Bayesian HMM clustering in speaker diarization

Mireia Diez

Brno University of Technology
Faculty of Information Technology Brno - Czechia
mireia@fit.vutbr.cz

January 23, 2021

Table of contents

Speaker Diarization based on Bayesian HMM
(VB diarization, BHMM diarization, VBx)

- Historical perspective
- Evolution of the algorithm
- Two flavours of this diarization method:

  Bayesian HMM with eigenvoice priors operating in frame-by-frame basis

  Winning system in DIHARD I - 2018 JHU

  Bayesian HMM based x-vector clustering

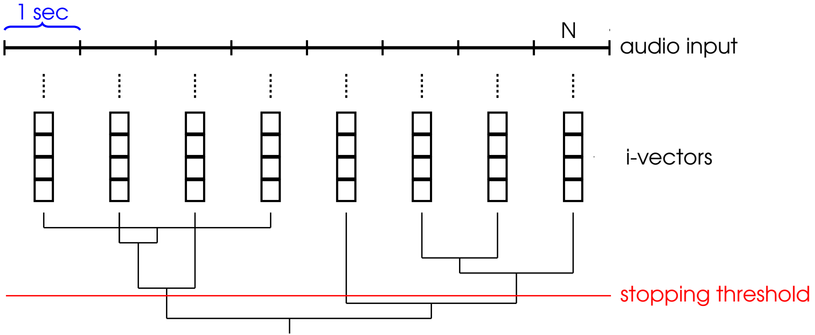
  Winning system in DIHARD II - 2019 BUT

- Things were made publicly available, it could serve as baseline:

	Core	Full
DIHARD III baseline	20.65	19.25
VBx Github recipe*	17.25	16.01

- Last journal paper, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks":
  VBx is used to set baselines in CALLHOME, AMI and DIHARDII

- Cut audio input utterances into short equal length segments
- extract i-vectors (speaker factors from JFA model)
- Cluster these low dimensional representations



This model operates on frame-by-frame MFCC features, although to avoid frequent speaker turns it allows speaker changes only every second

It models speaker specific distributions using i-vector like model: Speaker specific distributions are:

$$p(\mathbf{x}_t|\mathbf{y}_s) = \text{GMM}(\mathbf{x}_t; \{\boldsymbol{\mu}_{sc}\}, \{\boldsymbol{\Sigma}_c^{ubm}\}, \{w_c^{ubm}\})$$

All speaker specific GMMs share the w_c^{ubm} and the $\boldsymbol{\Sigma}_c^{ubm}$.

For a speaker s, the super-vector of concatenated component means $\boldsymbol{\mu}_s = [\boldsymbol{\mu}_{s1}^T \boldsymbol{\mu}_{s2}^T \ldots \boldsymbol{\mu}_{sC}^T]^T$ is constrained to live in a linear subspace:

$$\boldsymbol{\mu}_s = \boldsymbol{\mu}^{ubm} + \mathbf{V}\mathbf{y}_s$$

The low dimensional vectors $\mathbf{y}_s$ are treated as latent random variables with standard normal prior $p(\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s; \mathbf{0}, \mathbf{I})$.

Both models use the same informative prior on speaker distributions –pre-trained i-vector extractor model– but:

