BUT System Description 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, Speech@FIT, Czechia
[2]Omilia - Conversational Intelligence, Greece
{*landini, lozano, mireia*}*@fit.vutbr.cz*

***Abstract*—This is the system description corresponding to the systems developed by the BUT team for The Third DIHARD Speech Diarization Challenge. The systems for both tracks consist of a DOVERlap fusion of an end-to-end NN system with x-vector based clustering systems in the form of spectral clustering and VBx. Given that the x-vector clustering systems do not provide overlapping speakers, overlapped speech is detected by a TasNet-based detector before the final fusion with the end-to-end approach.**

***Index Terms*—Speaker Diarization, DIHARD, VBx diarization, end-to-end diarization, overlapped speech detection**

I. Notable highlights

Our best submitted system for Track 1 follows the pipeline described in Figure 1. Four systems are run in parallel on the input to produce diarization labels. Three of those systems are based on clustering of x-vectors and the remaining is an end-to-end system. The outputs of all systems are fused; then, an overlapped speech detector is run and an heuristic is used to assign a second speaker in those segments determined by the overlap detector. Finally, since the end-to-end system performs substantially better than the rest on telephone conversations, a telephone channel detector is used to detect recordings of that domain. Telephone recordings are then processed only with the end-to-end system while the other recordings are processed with the whole pipeline.

Details about the methods and citations to relevant works are presented in the following sections.

II. Data resources

The list of datasets used to produce our systems is the following:

- VoxCeleb 2[1]
- AMI[2]
- ICSI: LDC2004S02, LDC2004T04
- ISL: LDC2004S05, LDC2004T10
- DIHARD III development set [1]
- CALLHOME: LDC96S34, LDC96S35, LDC96S37, LDC96T16, LDC96T17, LDC96T18, LDC97S42, LDC97S43, LDC97S45, LDC97T14, LDC97T15, LDC97T19
- The artificial conversations to train the end-to-end system were created from NIST SRE and SWITCHBOARD: LDC2006S44, LDC2011S01, LDC2011S04, LDC2011S09, LDC2011S10, LDC2012S01, LDC2011S05, LDC2011S08, LDC98S75, LDC99S79, LDC2002S06, LDC2001S13, LDC2004S07
- VoxConverse development set[3]

They were used for different parts of the system so we detail in each one which datasets are used.

III. Detailed description of algorithm and results

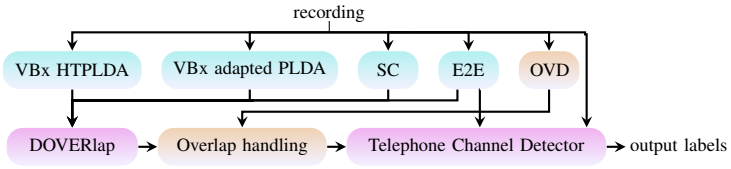

Fig. 1. Diagram of the final diarization system for Track 1.

For Tracks 1 and 2 we used very similar pipelines. For being Track 1 the track where we spent most of our effort, we describe it in more detail. For the final system for Track 2 we simply explain the differences. Figure 1 presents the pipeline for Track 1. We explain first the overall system and then each subsystem below.

Four different diarization systems are used first: three of them based on clustering of x-vectors and one end-to-end (E2E) approach. The x-vectors are extracted only on speech segments according to the oracle voice activity detection (VAD). The end-to-end approach uses the oracle VAD only as a post-processing step. The outputs of the four systems are then fused using DOVERlap [2]. Since only the end-to-end approach produces overlapped speech labels by default, we apply an overlapped speech detector (OVD) and add a second simultaneous speaker in some segments according to

The work was supported by Czech National Science Foundation (GACR) project "NEUREM3" No. 19-26934X, European Union's Horizon 2020 project No. 833635 ROXANNE and European Union's Marie Sklodowska-Curie grant agreement No. 843627. Some of the methods were implemented during the JSALT2020 workshop, hosted by JHU.

[1]https://www.robots.ox.ac.uk/ vgg/data/voxceleb/vox2.html

[2]http://groups.inf.ed.ac.uk/ami/corpus/

[3]http://www.robots.ox.ac.uk/~vgg/data/voxconverse/

its output. Finally, given the superior performance of the E2E system on telephone conversations, we use a telephone channel detector to find which recordings are telephonic and for those we only apply the E2E approach. For the rest, we use the pipeline already described.

