BUT System for The Third DIHARD Speech Diarization Challenge

Federico Landini[1], Alicia Lozano-Diez[1], Lukáš Burget[1], Mireia Diez[1], Anna Silnova[1], Kateřina Žmolíková[1], Ondřej Glembek[1], Pavel Matějka[1], Themos Stafylakis[2], Niko Brümmer[2]

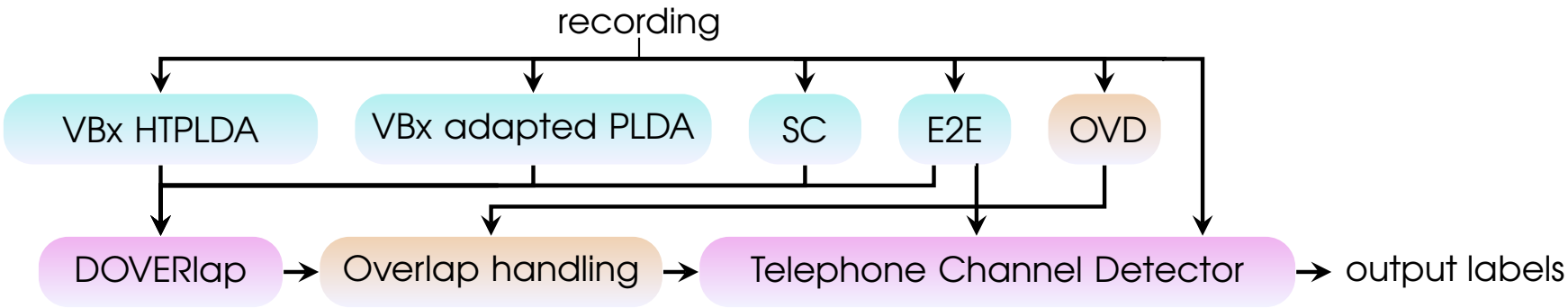
[1]Brno University of Technology, Faculty of Information Technology Brno - Czechia

[2]Omilia - Conversational Intelligence, Greece

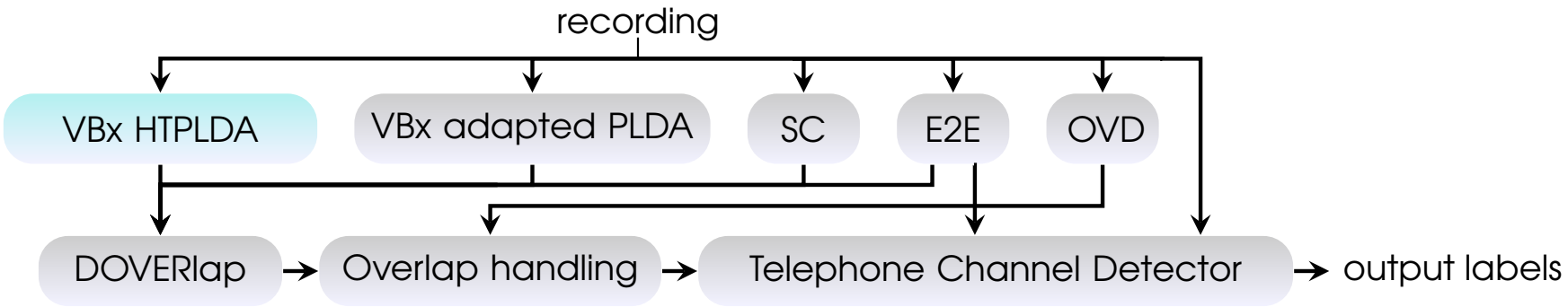
{landini, lozano, mireia}@fit.vutbr.cz

January 23, 2021

System Overview



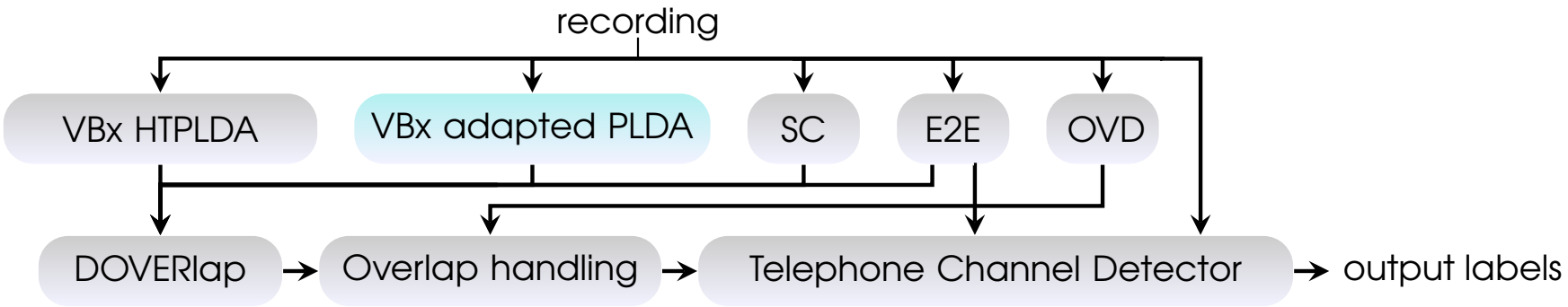
VBx heavy-tailed PLDA



- TDNN-based x-vectors clustered with AHC (initialization)
- Bayesian HMM that infers number of speakers, speaker models and assignment of x-vectors to speakers (VBx)
- Core of the winning system of DIHARD II [1]
- But state distributions derived from a heavy-tailed PLDA model instead of a Gaussian one

