BUT System for the Second DIHARD Speech Diarization Challenge

Federico Landini[1], Shuai Wang[1,2], Mireia Diez[1], Lukáš Burget[1],
Pavel Matějka[1], Kateřina Žmolíková[1], Ladislav Mošner[1], Anna Silnova[1],
Oldřich Plchot[1], Ondřej Novotný[1], Hossein Zeinali[1], Johan Rohdin[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

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

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

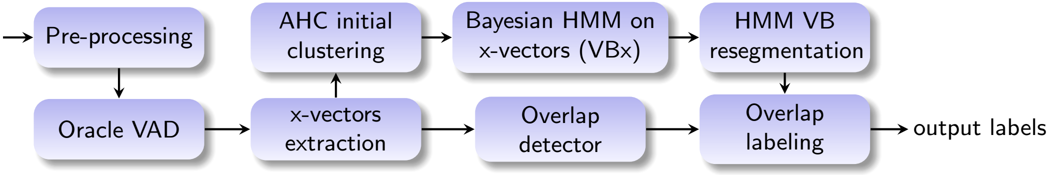
DIHARD II corpus

- 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

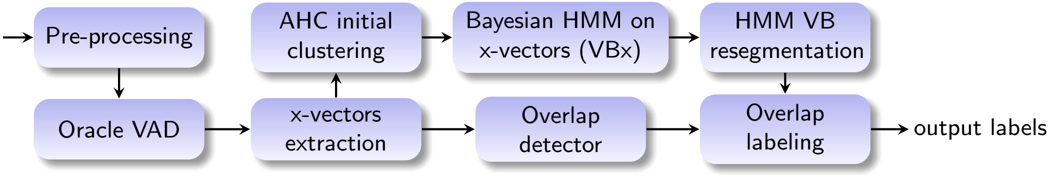
CHiME-5 corpus

- 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

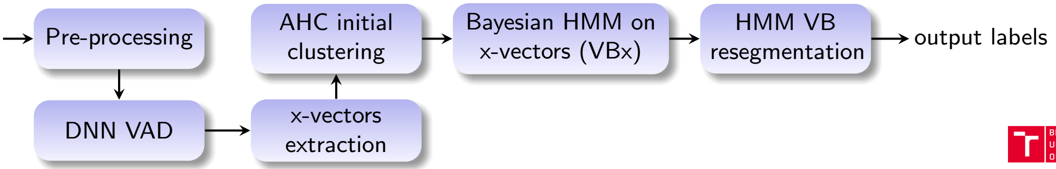
- Track 1


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

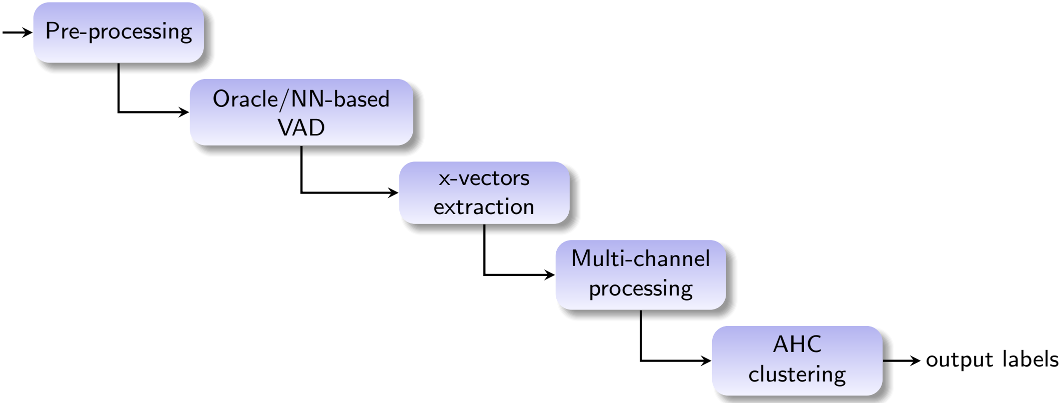


- Track 2

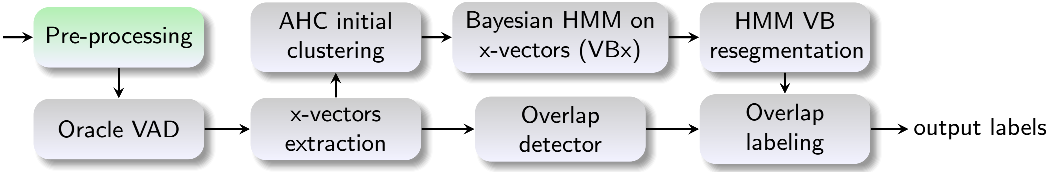


- Tracks 3 and 4

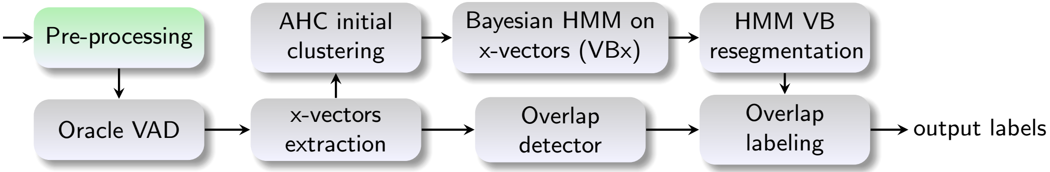
