<td>47.93</td>
</tr>
<tr>
<td><strong>DER</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eval</td>
<td>45.65</td>
<td></td>
<td></td>
<td>58.92</td>
<td></td>
</tr>
</tbody>
</table>
x-vectors have become the cornerstone for top-performing diarization systems

VBx allows for better performance than simple AHC
  - Even more when a better PLDA model is used to compare the x-vectors
  - Thus, adapting the PLDA model to in-domain data fosters performance

With the current performance on DIHARD II data, overlapped speech accounts for more than 50% of DER meaning this has to be addressed in the future

Recipe for Track 1: https://github.com/BUTSpeechFIT/VBx

CHiME presents a challenging scenario with considerable room for improvement
**Track 1**

1. **Pre-processing**
2. **Oracle VAD**
3. **AHC initial clustering**
4. **x-vectors extraction**
5. **Bayesian HMM on x-vectors (VBx)**
6. **Overlap detector**
7. **HMM VB resegmentation**
8. **Overlap labeling**
9. **output labels**

---

### Challenge and Datasets

**Systems Overview**
- **Track 1**
- **Track 2**
- **Tracks 3 and 4**

---

### Track 1

<table>
<thead>
<tr>
<th>DER</th>
<th>PLDA model</th>
<th>% files</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VoxCeleb</td>
<td>Interpolated</td>
</tr>
<tr>
<td>AHC</td>
<td>dev</td>
<td>20.46</td>
</tr>
<tr>
<td></td>
<td>eval</td>
<td>21.12</td>
</tr>
<tr>
<td>VBx</td>
<td>dev</td>
<td>18.34</td>
</tr>
<tr>
<td></td>
<td>eval</td>
<td>19.14</td>
</tr>
</tbody>
</table>

---

**BUT System for the Second DIHARD Speech Diarization Challenge**