ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097049

MULTI-SPEAKER AND WIDE-BAND SIMULATED CONVERSATIONS AS TRAINING DATA FOR END-TO-END NEURAL DIARIZATION

Federico Landini[1], *Mireia Diez*[1], *Alicia Lozano-Diez*[2], *Lukáš Burget*[1]

[1]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
[2]AUDIAS (Audio, Data Intelligence and Speech), Universidad Autónoma de Madrid, Spain
{*landini,mireia,burget*}*@fit.vutbr.cz, alicia.lozano@uam.es*

ABSTRACT

End-to-end diarization presents an attractive alternative to standard cascaded diarization systems because a single system can handle all aspects of the task at once. Many flavors of end-to-end models have been proposed but all of them require (so far non-existing) large amounts of annotated data for training. The compromise solution consists in generating synthetic data and the recently proposed simulated conversations (SC) have shown remarkable improvements over the original simulated mixtures (SM). In this work, we create SC with multiple speakers per conversation and show that they allow for substantially better performance than SM, also reducing the dependence on a fine-tuning stage. We also create SC with wide-band public audio sources and present an analysis on several evaluation sets. Together with this publication, we release the recipes for generating such data and models trained on public sets as well as the implementation to efficiently handle multiple speakers per conversation and an auxiliary voice activity detection loss.

Index Terms— Speaker diarization, end-to-end neural diarization, simulated conversations

1. INTRODUCTION

End-to-end neural diarization (EEND) is a popular alternative to the traditional cascaded speaker diarization system composed of different submodules, i.e. voice activity detection (VAD), uniform segmentation, speaker embedding extraction, clustering and overlapped speech detection and handling. Instead of having separate models trained independently, EEND formulates the speaker diarization problem as a per-speaker-per-time-frame binary classification problem where a permutation-free objective is used to minimize the speech activity error for all speakers. This has the advantage of training the model directly for the task of interest; but unlike in the cascaded paradigm, end-to-end models require large amounts of data with diarization annotations. Given that currently available data with manual annotations for diarization are scarce, the usual strategy is to create audios where segments of different recordings are combined and VAD annotations define the ground truth annotations.

When EEND was proposed, Fujita et al. defined a procedure to generate training data in the form of simulated mixtures (SM) [1] Algorithm 1. The method selects the number of speakers that the mixture will have and for each speaker, speech segments are selected from a pool. Then, pauses are introduced between the segments of a speaker. This is repeated for each speaker independently, producing one channel per speaker. Finally, the channels are summed to produce the final mixture. In order to enrich the mixtures, the speech from each speaker is reverberated and background noise is added to the final mixture. The main disadvantage of this approach is that the pauses between the turns of each speaker (and resulting overlaps) are defined independently and this does not resemble real conversations. In order to tackle this problem, approaches that consider the interaction between speakers were proposed [2, 3]. Instead of treating speakers independently, simulated conversations (SC) [2] Algorithm 1, are defined by interleaving speech segments and determining pauses or overlaps between them with probabilities and lengths drawn from distributions estimated on real interactions. SC can also be optionally enriched with reverberation and background noises. It has been shown [2] that SC allow for better performing EEND models on 2-speaker telephone conversations and reduce the dependence on a fine-tuning stage, i.e. after training the model with SC, it is further trained using a small development set with low learning rate.

Most works with end-to-end models so far have focused on the telephony scenario where large amounts of data are available and where each speaker is recorded in a separate channel. This allows to create diarization segment annotations by simply running a VAD system on each channel. This scenario presents certain advantages such as (usually) one speaker per channel, conversational speech and relatively similar channel characteristics across recordings. This last aspect allows for a relatively easy amalgamation of different recordings when creating synthetic data where the channel differences are less relevant than the speaker ones and, thus, the model can focus on learning to separate speakers rather than channels.

Contrarily, wide-band data can have more variability in terms of recording devices, settings and types of interactions. However, it is rare to have a single speaker per recording, even less in a conversational scenario. Thus, applying the same techniques for creating synthetic data becomes more challenging since the mismatch between (artificial) training and (authentic) test data can become much larger. There have been efforts [4, 5, 6, 7] in generating training data for EEND using wide-band recordings from LibriSpeech [8] mostly based on SM. In this work, we aim to extend the usage of SC to wide-band data while exploring different available corpora . For the sake of keeping the focus on the data sources for SC, we perform the analysis under the scenario of two speakers per conversation.

In order to analyze the advantages of SC for more than two speakers, we focus on the already studied telephony case where SC has been shown to be more realistic than SM [2]. Furthermore, a modification in the loss is introduced in order to be able to train models with several speakers efficiently and an auxiliary loss is added to improve the model's VAD capabilities.

The work was supported by Czech Ministry of Interior project No. VJ01010108 "ROZKAZ", Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, Horizon 2020 Marie Sklodowska-Curie grant ESPERANTO, No. 101007666 and grant PID2021-125943OB-I00, MCIN /AEI /10.13039/501100011033 / FEDER, UE. Computing on IT4I supercomputer was supported the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140).