- Tracks 3 and 4



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

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

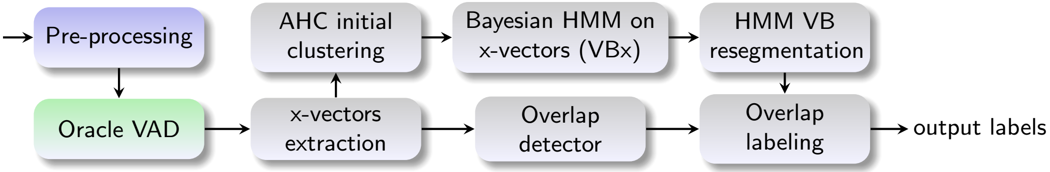
[1] https://github.com/staplesinLA/denoising_DIHARD18

[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*

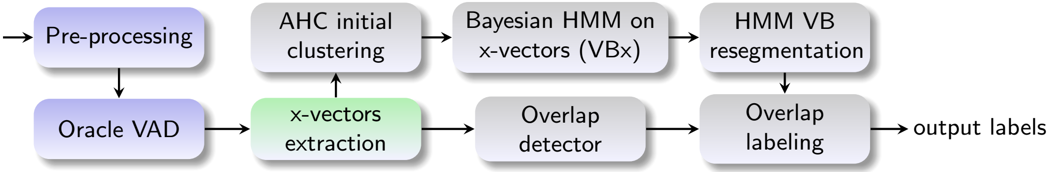
[4] T. Nakatani et al., *Speech dereverberation based on variance-normalized delayed linear prediction*, and L. Drude et al., *NARA-WPE: A Python package forweighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing*

Track 1



- For Track 1 the oracle voice activity detection labels are used

Track 1

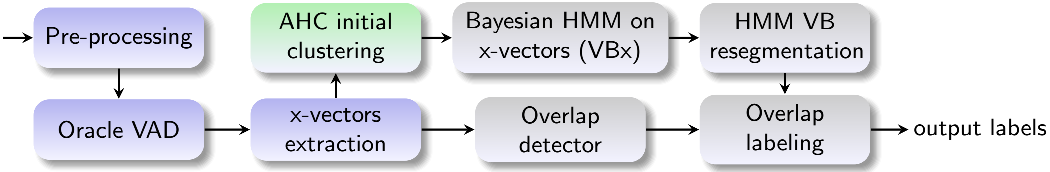


- x-vectors: DNN based speaker embeddings[5]
- Extractor trained on VoxCeleb 1 and 2 with augmentations with some tweaks with respect to Kaldi SRE16 recipe[6]
- x-vectors extracted on 1.5s windows every 0.25s[7]
  - Instead of standard 1.5s windows every 0.75s