IV-Clust Extract i-vectors from short segments $\rightarrow$ noisy estimates

DFA Estimates speaker models on all the frames coming from a given speaker

---

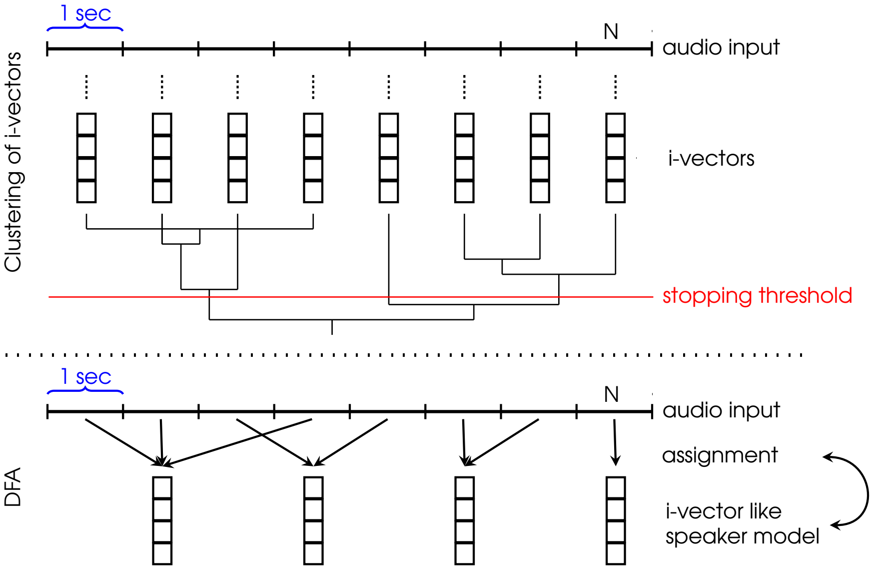
IV-clust Makes hard decisions

- i-vectors represent point estimates of "speaker models"
- Clustering of i-vectors making hard decisions cannot recover from prior errors

DFA No hard decisions in iterative VB inference:

- Re-estimates the speaker models, keeping the uncertainty of the models (posterior distributions)
- Re-estimates the soft-assignments of frames to the speaker models

---


Clustering of i-vectors
1 sec
N
audio input
i-vectors
stopping threshold
DFA
1 sec
N
audio input
assignment
i-vector like
speaker model

i-vector estimation:

$$\mathbf{i}_n = \mathbf{L}_n^{-1} \sum_{t=1}^{T_n} \sum_c \zeta_{tc} \left(\mathbf{x}_{nt} - \boldsymbol{\mu}_c^{ubm} \right)^T \boldsymbol{\Sigma}_c^{ubm^{-1}} \mathbf{V}_c$$

$$\mathbf{L}_n = \mathbf{I} + \sum_c \zeta_{tc} \mathbf{V}_c^T \boldsymbol{\Sigma}_c^{ubm^{-1}} \mathbf{V}_c$$

i-vector like update formulas:

$$q(\mathbf{y}_s) = \mathcal{N} \left(\mathbf{y}_s | \boldsymbol{\alpha}_s, \mathbf{L}_s^{-1} \right)$$

$$\boldsymbol{\alpha}_s = \mathbf{L}_s^{-1} \sum_t \gamma_{ts} \sum_c \zeta_{tc} \left(\mathbf{x}_t - \boldsymbol{\mu}_c^{ubm} \right)^T \boldsymbol{\Sigma}_c^{ubm^{-1}} \mathbf{V}_c$$

$$\mathbf{L}_s = \mathbf{I} + \sum_t \gamma_{ts} \sum_c \zeta_{tc} \mathbf{V}_c^T \boldsymbol{\Sigma}_c^{ubm^{-1}} \mathbf{V}_c$$

where $\zeta_{tc} = p_{ubm}(c|\mathbf{x}_t)$ and $\gamma_{ts} = ...$ are the probabilistic soft assignment of frames to speakers

Bayesian HMM with Eigenvoice priors

- Extension of the diarization using factor analysis
- Building it into a HMM to model the speaker turns and avoid pre-clustering into fixed length segments
- Implemented by Lukas Burget BUT in 2010, but not officially released

Bayesian HMM with Eigenvoice priors

Structure of the model

Our model is a Bayesian Hidden Markov Model

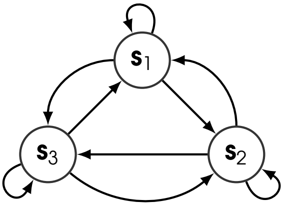
- States represent speaker specific distributions
- Transitions between states represent speaker turns
- Speaker distributions are modeled by GMMs with parameters constrained by eigenvoice priors (as in i-vector or JFA models)

$$p(\mathbf{x}_t|z_t = s) = p(\mathbf{x}_t|\mathbf{y}_s)$$
$$= \mathrm{GMM}(\mathbf{x}_t; \{\boldsymbol{\mu}_{sc}\}, \{\boldsymbol{\Sigma}_c^{ubm}\}, \{w_c^{ubm}\})$$