[1] Landini et al., BUT System for the Second DIHARD Speech Diarization Challenge
https://github.com/BUTSpeechFIT/VBx/tree/v1.0_DIHARDII

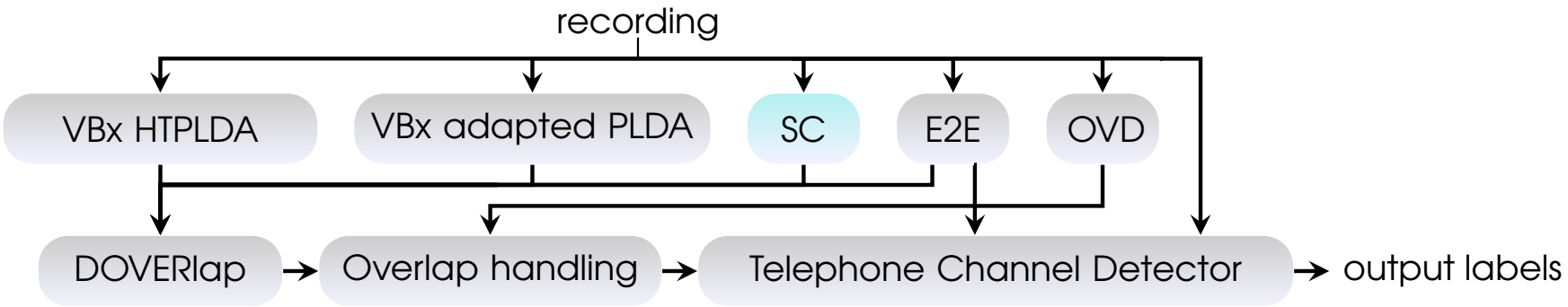
VBx adapted PLDA



- VBx using ResNet152-based x-vectors
- Core of BUT system for VoxConverse 2020 [2]
- But the PLDA model is an interpolation of
  - a PLDA trained on speakers from VoxCeleb
  - a PLDA trained on speakers from DIHARD 2020 dev set

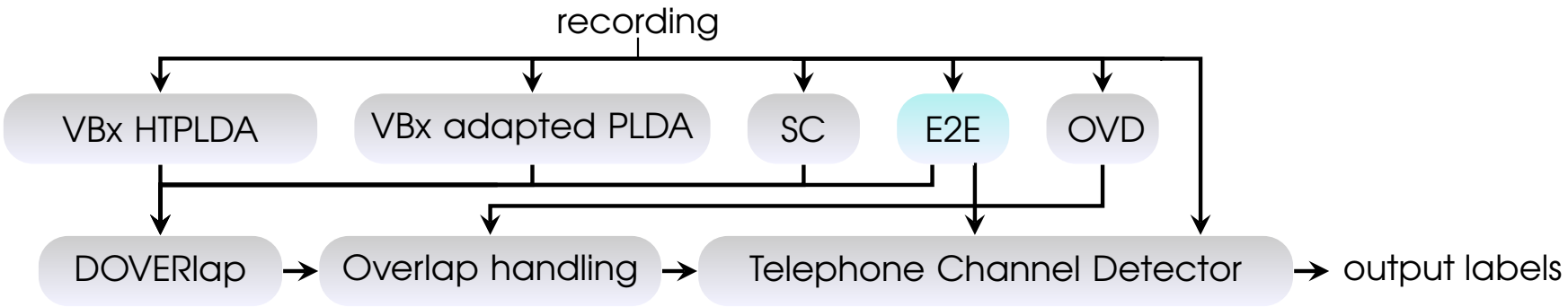
[2] Landini et al., Analysis of the BUT Diarization System for VoxConverse Challenge

Spectral Clustering



- ResNet152-based x-vectors clustered by means of spectral clustering + k-means
  - Affinity matrix based on cosine similarity
  - Only n largest elements are kept in each column/row of the affinity matrix
  - Number of speakers decided based on the largest eigen-gap