---

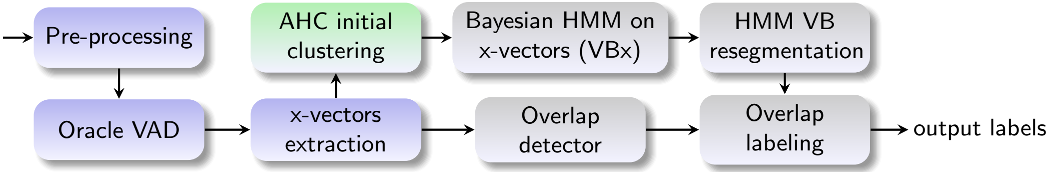
[5] D. Snyder et al., *Deep Neural Network Embeddings for Text-Independent Speaker Verification*
[6] More details in *BUT System Description for DIHARD Speech Diarization Challenge 2019*
[7] Comparative analysis in *Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge*

Track 1



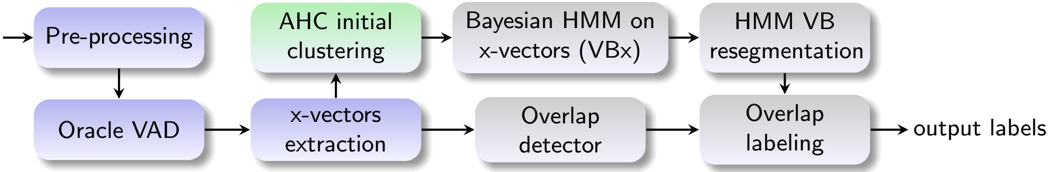
- Agglomerative hierarchical clustering with similarity matrix

Track 1



- 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

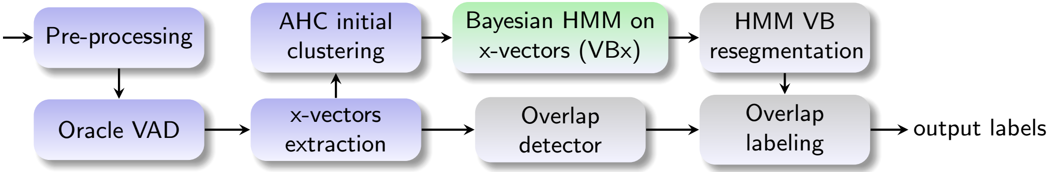


- Agglomerative hierarchical clustering with similarity matrix
  Based on the interpolation of two PLDA models:
  1. trained on VoxCeleb segments
  2. trained on DIHARD II development segments

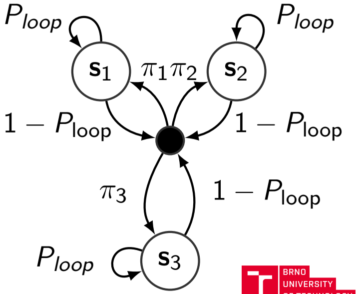
	PLDA model	
DER	VoxCeleb	Interpolated
dev	20.46	19.74
eval	21.12	20.96