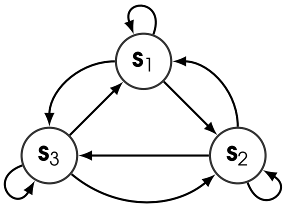
$$\boldsymbol{\mu}_s = \boldsymbol{\mu}^{ubm} + \mathbf{V}\mathbf{y}_s$$

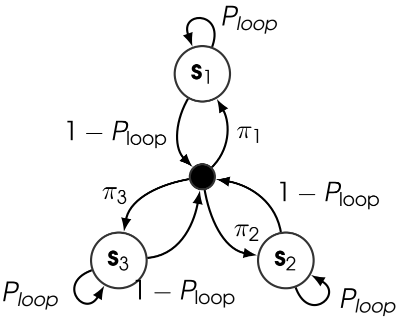
$$p(\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s; \mathbf{0}, \mathbf{I}).$$

$$p(z_t = s|z_{t-1} = s')$$



HMM model for 3 speakers with a single state per speaker, with a dummy non-emitting (initial) state.





$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ observed vectors (i.e. MFCC features)

$\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T\}$ alignment of frames to HMM states

$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_S\}$ set of all the speaker-specific latent variables

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) &= p(\mathbf{X}|\mathbf{Z}, \mathbf{Y}) p(\mathbf{Z}) p(\mathbf{Y}) \\ &= \prod_t p\left(\mathbf{x}_t | z_t\right) \prod_t p\left(z_t | z_{t-1}\right) \prod_s p\left(\mathbf{y}_s\right), \end{aligned} \tag{1}$$

To address the SD task, the speaker distributions and latent variables **Z** are jointly estimated given the input sequence **X**

The solution to the SD task is given by the most likely sequence **Z**:

$$p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) d\mathbf{Y}$$

We need to :

- Pre-estimate the i-vector extractor model: **V**, UBM
- Initialize the assignment of frames to speakers **Z**.
  It can be randomly, but initializing with another (good) system is better

VB iteratively updates:

- The speaker models $\mathbf{y}_s$ (same as in DFA model)

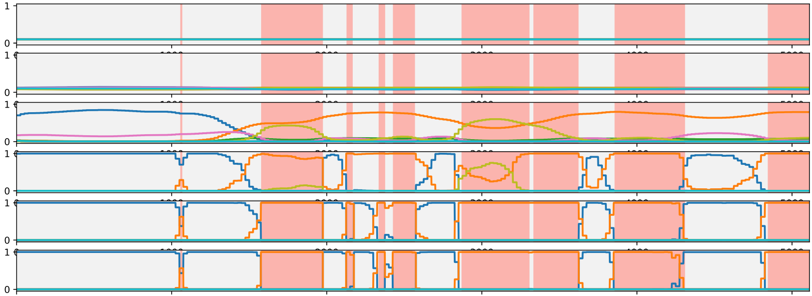
$$q(\mathbf{y}_s) = \mathcal{N}\left(\mathbf{y}_s | \boldsymbol{\alpha}_{\mathbf{s}}, \mathbf{L}_s^{-1}\right)$$

$$\boldsymbol{\alpha}_{\mathbf{s}} = \mathbf{L}_s^{-1} \sum_t \gamma_{ts} \sum_c \zeta_{tc} \left(\mathbf{x}_t - \boldsymbol{\mu}_c^{ubm}\right)^T \boldsymbol{\Sigma}_c^{ubm^{-1}} \mathbf{V}_c$$

$$\mathbf{L}_s = \mathbf{I} + \sum_t \gamma_{ts} \sum_c \zeta_{tc} \mathbf{V}_c^T \boldsymbol{\Sigma}_c^{ubm^{-1}} \mathbf{V}_c$$

- $\gamma_{ts} =$ ... using the forward-backward on HMM model
- The speaker priors π_s (number of speakers) – dropping redundant speakers by automatic relevance determination principle

Results
Convergence of the model



This approach addresses the complete SD problem by means a straightforward and efficient Variational Bayes (VB) inference in a single probabilistic model.
A single model will be used to infer:

- The assignment of frames to speakers
- Number of speakers
- Speaker specific models

Unlike other newer (End2End) approaches, the method:

- Requires input after VAD
- Does not deal with overlapped speech

2008 — Clustering of Speaker factors from JFA
Castaldo et.al (1)