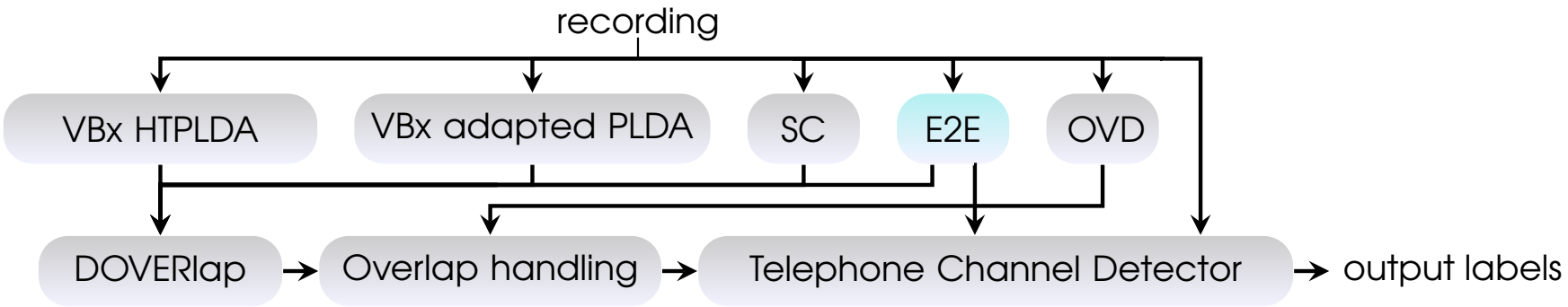
End-to-end Diarization



- Recordings downsampled to 8 kHz
- System based on self-attention and encoder-decoder LSTM-based attractors [3]
- Model trained on artificially created telephone conversations and fine-tuned to CALLHOME conversations

[3] Horiguchi et al., End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors

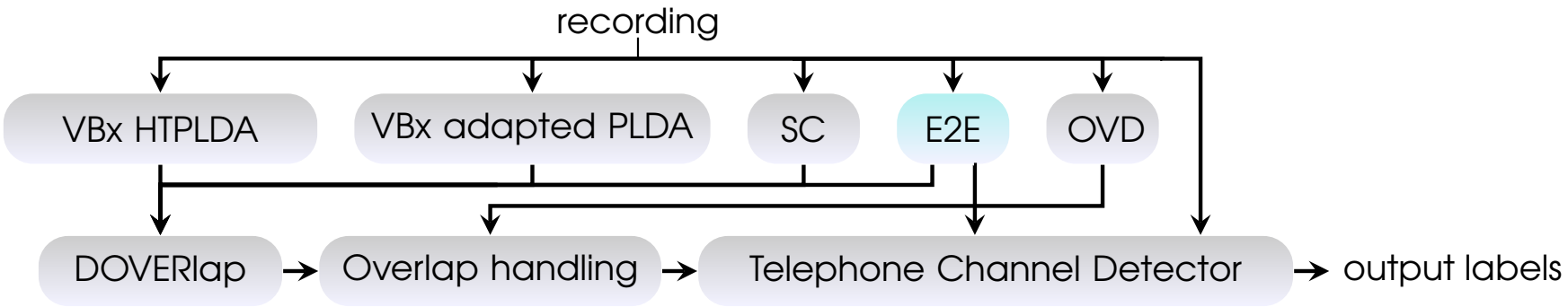
End-to-end Diarization



- Recordings downsampled to 8 kHz
- System based on self-attention and encoder-decoder LSTM-based attractors [3]
- Model trained on artificially created telephone conversations and fine-tuned to CALLHOME conversations
- By setting a threshold on the outputs, it is possible to predict silence and overlapped speech

[3] Horiguchi et al., End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors

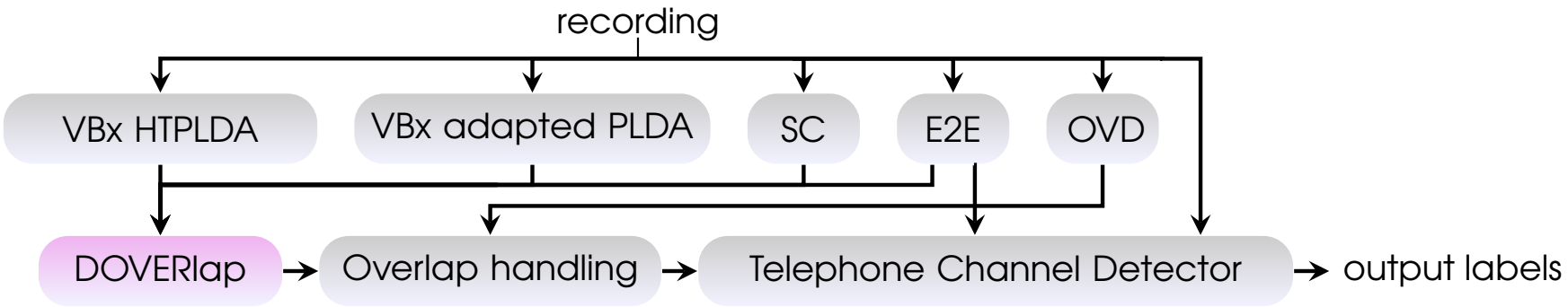
End-to-end Diarization



- Recordings downsampled to 8 kHz
- System based on self-attention and encoder-decoder LSTM-based attractors [3]
- Model trained on artificially created telephone conversations and fine-tuned to CALLHOME conversations
- By setting a threshold on the outputs, it is possible to predict silence and overlapped speech
  - Use oracle VAD for post-processing
  - Output always the most likely speaker
  - Tune threshold to find overlap (two or more speakers)

[3] Horiguchi et al., End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors