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

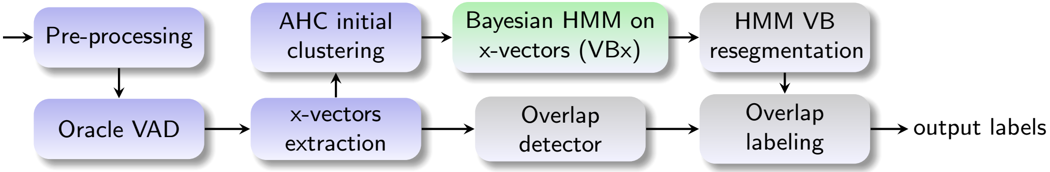
Track 1



- 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*

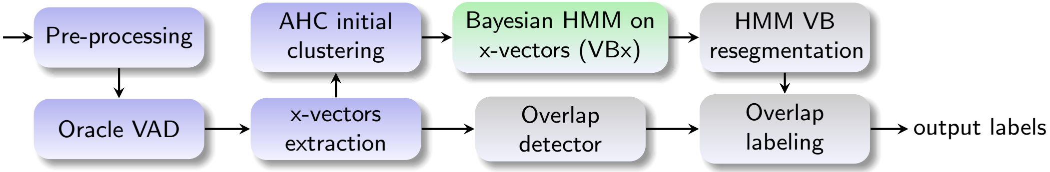


Track 1



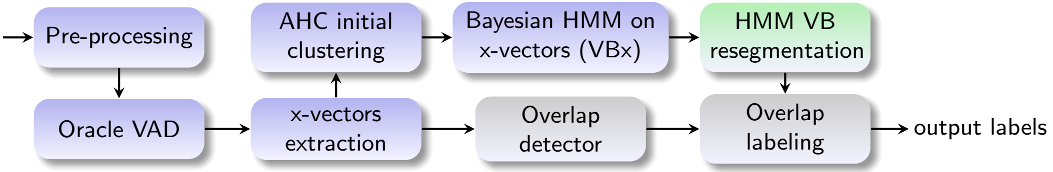
DER		PLDA model: VoxCeleb	PLDA model: Interpolated
AHC	dev	20.46	19.74
	eval	21.12	20.96
VBx	dev	18.34	17.90
	eval	19.14	18.39

Track 1



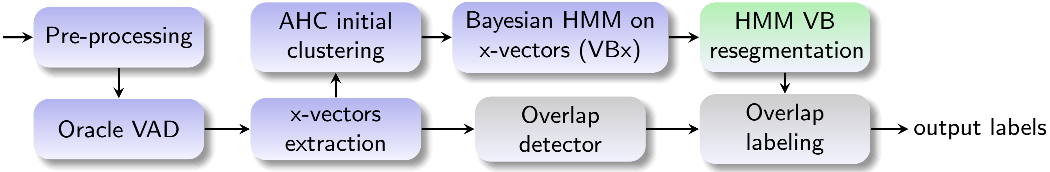
DER		PLDA model	
		VoxCeleb	Interpolated
AHC	dev	20.46	19.74
	eval	21.12	20.96
VBx	dev	18.34	17.90
	eval	19.14	18.39

Track 1



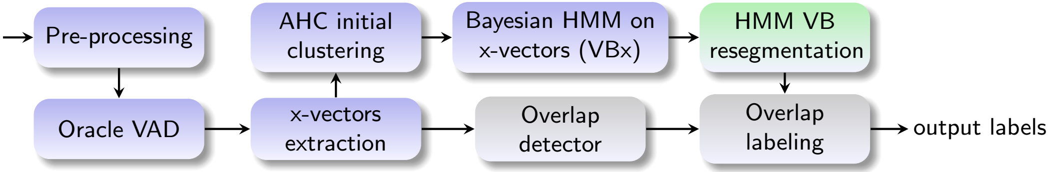
- VBx has a 0.25s resolution so we use VB resegmentation with MFCCs every 10ms

Track 1



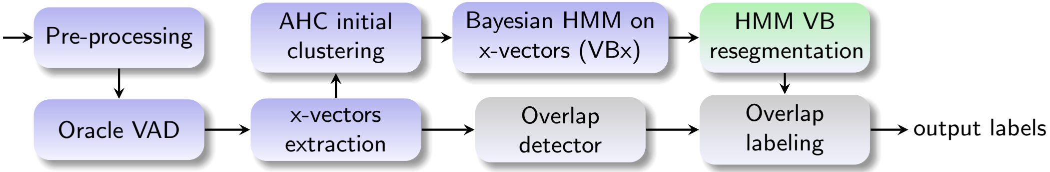
- VBx has a 0.25s resolution so we use VB resegmentation with MFCCs every 10ms
- Same modeling as before in terms of states and transitions