2. WIDE-BAND SIMULATED CONVERSATIONS

Although there are nowadays numerous wide-band collections with thousands of hours of speech, they contain normally more than one speaker in the same channel. This makes it impossible to use the same strategy as with telephone for deriving the speaker turns. Instead, it is necessary to have some kind of segmentation already available that contains information about the speakers. Next, we describe the freely available datasets that we used for this work[1].

- LibriSpeech [8] consists of 1000 hours of read English speech from almost 2500 speakers. Each recording is expected to contain speech of a single speaker; thus, the original strategy of running VAD to obtain segmentation is possible. Recordings are of good quality and without background noises meaning that channel characteristics are relatively similar across recordings. However, all speech is read and not conversational.
- VoxCeleb2 [10] consists of more than 2400 hours of recordings from more than 6000 speakers speaking mostly English. Originally prepared as a training set for training speaker recognition systems, the recordings are partially annotated. This means that for the speakers of interest, some of their segments are identified. Thus, it is possible to derive speech segments for a given speaker without the need for any VAD system. At the same time, these annotations are automatically generated, possibly introducing errors such as including small excerpts from other speakers. Furthermore, the speech is collected "in the wild" so the recordings can have different noises. We observed better results if recordings with a poor signal-to-noise ratio (SNR) were filtered out.
- VoxPopuli [11] consists of recordings from the European Parliament in different languages. For this work, we used the subset for which transcriptions exist. Each recording is expected to contain speech from a single speaker and the transcription timestamps are used to derive the segmentation used for SC. Since the annotations were automatically derived, in some rare cases there are short excerpts from other speakers. This subset contains approximately 2700 hours from 2600 speakers. The recordings correspond to speakers' turns during plenary sessions which are monologues. Therefore, this corpus presents speech that is not read nor conversational but has exclusive turns and spontaneous speech. Recordings are normally of good quality and without background noises.

As in the telephony scenario, SC were augmented with background noises from the MUSAN collection [12] scaled with an SNR selected randomly from $\{5, 10, 15, 20\}$ dB. Room impulse responses and leveling of relative energy between speakers were evaluated as mentioned in [2] but the performance was about the same so we present results only when adding background noises.

3. SIMULATED CONVERSATIONS WITH SEVERAL SPEAKERS AND MODEL ADJUSTMENTS

SC are more realistic than SM in terms of the percentages of silences and overlap in the produced recordings[2]. Given that each speaker is modeled independently in SM, the percentages of overlapped speech are usually much larger than in real conversations. This is exacerbated the more speakers per mixture are used. On the contrary, SC contain proportions of silence and overlap that still resemble those seen in real data. We compare the performance of SM and SC with recordings with more than two speakers.

[1]It should be noted that GigaSpeech [9] was considered but, at the time of writing, segments do not have speaker labels so it is not usable for our purposes.

We carry out our analysis using the self-attention EEND with encoder-decoder attractors (EEND-EDA) [13]. One of its known limitations is the inability to handle more speakers per utterance than those seen during training, i.e. if the model is trained with recordings that contain up to 4 speakers, it will not perform well for recordings with more than 4. At the same time, one of the limitations in the setup originally proposed in [13] is that with the permutation invariant training (PIT) scheme, naively calculating all permutations becomes prohibitive in practice for more than 4 or 5 speakers per sequence. Faster alternatives have been studied [14] and in this work, we use PIT to find the best assignment between speakers and reference labels in polynomial time using the Hungarian algorithm[2]. This allows us to be able to train the models on utterances with more speakers without increasing the computational cost considerably.

In previous experiments with SC [2] it was observed that the outputs of the model have relatively high missed and false alarm speech. For this reason, and following a similar idea to that proposed in [15], we introduced the auxiliary VAD loss in Eq. 1 which reinforces per frame speech/non-speech classification.

$$L_{VAD} = -\frac{1}{F}\sum_{f}^{F} s_f \log(p(sil_f)) + (1-s_f)\log(1-p(sil_f)) \quad (1)$$

where $p(sil_f) = \prod_s (1 - y_f^s)$ represents the probability of silence given by the model for frame f, $s_f = \mathbb{1}[(\sum_s t_f^s) = 0]$ represents the silence label for frame f, where t_f^s and y_f^s are the label and model probability for activity of speaker s at frame f respectively. This loss is combined with the diarization and attractor existence losses in the following fashion $L = L_{diarization} + L_{attractors} + \alpha L_{VAD}$.

4. EXPERIMENTAL SETUP

4.1. Diarization model

All experiments were performed using the EEND-EDA for showing superior performance in previous works [13]. The architecture used was exactly the same as that described in [13][3]. For the sake of making the code more efficient, we used our PyTorch implementation[4]. 15 consecutive 23-dimensional log-scaled Mel-filterbanks (computed over 25 ms every 10 ms) are stacked to produce 345-dimensional features every 100ms. These are transformed by 4 self-attention encoder blocks (with 4 attention heads each) into a sequence of 256-dimensional embeddings. These are then shuffled in time and fed into the LSTM-based encoder-decoder module that decodes attractors, which are deemed as valid if their existence probability is above a certain threshold. A binary linear classifier is used to obtain speech activity probabilities for each speaker (represented by a valid attractor) at each time step (represented by an embedding).