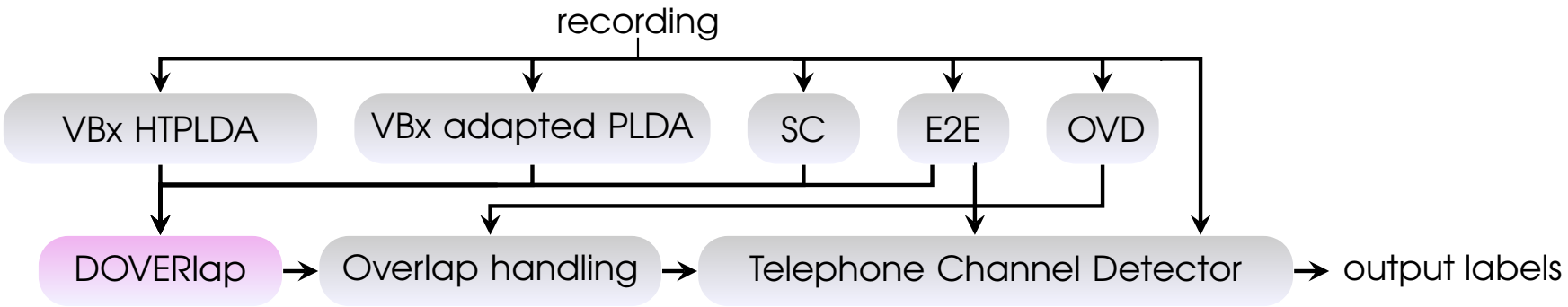
System Fusion



- The outputs of the four systems were fused using DOVERlap [4]
  - Speaker labels from different systems are globally mapped
  - Fusion labels are obtained with weighted majority voting
  - The voting scheme can handle overlapping labels

[4] Raj et al., DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs

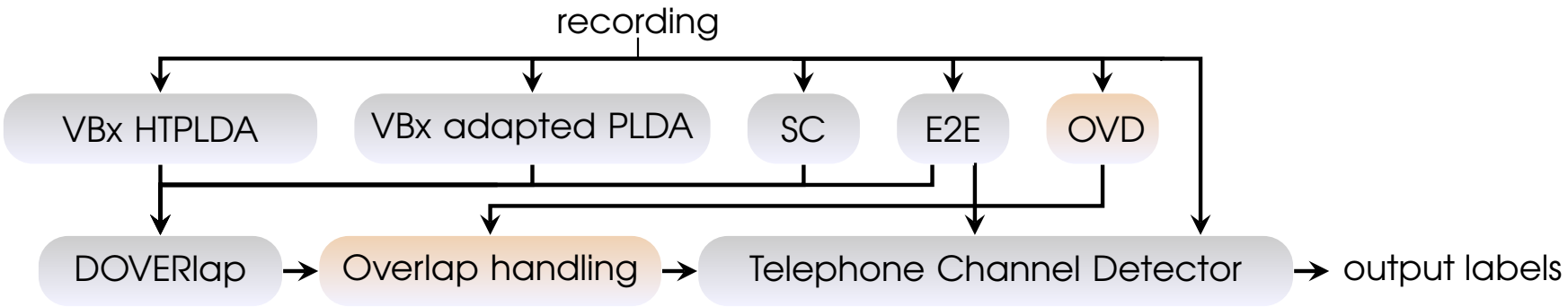
System Fusion



- The outputs of the four systems were fused using DOVERlap [4]
  - Speaker labels from different systems are globally mapped
  - Fusion labels are obtained with weighted majority voting
  - The voting scheme can handle overlapping labels
- However, only one of the systems accounts for overlapped speech

[4] Raj et al., DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs

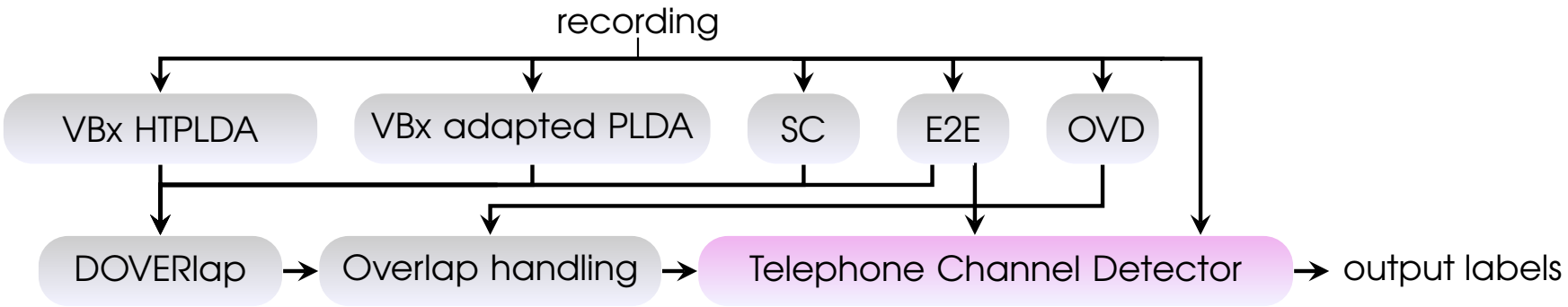
Overlapped Speech Handling



- Second speaker obtained using an heuristic: closest in time
- OVD uses the encoder and separator of Conv-TasNet [5]
- It was trained on DIHARD III dev set, VoxConverse dev set and three meeting datasets: ICSI, ISL and AMI train set
- Both real data and artificial overlaps were used for training