Track 1



- 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

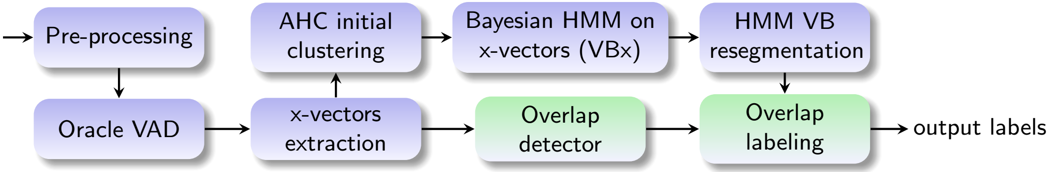
Track 1



- 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

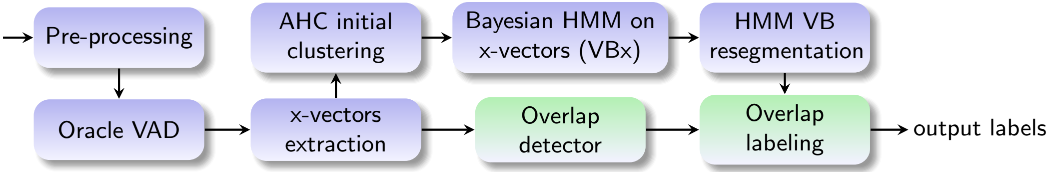
DER	VBx	+ resegmentation
dev	17.90	18.23
eval	18.39	18.38

Track 1



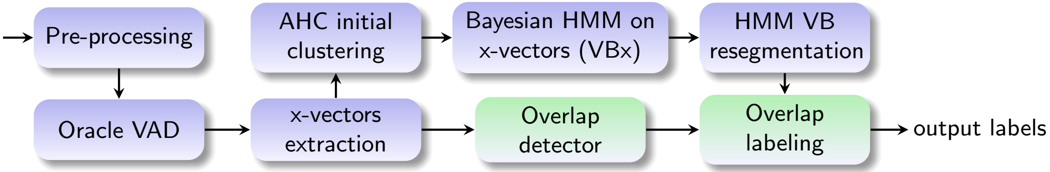
- Previous steps output one speaker per frame but there could be overlapped speech

Track 1



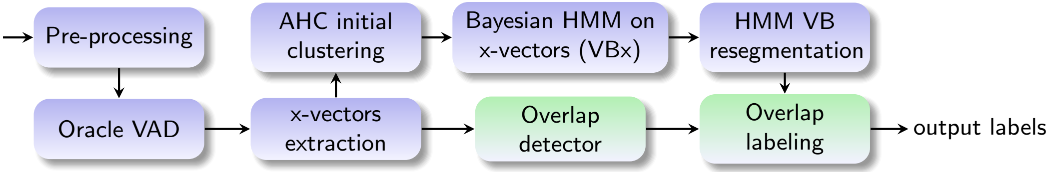
- Previous steps output one speaker per frame but there could be overlapped speech
- We used a logistic regression classifier to determine if x-vectors correspond to overlapped speech or not

Track 1



- 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

Track 1



- 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

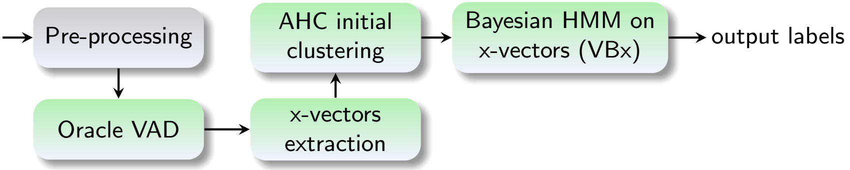
DER	No ov. proc.	With ov. proc.
dev	18.23	18.02
eval	18.38	18.21