2013 — Bayesian GMM clustering of i-vectors
Shum et.al. (4)

2014 — PLDA AHC i-vector clustering
Sell et.al. (5)

2008 — Clustering of Speaker factors from JFA
Castaldo et.al (1)

2008 — Diarization using Factor analysis
Kenny et.al. (2)

2013 — Bayesian GMM clustering of i-vectors
Shum et.al. (4)

2014 — PLDA AHC i-vector clustering
Sell et.al. (5)

- 2008 — Clustering of Speaker factors from JFA
  Castaldo et.al (1)
- 2008 — Diarization using Factor analysis
  Kenny et.al. (2)
- 2010 — Bayesian HMM software
  Burget et.al. (3)
- 2013 — Bayesian GMM clustering of i-vectors
  Shum et.al. (4)
- 2014 — PLDA AHC i-vector clustering
  Sell et.al. (5)

- 2008 — Clustering of Speaker factors from JFA
  Castaldo et.al (1)
- 2008 — Diarization using Factor analysis
  Kenny et.al. (2)
- 2010 — Bayesian HMM software
  Burget et.al. (3)
- 2013 — Bayesian GMM clustering of i-vectors
  Shum et.al. (4)
- 2014 — PLDA AHC i-vector clustering
  Sell et.al. (5)
- 2015 — Combine PLDA AHC i-vector clustering + Bayesian HMM diarization
  Sell et.al. (6)

Historical perspective
Timeline

2008	Clustering of Speaker factors from JFA Castaldo et.al (1)
2008	Diarization using Factor analysis Kenny et.al. (2)
2010	Bayesian HMM software Burget et.al. (3)
2013	Bayesian GMM clustering of i-vectors Shum et.al. (4)
2014	PLDA AHC i-vector clustering Sell et.al. (5)
2015	Combine PLDA AHC i-vector clustering + Bayesian HMM diarization Sell et.al. (6)
2018	First Publication describing Bayesian HMM diarization Diez et.al. (8)

2008 — Clustering of Speaker factors from JFA
Castaldo et.al (1)

2008 — Diarization using Factor analysis
Kenny et.al. (2)

2010 — Bayesian HMM software
Burget et.al. (3)

2013 — Bayesian GMM clustering of i-vectors
Shum et.al. (4)

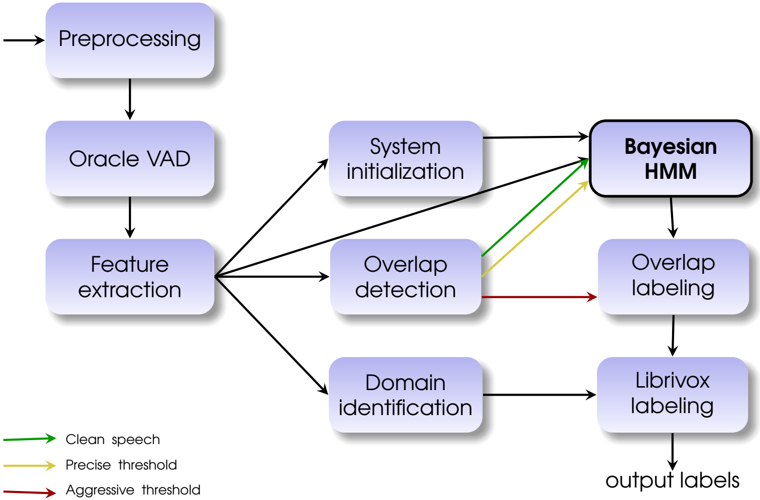
2014 — PLDA AHC i-vector clustering
Sell et.al. (5)

2015 — Combine PLDA AHC i-vector clustering + Bayesian HMM diarization
Sell et.al. (6)