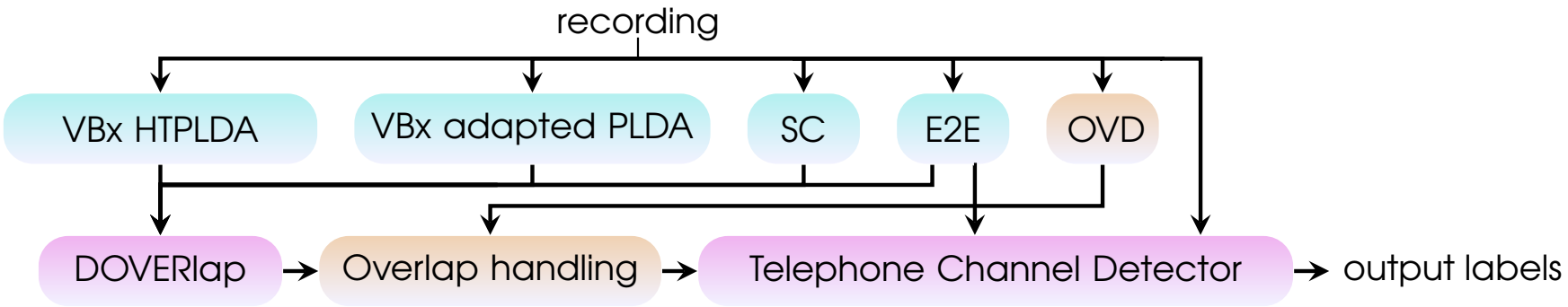
[5] Luo et al., Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation

Telephone Channel Detector

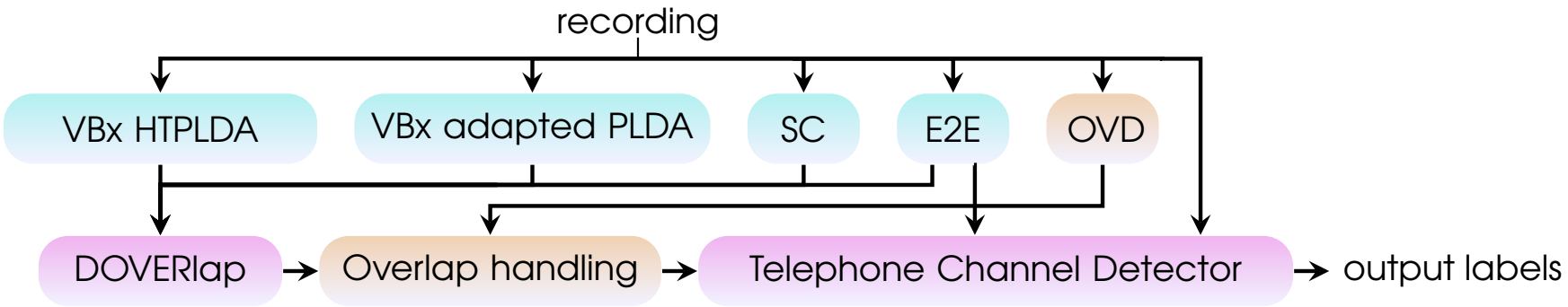


- Analyzing the average energy levels in spectrogram, utterances are classified as telephone or wide-band
- Telephone utterances are processed with the E2E system
- Other utterances are processed with the fusion+overlap

Results


recording
VBx HTPLDA
VBx adapted PLDA
SC
E2E
OVD
DOVERlap
Overlap handling
Telephone Channel Detector
output labels