Track 1 recipe

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

Track 1 recipe

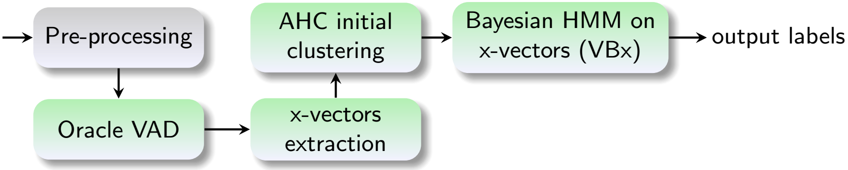
- `https://github.com/BUTSpeechFIT/VBx`



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

Track 1 recipe

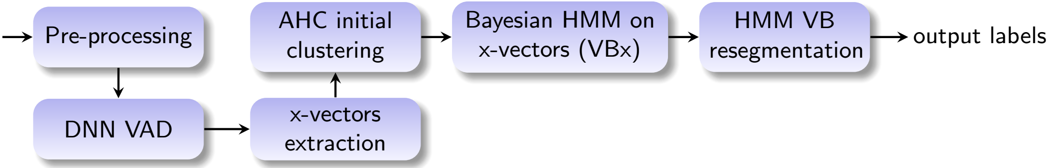
- `https://github.com/BUTSpeechFIT/VBx`



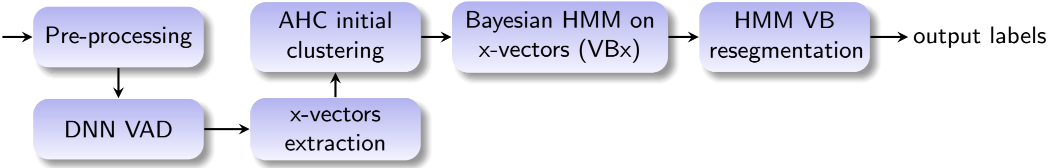
- Only the most relevant modules are included
- Simplification in PLDA interpolation which improves results

DER	No WPE	With WPE
dev	17.87	17.64
eval	18.31	18.09

Track 2

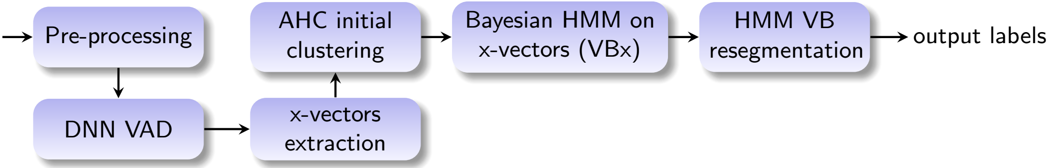

Pre-processing
DNN VAD
x-vectors extraction
AHC initial clustering
Bayesian HMM on x-vectors (VBx)
HMM VB resegmentation
output labels

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

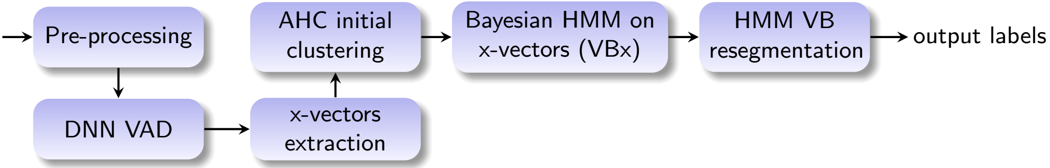
Track 2



- 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

DER	Track 1	Track 2
dev	18.23	23.81
eval	18.38	27.11

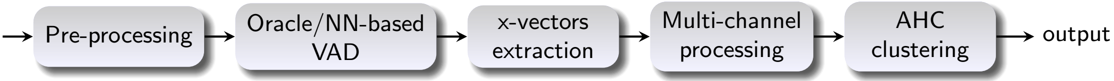
Track 2



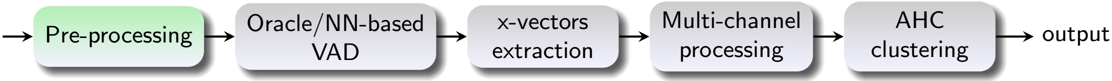
- DNN-based VAD instead of oracle:
  - trained for binary, speech/non-speech, classification of 10ms speech frames
  - trained on the development set
- Slightly simpler pipeline: no overlap detection and PLDA trained on VoxCeleb