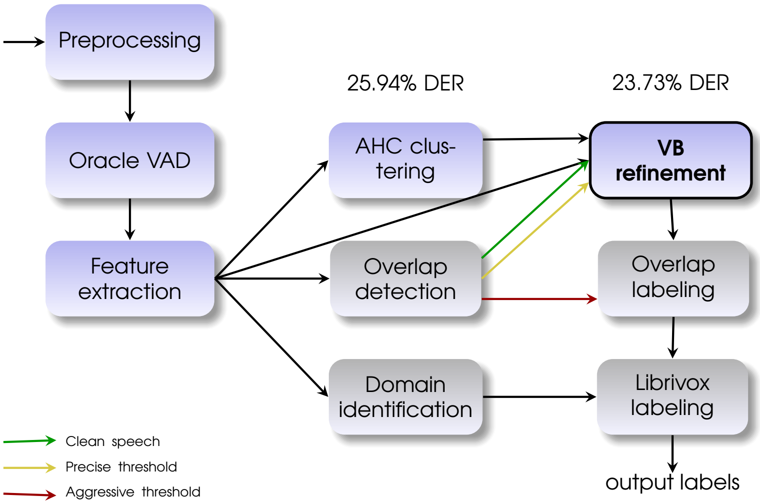
2016 — x-vectors
Snyder et.al. (7)

2018 — First Publication describing Bayesian HMM diarization
Diez et.al. (8)

BUT - 25.06% DER (Diez et.al. (9))



JHU - 23.73% DER (Sell et.al. (10))



New simplified version Alternative to AHC clustering of x-vectors
States modeled by PLDA like model

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

$$p(\mathbf{m}_s) = \mathcal{N}(\mathbf{m}_s; \mathbf{m}, \boldsymbol{\Sigma}_{ac})$$

or equivalently

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

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

where $\Sigma_{ac} = \mathbf{V}\mathbf{V}^T$
Same model and inference as the original Bayesian HMM with a single Gaussian per state and $\mathbf{V}$, $\mathbf{m}$ and $\boldsymbol{\Sigma}_{ac}$ initialized from the **PLDA** model pertained on large amount of x-vectors

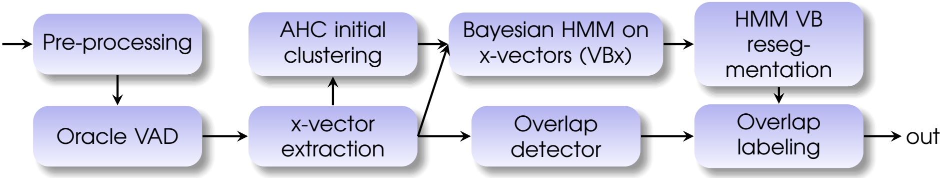
Parameters of the model:

- P_{loop} Loop probability to model speaker turns.
- *Acoustic scaling factor* F_A: introduced to counteract the assumption of statistical independence between observations.
- *Speaker regularization coefficient* F_B is a regularization term penalizing the complexity of the speaker models, high value of F_B results in the VB inference dropping more speakers.

  The ELBO can be split into three terms, where we scale the first two terms by the constant factors F_A and F_B.

$$\begin{aligned}\hat{\mathcal{L}}\left(q(\mathbf{X},\mathbf{Y})\right) &= F_A E_{q(\mathbf{Y},\mathbf{Z})}\left[\ln p(\mathbf{X}|\mathbf{Y},\mathbf{Z})\right] \\ &+ F_B E_{q(\mathbf{Y})}\left[\ln \frac{p(\mathbf{Y})}{q(\mathbf{Y})}\right] + E_{q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{Z})}{q(\mathbf{Z})}\right], \end{aligned} \tag{2}$$

BUT - 18.42% DER


Pre-processing
Oracle VAD
AHC initial clustering
x-vector extraction
Bayesian HMM on x-vectors (VBx)
Overlap detector
HMM VB reseg-mentation
Overlap labeling
out

Latest BHMM related publication

Bayesian HMM clustering of x-vector sequences (VBx)in speaker diarization: theory, implementation and analysis on standard tasks

- We provide the derivation and update formulas for the inference in the simpler VBx model
- We establish new baseline results in CALLHOME, AMI and DIHARDII datasets with VBx

How do people evaluate diarization systems?