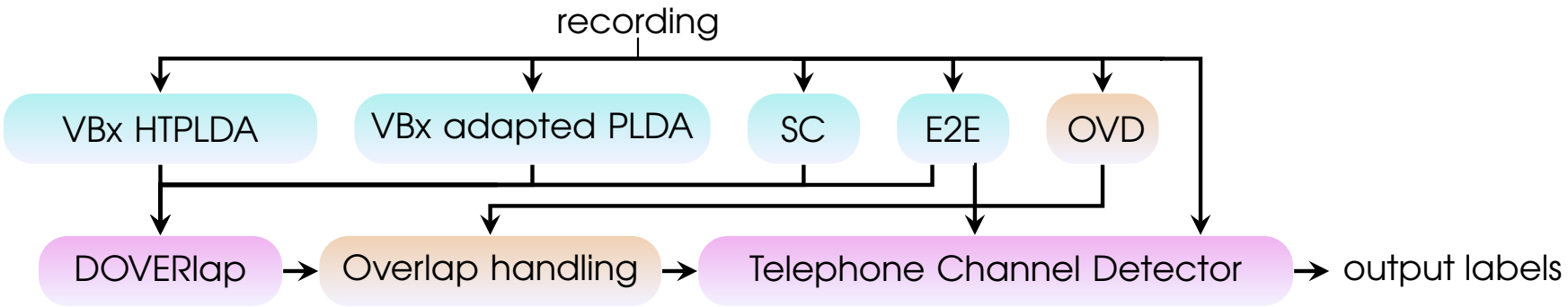
Results



Track 1

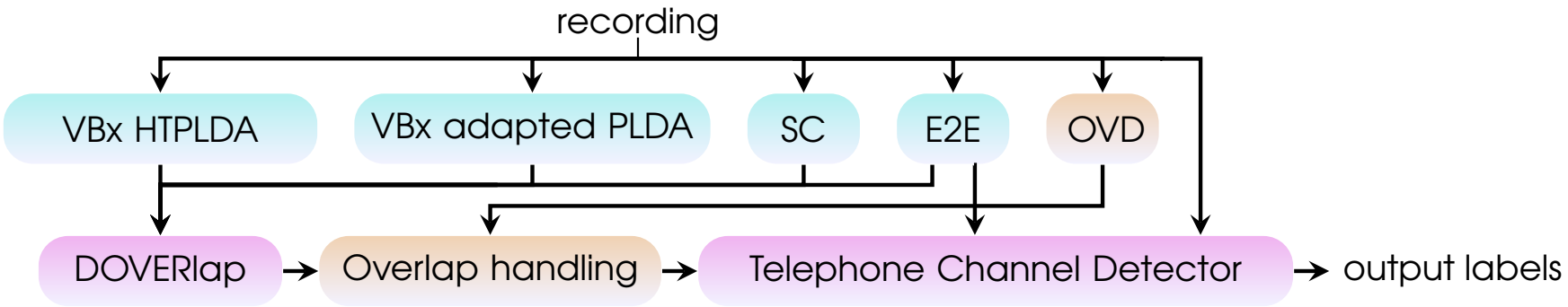
System	Development								Evaluation	
	Core				Full				Core	Full
	DER	Miss	FA	SER	DER	Miss	FA	SER	DER	DER
VBx HTPLDA	16.33	10.95	0	5.38	15.98	10.93	0	5.05	16.54	15.5
VBx adapted PLDA	16.66	10.95	0	5.72	16.26	10.93	0	5.33	16.67	15.74
SC	16.63	10.95	0	5.69	16.51	10.93	0	5.58	16.56	15.79
E2E	24.17	8.89	1.69	13.59	20.59	7.82	1.88	10.89	23.51	19.06

Results



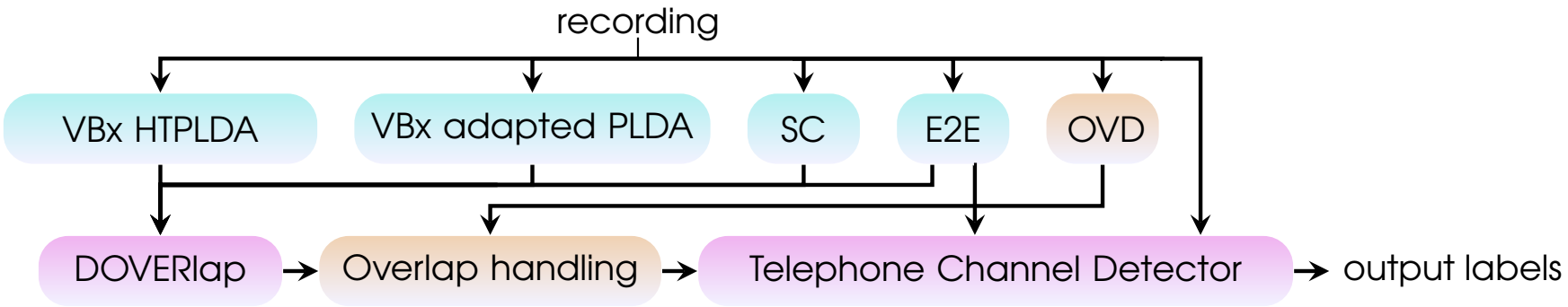
Track 1

System	Development								Evaluation	
	Core				Full				Core	Full
	DER	Miss	FA	SER	DER	Miss	FA	SER	DER	DER
VBx HTPLDA	16.33	10.95	0	5.38	15.98	10.93	0	5.05	16.54	15.5
VBx adapted PLDA	16.66	10.95	0	5.72	16.26	10.93	0	5.33	16.67	15.74
SC	16.63	10.95	0	5.69	16.51	10.93	0	5.58	16.56	15.79
E2E	24.17	8.89	1.69	13.59	20.59	7.82	1.88	10.89	23.51	19.06
DOVERlap	15.86	10.94	0.01	4.92	15.57	10.92	0	4.65	16.22	15.26