A. VBx HTPLDA

This system follows the approach explained in [3] and released in the recipe of VBx https://github.com/BUTSpeechFIT/VBx/tree/v1.0_DIHARDII. In particular, the x-vectors are extracted on 1.5 s long segments every 0.25 s and over shorter segments when they are shorter than 1.5 s. The x-vector extractor is based on a time-delay neural network (TDNN) architecture and more details about its training can be found in section "X-vector extractor" in the first page of [3]. The x-vectors are clustered using agglomerative hierarchical clustering (AHC) with similarity metric based on probabilistic linear discriminant analysis (PLDA) [4] log-likelihood ratio scores as used for speaker verification and then the Bayesian hidden Markov model (HMM) for diarization is applied. The difference with the VBx recipe is that in this system a heavy tailed PLDA was used instead of a Gaussian one; hence, the name "VBx HTPLDA". The rest of the recipe is explained in [5], [6]. The parameters used in this system are as shared in the recipe except for F_A, F_B and P_{loop} which were tuned to obtain the best performance on the DIHARD III development set ($F_A = 0.5$, $F_B = 10$ $P_{loop} = 0$). Results for this system are presented in Table I.

TABLE I
RESULTS (%) FOR THE SYSTEM VBX HTPLDA FOR TRACK 1 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

	Development		Evaluation	
	Core	Full	Core	Full
DER	16.33	15.98	16.54	15.5
JER	36.82	33.35	37.82	33.61

B. VBx adapted PLDA

This system follows the VBx part of the system described in [7]. The x-vectors are extracted using a system based on a ResNet152 architecture and the details on how it was trained can be found on that publication. The only difference in VBx for diarization is that in the system for DIHARD III an interpolation of PLDA models was used following the approach described in [6]. The model trained on speakers from VoxCeleb 2 was assigned a weight of 0.95 and the PLDA trained on speakers from the development set of DIHARD III was assigned a weight of 0.05. The only hyperparameter of the model different to those shared in [7] is F_B after tuning on the development set of DIHARD III ($F_A = 0.3$, $F_B = 14$ and $P_{loop} = 0.9$). Results for this system are presented in Table II.

C. Spectral Clustering (SC)

This system makes use of the ResNet152-based x-vectors described in the previous section. The PLDA model is used to transform the x-vectors and reduce their dimensionality into 160. However, the x-vectors are compared by means of cosine similarity in order to produce the affinity matrix used for clustering. The affinity matrix is modified so that each x-vector has 0.3 extra affinity with the 4 following x-vectors in order to favor x-vectors close to each other in the time-domain to be part of the same cluster. Then, only the affinities between each x-vector and the 27 most affine ones are kept (the rest are zeroed). Finally, the number of clusters is estimated using the largest eigen-gap in spectral clustering [8] assuming a maximum number of clusters to be 20. The x-vectors are then clustered using k-means. All hyperparameters were tuned to reach the best performance on the development set of DIHARD III. Results for this system are presented in Table III.

TABLE II
RESULTS (%) FOR THE SYSTEM VBX ADAPTED PLDA FOR TRACK 1 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

	Development		Evaluation	
	Core	Full	Core	Full
DER	16.66	16.26	16.67	15.74
JER	37.19	33.68	37.69	33.75

TABLE III
RESULTS (%) FOR THE SYSTEM SC FOR TRACK 1 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

	Development		Evaluation	
	Core	Full	Core	Full
DER	16.63	16.51	16.56	15.79
JER	38.67	34.97	38.72	34.46

D. E2E

This system follows the approach described in [9] based on self-attention and encoder-decoder long short-term memory based attractors. For training we follow the same approach as in the paper: we train a model on 100000 simulated mixtures of two speakers for 100 epochs, followed by the training of a model with 400000 mixtures of up to 4 speakers for 25 epochs. Then, we finetune the model to the whole CALLHOME dataset for another 100 epochs and with up to 7 speakers. For evaluation, we combine the outputs of the system with some external VAD segmentation as follows: the system outputs the most likely speaker for each time frame and a threshold is used to indicate the presence of other speakers in the same frame; then, the false alarm is compensated with the corresponding external VAD labels (the oracle VAD for Track 1 and the baseline VAD for Track 2). This system operates on 8 kHz data, and therefore, the DIHARD III development and evaluation sets are downsampled to this frequency from the original 16 kHz. Results for this system are presented in Table IV.