- DER

$$DER = \frac{SER + FA + Miss}{Total\_speech}$$

  - *SER*: speaker error
  - *FA*: false alarm
  - *Miss*: missed speech
  - Total_speech (accounts also for speaker overlaps)
- Collar, yes/no?
- Overlapped speech, yes/no?
- JER

Callhome Results

Evaluation setup		System	SER	DER	JER
Collar	Overlap				
0.25	No	Kaldi (Sell et al. (10))	6.48		–
		Zhang et al. (11)	7.60		–
		Lin et al. (12)	6.63		–
		Pal et al. (13)	6.76		–
		Aronowitz et al. (14)	5.10		–
		AHC	8.10		–
		VBx	**4.42**		–
0.25	Yes	Horiguchi et al. (15)	–	15.29*	–
		AHC	7.53	17.64	–
		VBx	4.10	14.21	–
0	Yes	AHC	11.06	25.61	35.48
		VBx	7.22	21.77	34.02

How do people evaluate?

- DER Collar, yes/no?
- Overlapped speech, yes/no?

AMI dataset, much more chaos than that:

- Type of audio (Beamformed, Mix-headset, single mic)
- Partition (train/dev/test)
- References (different derivation of rttms from official AMI transcription files)

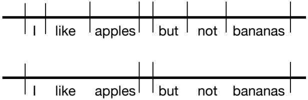
Partition	References	Audio type	Evaluation setup: Collar	Evaluation setup: Overlap	Scored speech dev/eval (s)	System	development SER	development DER	evaluation SER	evaluation DER
Pyannote	*Pyannote*	*Mix-Headset*	0.25	No	29200/29609	Bredin et al.(16)	–		4.6	
						VBx	2.14		2.17	
			0	Yes	54051/52317	Bredin et al. (16)	–	–	–	24.8
						Bullock et al. (17)	–	–	7.2	23.8
						VBx	3.33	22.95	3.86	22.85
Kaldi	*Force Aligned*	*Beamformed mic-array*	0.25	No	15053/14080	Sun et al. (18)	16.4		15.4	
						VBx	1.32		1.84	
			0.25	Yes	16241/14886	Sun et al. (18)	–	19.4	–	17.8
						VBx	1.26	4.96	1.92	4.67
	Kaldi	*Mix-Headset*	0.25	No	18743/18219	Maciejewski et al. (19)	–		– / (4.8*)	
						VBx	2.14		3.02/(2.58*)	
	Pyannote	*Mix-Headset*	0	Yes	35495/33953	Raj et al. (20)	–	–	10.1	23.6
						Raj et al. (21)	–	21.6	–	20.5
						VBx	3.12	22.63	3.56	23.47
		–	0	No	22812/21911	Raj et al. (21)	7.7		5.2	
						VBx	4.08		3.80	
Kaldi no TNO	*Work specific*	*Beamformed mic-array*	0.25	No	14545/13309	Pal et al. (13)	5.02 6.21		4.92 2.87	
						VBx	4.27		4.58	

New AMI Evaluation protocol

We consider AMI full-corpus ASR partition for train/dev/eval
References derived from manual annotations v 1.6.2.

- All words are considered as speech and included in the references
- Well defined, consistent and conservative approach in which all vocal sounds are discarded:
  - Very different sounds labeled as vocal-sounds
  - Not all vocal sounds were time-labeled
  - More consistent with the task of speaker-attributed ASR
- Adjacent speech segments (words) of the same speaker are merged not to create false "break" points.

starttime="0.86" endtime="0.98" word=I
starttime="0.98" endtime="1.1" word=like
starttime="1.1" endtime="1.40" word=apples
starttime="1.45" endtime="1.55" word=but
starttime="1.55" endtime="1.62" word=not
starttime="1.62" endtime="2.0" word=bananas