Track 1

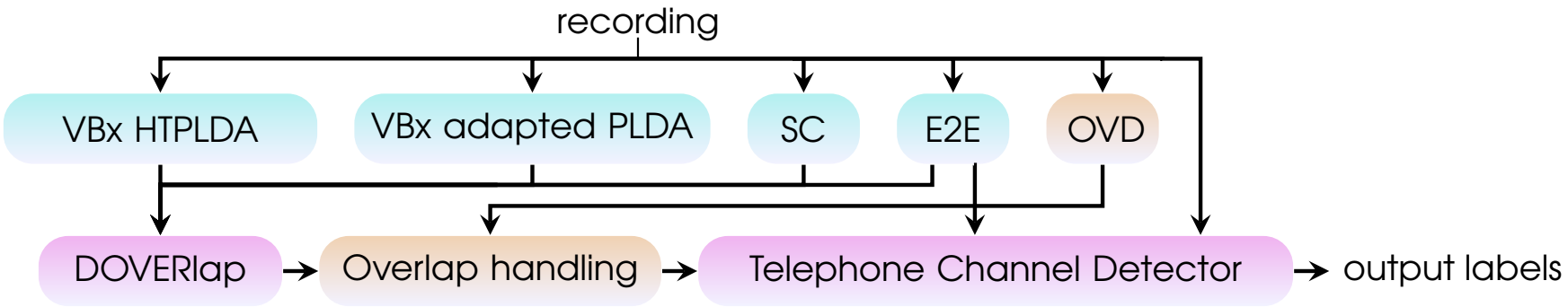
System	Development								Evaluation	
	Core				Full				Core	Full
	DER	Miss	FA	SER	DER	Miss	FA	SER	DER	DER
VBx HTPLDA	16.33	10.95	0	5.38	15.98	10.93	0	5.05	16.54	15.5
VBx adapted PLDA	16.66	10.95	0	5.72	16.26	10.93	0	5.33	16.67	15.74
SC	16.63	10.95	0	5.69	16.51	10.93	0	5.58	16.56	15.79
E2E	24.17	8.89	1.69	13.59	20.59	7.82	1.88	10.89	23.51	19.06
DOVERlap	15.86	10.94	0.01	4.92	15.57	10.92	0	4.65	16.22	15.26
+ ov. handling	15.03	9.76	0.09	5.18	14.30	9.38	0.11	4.82	16.07	14.25



Track 1

System	Development								Evaluation	
	Core				Full				Core	Full
	DER	Miss	FA	SER	DER	Miss	FA	SER	DER	DER
VBx HTPLDA	16.33	10.95	0	5.38	15.98	10.93	0	5.05	16.54	15.5
VBx adapted PLDA	16.66	10.95	0	5.72	16.26	10.93	0	5.33	16.67	15.74
SC	16.63	10.95	0	5.69	16.51	10.93	0	5.58	16.56	15.79
E2E	24.17	8.89	1.69	13.59	20.59	7.82	1.88	10.89	23.51	19.06
DOVERlap	15.86	10.94	0.01	4.92	15.57	10.92	0	4.65	16.22	15.26
+ ov. handling	15.03	9.76	0.09	5.18	14.30	9.38	0.11	4.82	16.07	14.25
Final fusion	14.56	9.37	0.27	4.91	13.49	8.17	0.82	4.49	15.46	13.29

Results



Track 1

System	Development								Evaluation	
	Core				Full				Core	Full
	DER	Miss	FA	SER	DER	Miss	FA	SER	DER	DER
VBx HTPLDA	16.33	10.95	0	5.38	15.98	10.93	0	5.05	16.54	15.5
VBx adapted PLDA	16.66	10.95	0	5.72	16.26	10.93	0	5.33	16.67	15.74
SC	16.63	10.95	0	5.69	16.51	10.93	0	5.58	16.56	15.79
E2E	24.17	8.89	1.69	13.59	20.59	7.82	1.88	10.89	23.51	19.06
DOVERlap	15.86	10.94	0.01	4.92	15.57	10.92	0	4.65	16.22	15.26
+ ov. handling	15.03	9.76	0.09	5.18	14.30	9.38	0.11	4.82	16.07	14.25
Final fusion	14.56	9.37	0.27	4.91	13.49	8.17	0.82	4.49	15.46	13.29

- Our VBx system for DIHARD II[6] obtains 16.89% DER on development core and 16.46% DER on development full

[6] https://github.com/BUTSpeechFIT/VBx/tree/v1.0_DIHARDII

Results

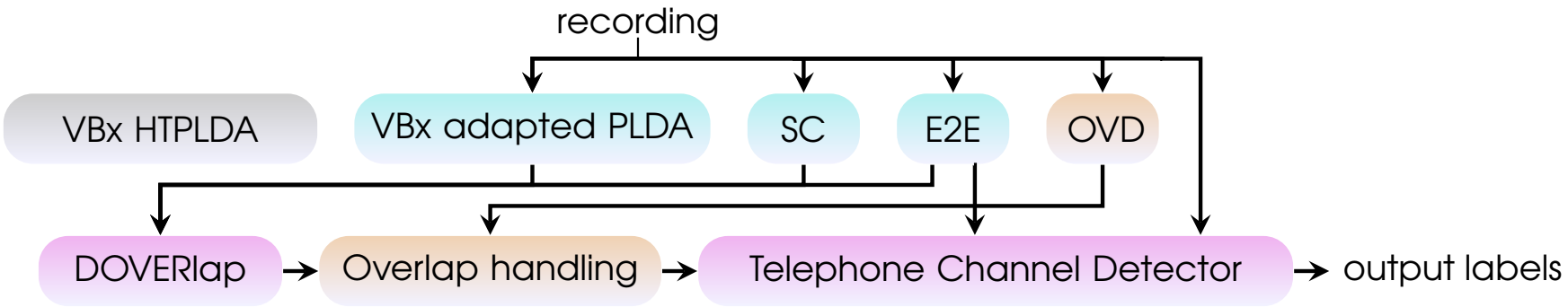
System	ALL	audiobooks	broadcast	clinical	court	cts
VBx HTPLDA	16.33	2	2.41	10.04	2.9	16.52
VBx adapted PLDA	16.66	3.83	2.11	10.32	2.73	17.24
SC	16.63	0.38	3.13	11.2	3.5	16.7
E2E	24.17	0.56	14.42	21.62	25.31	9.29
DOVERlap	15.86	0	2.42	9.43	3.01	16.29
+ ov. handling	15.03	0	2.32	9.17	2.77	13.78
Final fusion	14.56	0	2.32	9.17	2.77	9.29