TABLE IV
RESULTS (%) FOR THE SYSTEM E2E FOR TRACK 1 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

	Development		Evaluation	
	Core	Full	Core	Full
DER	24.17	20.59	23.51	19.06
JER	56.68	49.76	53.45	45.87

E. DOVERlap

The outputs of the four systems described above are fused using DOVERlap [2] which is an improved version of the original DOVER [10] approach for fusion of diarization systems. We used the official implementation of DOVERlap as stated in the paper with the default configuration. Although DOVERlap uses the outputs of the systems, it does not take any of the systems as primary and can also deal with overlapping segments. Results for the output of DOVERlap are presented in Table V.

TABLE V
RESULTS (%) FOR THE DOVERLAP FUSION FOR TRACK 1 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

	Development		Evaluation	
	Core	Full	Core	Full
DER	15.86	15.57	16.22	15.26
JER	38.36	34.5	39.47	35.08

F. Overlapped speech detection and handling

The OVD is based on the Conv-TasNet architecture [11] and its implementation in Asteroid [12]. It uses the encoder and separator parts of Conv-TasNet followed by softmax to classify 2 ms frames into three classes: silence, single-speaker speech and overlapped speech. The hyper-parameters used for the architecture are: $N = 192, L = 64, B = 128, Sc = 128, H = 192, X = 4, R = 3$. For training, we use DIHARD III dev set, VoxConverse [13] dev set and three meeting datasets: ICSI [14], ISL [15] and AMI [16] train set (both beamformed and Mix-Headset). At first, we sample from the datasets with ratio $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ for DIHARD:VoxConverse:meeting datasets and anneal towards 1:0:0 in each training iteration. In half of the samples, we use the real data directly and in the other half, we artificially mix two segments to create more overlap examples. We use 150k iterations with batch size 24, Adam optimizer and learning rate 1e-3. For both training and inference, we apply the network on 3 s segments. We tuned the overlap detection threshold on the development set but, expecting a slightly different threshold would work better on the evaluation set (as the class priors could be different), we tried more than one threshold. The final threshold was 0.93.

We handle overlaps by assigning a second speaker to those segments detected as overlaps, chosen as the one with the highest temporal proximity [17]. We compared this heuristic with using the second most likely speaker given by VBx (based on [18]) but, as in [7], this approach worked slightly worse than the heuristic. Results for the output of DOVERlap with overlap detection and handling are presented in Table VI.

TABLE VI
RESULTS (%) FOR THE DOVERLAP FUSION WITH OVERLAP DETECTION FOR TRACK 1 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

	Development		Evaluation	
	Core	Full	Core	Full
DER	15.03	14.30	16.07	14.25
JER	37.72	33.62	39.09	34.32

G. Telephone channel detection

Due to the superior performance of the end-to-end system with respect to the rest of the systems on the telephone domain, we decided to use a telephone channel detector. Recordings identified as telephone were processed only with the E2E system and the others with the rest of the pipeline. The detector consisted in averaging the upper part of the spectrogram of the recording. If the level was below 125, the files were classified as telephone and as non-telephone otherwise. Results for the output of DOVERlap with overlap detection and handling for non-telephone recordings but the E2E system for telephone ones are presented in Table VII.

TABLE VII
RESULTS (%) FOR THE FINAL SYSTEM FOR TRACK 1 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

	Development		Evaluation	
	Core	Full	Core	Full
DER	14.56	13.49	15.46	13.29
JER	37.42	32.92	38.68	33.45

H. Track 2 system

For Track 2, we used a similar pipeline as for Track 1 except for VBx HTPLDA which was not used at all. Also, instead of the oracle VAD, we used the baseline VAD provided by the organizers [1] which consists of a TDNN trained on DIHARD III development set. Since we only submitted the final system, we do not have results for the intermediate systems for this track. Results for the system on Track 2 are presented in Table VIII.

TABLE VIII
RESULTS (%) FOR THE FINAL SYSTEM FOR TRACK 2 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

	Development		Evaluation	
	Core	Full	Core	Full
DER	17.52	16.32	24.62	21.09
JER	40.72	36.17	44.49	39.28

IV. System comparison

A comparison of the results obtained by each system for Track 1 detailed by the domains on the development set is presented in Table IX. We observe that the E2E approach trained on telephone speech obtains the best result in the cts domain (telephone) by far. However, it also has a performance on par with the others for audiobooks and sociolinguistic lab domains. The reason for this is still unclear and requires further experiments. On one side, these are among the cleanest domains in terms of background noise; however, the same could be said about the broadcast interviews or maptask and yet this approach shows notable performance degradation on those domains. As for the x-vector based approaches, although all of them have similar performance for the whole core set, we see differences of 1 point between the best and worst approach for almost every domain. With the fusion, we see that in all domains the resulting system has either better performance than the best of the four or the performance is close to that of the best system.