AMI II Results

<table>
<tr><th rowspan="2">Audio type</th><th colspan="2">Evaluation setup</th><th rowspan="2">System</th><th colspan="3">development</th><th colspan="3">evaluation</th></tr>
<tr><th>Collar</th><th>Overlap</th><th>SER</th><th>DER</th><th>JER</th><th>SER</th><th>DER</th><th>JER</th></tr>
<tr><td rowspan="6">Beamformed</td><td rowspan="2">0.25</td><td rowspan="2">No</td><td>AHC</td><td colspan="2">6.32</td><td>–</td><td colspan="2">7.65</td><td>–</td></tr>
<tr><td>VBx</td><td colspan="2">2.80</td><td>–</td><td colspan="2">3.90</td><td>–</td></tr>
<tr><td rowspan="2">0.25</td><td rowspan="2">Yes</td><td>AHC</td><td>6.43</td><td>14.68</td><td>–</td><td>8.82</td><td>18.36</td><td>–</td></tr>
<tr><td>VBx</td><td>2.57</td><td>10.81</td><td>–</td><td>4.69</td><td>14.23</td><td>–</td></tr>
<tr><td rowspan="2">0</td><td rowspan="2">Yes</td><td>AHC</td><td>8.68</td><td>22.14</td><td>25.29</td><td>10.93</td><td>25.48</td><td>29.85</td></tr>
<tr><td>VBx</td><td>4.20</td><td>17.66</td><td>22.26</td><td>6.28</td><td>20.84</td><td>26.92</td></tr>
<tr><td rowspan="6">Mix-Headset</td><td rowspan="2">0.25</td><td rowspan="2">No</td><td>AHC</td><td colspan="2">3.90</td><td>–</td><td colspan="2">3.96</td><td>–</td></tr>
<tr><td>VBx</td><td colspan="2">1.56</td><td>–</td><td colspan="2">2.10</td><td>–</td></tr>
<tr><td rowspan="2">0.25</td><td rowspan="2">Yes</td><td>AHC</td><td>4.06</td><td>12.31</td><td>–</td><td>5.05</td><td>14.60</td><td>–</td></tr>
<tr><td>VBx</td><td>1.43</td><td>9.68</td><td>–</td><td>2.98</td><td>12.53</td><td>–</td></tr>
<tr><td rowspan="2">0</td><td rowspan="2">Yes</td><td>AHC</td><td>6.16</td><td>19.61</td><td>23.90</td><td>6.87</td><td>21.43</td><td>25.50</td></tr>
<tr><td>VBx</td><td>2.88</td><td>16.33</td><td>20.57</td><td>4.43</td><td>18.99</td><td>24.57</td></tr>
</table>

DIHARD II Results

Evaluation setup Collar	Overlap	System	development SER	DER	JER	evaluation SER	DER	JER
0	Yes	Landini et al.* (22)	–	17.90 (18.34)	–	–	18.21 (19.14)	–
		Lin et.al. (23)	–	21.36	–	–	18.84	–
		Lin et.al. (24)	–	18.76	–	–	18.44 (19.46)	–
		AHC	10.89	21.68	42.28	13.89	23.59	43.93
		VBx	7.41	18.19	42.53	8.85	18.55	43.91
0.25	Yes	AHC	8.22	14.91	–	10.94	16.67	–
		VBx	5.53	12.23	–	6.55	12.29	–

Evaluation setup: Collar	Evaluation setup: Overlap	System	evaluation: Core	evaluation: Full
0	Yes	VBx baseline (DIHARD II system)	17.25	16.01
		BUT	15.46	13.29
		Winning system	13.45	11.30

Summary (I)

- Bayesian HMM systems have been consistently the best or among the best performing systems in the last years
- Despite their weaknesses:
  - No overlap handling
  - Need for external VAD
  - VBx: no frame precission
  - We didn't win DIHARDIII with it :(

  The method is still competitive, a strong baseline or and an important component for fusions
- We establish new baseline results in CALLHOME, AMI and DIHARDII datasets with VBx
  - We provide a new evaluation protocol for AMI dataset, which we hope can become the new standard

- All the code is made publicly available:
  - Recipe for training the x-vector extractors (8 kHz and 16 kHz)
  - Trained x-vector extractors
  - Pipeline for applying BHMM diarization
- Future:
  - Seek of ways of combining it with other embeddings
  - Combine the benefits of frame-by-frame BHMM and VBx
  - Overlapped speech handling
  - Combine it with E2E approaches

First description Bayesian HMM with eigenvoice priors: (8)

Related talk: *Here*

Full derivation of the method and inference and introduction of Fetch Factors: (25)

Open source code: Here

Full description of Bayesian HMM for x-vector clustering (VBx) and latest results for CALLHOME, AMI and DIHARDII: (26)

Open source code:
`https://github.com/BUTSpeechFIT/VBx`

Contact: mireia@fit.vutbr.cz

F. Castaldo et al.,
"Stream-based speaker segmentation using speaker factors and eigenvoices,"
in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 4133–4136.