System	maptask	meeting	restaurant	soc. field	soc. lab	webvideo
VBx HTPLDA	4.89	26.52	39.89	12.82	8.13	35.12
VBx adapted PLDA	4.92	26.13	40.54	13.36	7.88	36.36
SC	6.09	26.87	38.93	13.77	8.33	36.32
E2E	16.97	39.02	53.96	18.86	7.18	40.36
DOVERlap	4.63	25.94	39.59	12.28	6.99	35.45
+ ov. handling	3.36	24.59	39.16	11.95	6.33	34.33
Final fusion	3.36	24.59	39.16	11.95	6.33	34.33

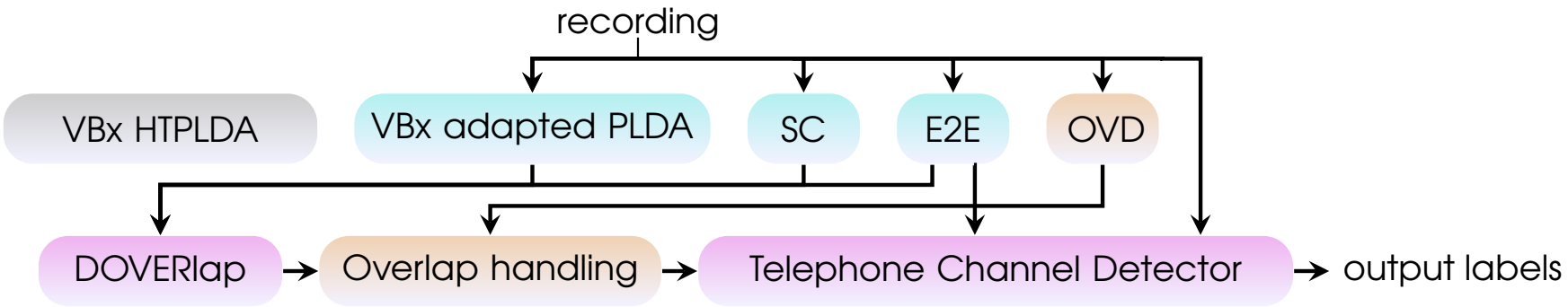
Results

System	ALL	audiobooks	broadcast	clinical	court	cts
VBx HTPLDA	16.33	2	2.41	10.04	2.9	16.52
VBx adapted PLDA	16.66	3.83	2.11	10.32	2.73	17.24
SC	16.63	0.38	3.13	11.2	3.5	16.7
E2E	24.17	0.56	14.42	21.62	25.31	9.29
DOVERlap	15.86	0	2.42	9.43	3.01	16.29
+ ov. handling	15.03	0	2.32	9.17	2.77	13.78
Final fusion	14.56	0	2.32	9.17	2.77	9.29

System	maptask	meeting	restaurant	soc. field	soc. lab	webvideo
VBx HTPLDA	4.89	26.52	39.89	12.82	8.13	35.12
VBx adapted PLDA	4.92	26.13	40.54	13.36	7.88	36.36
SC	6.09	26.87	38.93	13.77	8.33	36.32
E2E	16.97	39.02	53.96	18.86	7.18	40.36
DOVERlap	4.63	25.94	39.59	12.28	6.99	35.45
+ ov. handling	3.36	24.59	39.16	11.95	6.33	34.33
Final fusion	3.36	24.59	39.16	11.95	6.33	34.33



- Baseline VAD instead of oracle labels



- Baseline VAD instead of oracle labels

System	Development								Evaluation	
	Core				Full				Core	Full
	DER	Miss	FA	SER	DER	Miss	FA	SER	DER	DER
VBx adapted PLDA	19.49	12.6	0.91	5.98	19.14	12.59	0.96	5.58		
SC	19.58	12.61	0.91	6.06	19.52	12.6	0.96	5.95		
E2E	26.14	10.41	2.49	13.24	22.68	9.39	2.76	10.54		
DOVERlap	19.07	12.57	0.91	5.59	18.74	12.54	0.97	5.23		
+ ov. handling	17.89	10.32	1.35	6.22	16.89	9.84	1.4	5.65		
Final fusion	17.52	10.09	1.51	5.91	16.32	9.17	2.02	5.12	24.62	21.09
Final fusion Track 1	14.56	9.37	0.27	4.91	13.49	8.17	0.82	4.49	15.46	13.29

Final remarks

- Dealing with overlap using standard approaches is still challenging
- End-to-end approaches naturally model that aspect
- However, they still fall behind in overall performance against oracle VAD + standard approaches