DER	Track 1	Track 2
dev	18.23	23.81
eval	18.38	27.11

Tracks 3 and 4

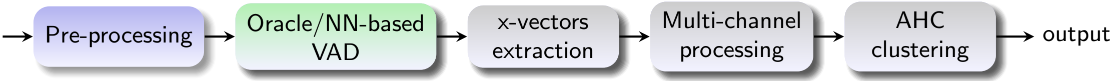


Tracks 3 and 4



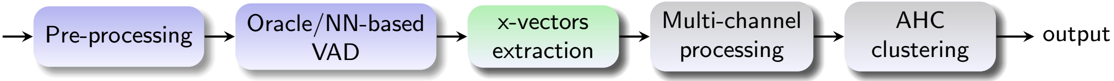
- WPE method applied on recordings from all channels

Tracks 3 and 4



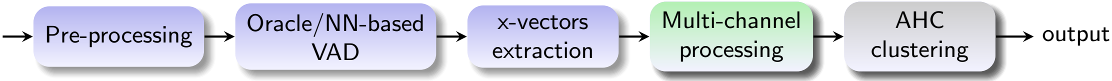
- WPE method applied on recordings from all channels
- NN-based VAD trained on Fisher English data for Track 4

Tracks 3 and 4



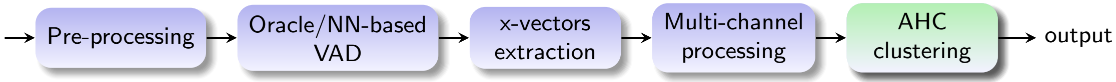
- WPE method applied on recordings from all channels
- NN-based VAD trained on Fisher English data for Track 4
- Features: x-vectors computed on 1.5s windows every 0.75s

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

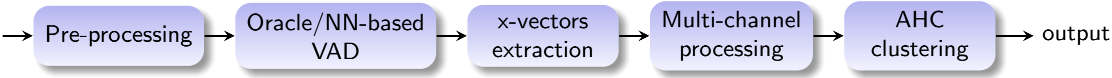
Tracks 3 and 4



- 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:

DER Track 3	CH1	CH2	CH3	CH4	Fusion
dev+train	55.43	55.34	55.78	54.95	53.58
eval	48.55	48.37	48.19	48.3	47.93

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:

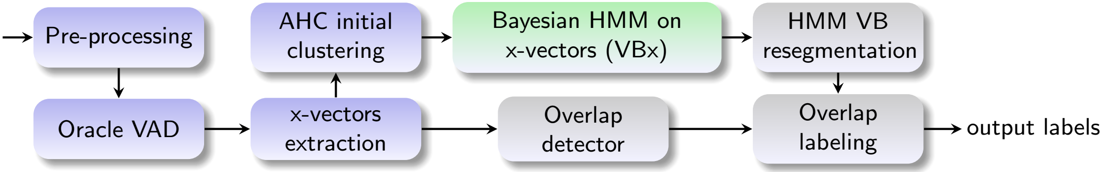
DER Track 3	CH1	CH2	CH3	CH4	Fusion
dev+train	55.43	55.34	55.78	54.95	53.58
eval	48.55	48.37	48.19	48.3	47.93

DER	Fusion Track 3	Fusion Track 4
eval	45.65	58.92

Summary

- 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



DER		PLDA model		% files	
		VoxCeleb	Interpolated	Same	Improved
AHC	dev	20.46	19.74	9%	59%
	eval	21.12	20.96	11%	45%
VBx	dev	18.34	17.90	14%	60%
	eval	19.14	18.39	22%	56%