When applying the overlapped speech handling, we obtain over 5% relative improvement in terms of DER, with some gain on all of the domains. However, the OVD system was partially trained on the development set so the results are over-optimistic. Comparing the results for development and evaluation on Table X, this is clear with only 0.15% DER improvement when applying the overlap handling on the core condition. The effect on the full condition is larger mainly because a large improvement is obtained on cts, a domain with a larger proportion in full than in core. However, the E2E approach has even better performance on such domain and, it is not over-optimistic; then, the gain on the evaluation set for full condition is even more.

Results for the different approaches on Track 2 are presented in Table XI. Although the performance of the final system shows 22% relative deterioration in Track 2 wrt Track 1 on the development set, the degradation is 55% relative on the evaluation set, showing that the results on the dev set are overoptimistic given that the VAD was trained on such set. When exploring other approaches for VAD, we saw that usually different domains require different parameters for voice detection, proving that producing a one-fits-all VAD system is indeed challenging.

V. Hardware requirements

The infrastructure used to run the experiments was, in the case of CPU, an Intel(R) Xeon(R) CPU 5675 @ 3.07GHz, with a total memory of 37GB unless specified otherwise. In the case of GPU, a Tesla P100 PCIe with 16GB of memory unless specified otherwise.

- The training of the TDNN-based x-vector extractor took approximately 215 hours on GPU. The extraction of x-vectors on CPU takes less than 19 s to process 1 minute of recording.
- The training of the ResNet152-based x-vector extractor was done on three NVidia Quadro RTX 8000 GPU cores in parallel and took approximately 60 hours. The training was done using Pytorch and Horovod parallelization library. Each job required 35GB of GPU memory and 2GB of CPU memory. The extraction of x-vectors on CPU takes less than 33 s to process 1 minute of recording.
- The training of each of the PLDA models takes less than 5 minutes.
- AHC has the time complexity $\mathcal{O}(n^3)$ and memory complexity $\mathcal{O}(n^2)$, which is in both cases the highest (theoretical) complexity out of all the processing steps. Therefore, AHC could become the bottleneck for very long utterances. However, even for the longest utterances in the DIHARD set (around 10 minutes), our non-optimized python implementation of AHC is still about 20 faster than realtime. On average, on all the DIHARD recordings (taking into account only speech and not silence), AHC is around 100 faster than real-time.
- Bayesian HMM based clustering of x-vectors initialized from the AHC is more than 200 times faster than real-time on the DIHARD recordings.
- Producing the diarization output with spectral clustering with k-means is more than 20 times faster than real-time on the DIHARD recordings.
- The training of the end-to-end system was performed in a single GPU and it took around 520 hours. The inference takes less than 0.1 s to process 1 minute of audio in GPU.
- The training of the OVD system took less than 8.5 hours in GPU and the evaluation takes around 5 s to process 1 minute of recording in CPU.
- Post-processing the speaker labels for overlap detection on a recording of 10 minutes varied from less than a second to 1 minute depending on the amount of overlapped speech segments found.
- For training the baseline VAD, the feature extraction was performed in parallel on 50 CPUs taking less than one hour. The training was run on GPU and the total time for the training was approximately 13 hours. For evaluating recordings, the script takes less than 1 s to process 1 minute of recording on CPU.

References

[1] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The Third DIHARD Diarization Challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[2] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs," *arXiv preprint arXiv:2011.01997*, 2020.

[3] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, O. Plchot, O. Novotnỳ, H. Zeinali *et al.*, "BUT System Description for DIHARD Speech Diarization Challenge 2019," *arXiv preprint arXiv:1910.08847*, 2019.

[4] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *in Proceedings of Odyssey*, Jun. 2010.

[5] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotnỳ *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.

[6] M. Diez, L. Burget, F. Landini, S. Wang, and H. Černockỳ, "Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge," in *ICASSP 2020-2020 IEEE International*

TABLE IX
DER (%) FOR THE DIFFERENT SYSTEMS FOR TRACK 1 ON EACH OF THE DOMAINS OF THE DEVELOPMENT SET, CORE CONDITION.