P. Kenny,
"Bayesian Analysis of Speaker Diarization with Eigenvoice Priors,"
Tech. Rep., Montreal: CRIM, 2008.

Lukáš Burget,
"VB Diarization with Eigenvoice and HMM Priors,"
`http://speech.fit.vutbr.cz/software/vb-diarization-eigenvoice-and-hmm-priors`, 2013,
(Online; Jan.-2017).

S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass,
"Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach,"
IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

G. Sell and D. Garcia-Romero,
"Speaker Diarization with PLDA i-vector scoring and unsupervised calibration,"
in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 413–417.

G. Sell and D. Garcia-Romero,
"Diarization resegmentation in the factor analysis subspace,"
in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4794–4798.

D. et al. Snyder,
"Deep neural network-based speaker embeddings for end-to-end speaker verification,"
in *Spoken Language Technology Workshop (SLT), 2016 IEEE*, 12 2016, pp. 165–170.

M. Diez, L. Burget, and P. Matějka,
"Speaker Diarization based on Bayesian HMM with Eigenvoice Priors,"

in *Proceedings of Odyssey 2018, The speaker and Language Recognition Workshop*, 2018.

- Mireia Diez et al.,
  "BUT System for DIHARD Speech Diarization Challenge 2018,"
  in *Proc. Interspeech*, 2018, pp. 2798–2802.

- Gregory Sell et al.,
  "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge,"
  in *Proc. Interspeech*, 2018, pp. 2808–2812.

- Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang,
  "Fully supervised speaker diarization,"
  in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

- Qingjian Lin, Ruiqing Yin, Ming Li, Hervé Bredin, and Claude Barras,

"LSTM based similarity measurement with spectral clustering for speaker diarization,"
arXiv preprint arXiv:1907.10393, 2019.

Monisankha Pal, Manoj Kumar, Raghuveer Peri, Tae Jin Park, So Hyun Kim, Catherine Lord, Somer Bishop, and Shrikanth Narayanan,
"Meta-learning with Latent Space Clustering in Generative Adversarial Network for Speaker Diarization," 2020.

Hagai Aronowitz, Weizhong Zhu, Masayuki Suzuki, Gakuto Kurata, and Ron Hoory,
"New Advances in Speaker Diarization,"
in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*. 2020, pp. 279–283, ISCA.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu,
"End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," 2020.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

Latan Bullock, Herv Bredin, and Leibny Paola Garcia-Perera, "Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection," 2019.

Guangzhi Sun, Chao Zhang, and Phil Woodland, "Combination of Deep Speaker Embeddings for Diarisation," 2020.

M. Maciejewski, D. Snyder, V. Manohar, N. Dehak, and S. Khudanpur, "Characterizing Performance of Speaker Diarization Systems on Far-Field Speech Using Standard Methods," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5244–5248.

Desh Raj, Zili Huang, and Sanjeev Khudanpur,
"Multi-class Spectral Clustering with Overlaps for Speaker Diarization," 2020.

Desh Raj, Leibny Paola Garcia-Perera, Zili Huang, Shinji Watanabe, Daniel Povey, Andreas Stolcke, and Sanjeev Khudanpur,
"DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs," 2020.

Federico Landini, Shuai Wang, Mireia Diez, Lukáš Burget, Pavel Matějka, Kateřina Žmolíková, Ladislav Mošner, Anna Silnova, Oldřich Plchot, Ondřej Novotný, et al.,
"BUT System for the Second DIHARD Speech Diarization Challenge,"
in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6529–6533.

Qingjian Lin, Weicheng Cai, Lin Yang, Junjie Wang, Jun Zhang, and Ming Li,
"DIHARD II is Still Hard: Experimental Results and Discussions from the DKU-LENOVO Team,"
in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 102–109.

Qingjian Lin, Yu Hou, and Ming Li,
"Self-Attentive Similarity Measurement Strategies in Speaker Diarization,"
in *Proc. Interspeech 2020*, 2020, pp. 284–288.

M. Diez, L. Burget, F. Landini, and J. ernock,
"Analysis of Speaker Diarization Based on Bayesian HMM With Eigenvoice Priors,"
IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 355–368, 2020.

Federico Landini, Jn Profant, Mireia Diez, and Luk Burget,
"Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," 2020.