System	ALL	audiob.	broadc.	clinical	court	cts	maptask	meeting	restaurant	soc. field	soc. lab	webvideo
VBx adapted PLDA	16.66	3.83	2.11	10.32	2.73	17.24	4.92	26.13	40.54	13.36	7.88	36.36
VBx HTPLDA	16.33	2	2.41	10.04	2.9	16.52	4.89	26.52	39.89	12.82	8.13	35.12
SC	16.63	0.38	3.13	11.2	3.5	16.7	6.09	26.87	38.93	13.77	8.33	36.32
E2E	24.17	0.56	14.42	21.62	25.31	9.29	16.97	39.02	53.96	18.86	7.18	40.36
DOVERlap	15.86	0	2.42	9.43	3.01	16.29	4.63	25.94	39.59	12.28	6.99	35.45
+ ov. handling	15.03	0	2.32	9.17	2.77	13.78	3.36	24.59	39.16	11.95	6.33	34.33
Final fusion	14.56	0	2.32	9.17	2.77	9.29	3.36	24.59	39.16	11.95	6.33	34.33

TABLE X
RESULTS (%) FOR THE DIFFERENT SYSTEMS FOR TRACK 1 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

System	Development										Evaluation			
	Core					Full					Core		Full	
	DER	Miss	FA	SER	JER	DER	Miss	FA	SER	JER	DER	JER	DER	JER
VBx adapted PLDA	16.66	10.95	0	5.72	37.19	16.26	10.93	0	5.33	33.68	16.67	37.69	15.74	33.75
VBx HTPLDA	16.33	10.95	0	5.38	36.82	15.98	10.93	0	5.05	33.35	16.54	37.82	15.5	33.61
SC	16.63	10.95	0	5.69	38.67	16.51	10.93	0	5.58	34.97	16.56	38.72	15.79	34.46
E2E	24.17	8.89	1.69	13.59	56.68	20.59	7.82	1.88	10.89	49.76	23.51	53.45	19.06	45.87
DOVERlap	15.86	10.94	0.01	4.92	38.36	15.57	10.92	0	4.65	34.5	16.22	39.47	15.26	35.08
+ ov. handling	15.03	9.76	0.09	5.18	37.72	14.30	9.38	0.11	4.82	33.62	16.07	39.09	14.25	34.32
Final fusion	14.56	9.37	0.27	4.91	37.42	13.49	8.17	0.82	4.49	32.95	15.46	38.68	13.29	33.45

TABLE XI
RESULTS (%) FOR THE DIFFERENT SYSTEMS FOR TRACK 2 ON DEVELOPMENT AND EVALUATION SETS, CORE AND FULL CONDITIONS.

System	Development										Evaluation			
	Core					Full					Core		Full	
	DER	Miss	FA	SER	JER	DER	Miss	FA	SER	JER	DER	JER	DER	JER
VBx adapted PLDA	19.49	12.6	0.91	5.98	39.83	19.14	12.59	0.96	5.58	36.43				
SC	19.58	12.61	0.91	6.06	41.78	19.52	12.6	0.96	5.95	38.18				
E2E	26.14	10.41	2.49	13.24	57.61	22.68	9.39	2.76	10.54	50.86				
DOVERlap	19.07	12.57	0.91	5.59	41.66	18.74	12.54	0.97	5.23	37.86				
+ ov. handling	17.89	10.32	1.35	6.22	40.99	16.89	9.84	1.4	5.65	36.76				
Final fusion	17.52	10.09	1.51	5.91	40.72	16.32	9.17	2.02	5.12	36.17	24.62	44.49	21.09	39.28

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6519–6523.

[7] F. Landini, O. Glembek, P. Matějka, J. Rohdin, L. Burget, M. Diez, and A. Silnova, "Analysis of the BUT Diarization System for VoxConverse Challenge," *arXiv preprint arXiv:2010.11718*, 2020.

[8] H. Ning, M. Liu, H. Tang, and T. S. Huang, "A spectral clustering approach to speaker diarization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[9] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," in *Proc. Interspeech 2020*, 2020, pp. 269–273. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1022

[10] A. Stolcke and T. Yoshioka, "Dover: A method for combining diarization outputs," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 757–763.

[11] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas *et al.*, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[13] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *arXiv preprint arXiv:2007.01216*, 2020.

[14] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The ICSI meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[15] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Seventh International Conference on Spoken Language Processing*, 2002.

[16] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[17] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 683–686.

[18] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.