The training scheme consists in training the model first on synthetic training data and then performing fine-tuning using a small development set of real data of the same domain as the test set. Usually, in the experiments with more than two speakers, a model initially trained on synthetic data with two speakers per recording is adapted to a synthetic set with more speakers and finally fine-tuned to a development set. The initial training is run for 100 epochs, the adaptation is run for 100 epochs on smaller sets or 25 epochs on (approximately four times) larger sets. The fine-tuning step is run for 100 epochs.

[2]In fact, we use scipy.optimize.linear_sum_assignment which implements the Jonker-Volgenant variant.

[3]We refer the reader to [13] for a scheme of the model.

[4]https://github.com/BUTSpeechFIT/EEND

Table 1. Evaluation sets for wide-band scenario. For AMI, numbers do not correspond to dev but to train set.

Name	Type	#files		Avg. length (s.)	
		dev	eval	dev	eval
CTS	telephone conversations	61	61	600	600
Clinical	interviews with children	48	51	≈300	≈300
Maptask	fast-paced interactions	23	19	≈400	≈400
Socio lab	interviews with adults	16	12	≈600	≈600
AMI	meetings (2-spk)	804	93	1190	1137
VoxConverse	broadcast (2-spk)	44	31	280	520

During inference time, classifiers' outputs are thresholded at 0.5 to determine speech activities. Each training was run on a single GPU. The batch size was set to 64 or 32 with 100000 or 200000 minibatch updates of warm-up respectively. Following [13], the Adam optimizer [16] was used and scheduled with noam [17] reaching a maximum learning rate of 10^{-3}. For fine-tuning on a development set, the Adam optimizer was used with learning rate 10^{-5}. For the inference as well as for obtaining the model from which to fine-tune or adapt, the models from the last 10 epochs were averaged. Unless otherwise specified, during training, adaptation and fine-tuning, batches were formed by sequences of 500 Mel-filterbank outputs, corresponding to 50 s. During inference, the full recordings are fed to the network one at a time. Diarization performance is evaluated in terms of diarization error rate (DER) as defined by NIST [18]. For evaluation sets where a forgiveness collar is used when calculating DER, a median filter with window 11 is applied as post-processing over the network's output. If the forgiveness collar is 0 s, no filtering is applied as this provides better definition in the output.

4.2. Evaluation sets for wide-band experiments

Different collections were used for evaluation. For the wide-band scenario, only two-speaker recordings were used. We evaluate results on the four domains from the Third DIHARD Challenge [19] that satisfy such condition: CTS, Clinical, Maptask and Sociolinguistic (lab). More information can be found in [19] and in Table 1.

Another domain of interest for diarization is meetings. However, there are no datasets in a meeting-like scenario with only two speakers. Given that many end-to-end diarization works still focus on the two-speaker scenario, we considered of relevance to derive from AMI meetings [20] all possible subsets of two speakers for each conversation and make it of public access[5]. For each recording, all pairs of speakers were drawn and, for each pair, all speech where another speaker spoke was removed from the waveforms using reference diarization annotations of the "only words" setup described in [21][6]. Then, for each original conversation with four speakers, six *conversations* of two speakers were created. We evaluate results on Mix-Headset audios and the beamformed microphone array N1, where BeamformIt [22] is applied using the specific AMI setup.

Finally, to add more diversity, recordings with two speakers from VoxConverse [23] were used as these come from varied broadcast sources and present different characteristics from those covered in previously mentioned sets. Following its corresponding evaluation setup, a forgiveness collar of 0.25 s was used to compute DER while all the sets mentioned above were scored with forgiveness collar 0 s. In all cases, all speech (including overlap) was evaluated.

[5]https://github.com/BUTSpeechFIT/AMI_2speaker_subset

[6]https://github.com/BUTSpeechFIT/AMI-diarization-setup

4.3. Evaluation sets for multi-speaker experiments

To evaluate the performance when training EEND-EDA with more than two speakers, the 2000 NIST Speaker Recognition Evaluation [24] dataset, usually referred as "Callhome" [25] was used. We report results using the standard Callhome partition[7]. We will refer to the parts as CH1 and CH2. The amounts of recordings for CH1/CH2 per amount of speakers in the recording are: with 2 speakers 155/148, with 3 61/74, with 4 23/20, with 5 5/5, with 6 3/3, with 7 2/0. Results on Callhome consider all speech (including overlap segments) for evaluation with a forgiveness collar of 0.25 s. Fine-tuning was done using CH1, and evaluation on CH2.

5. RESULTS

5.1. Wide-band experiments

One of the main aspects when generating synthetic training data is the choice of recordings to use. Figure 1 presents a comparison of SM and SC using different sets, namely telephone conversations, LibriSpeech downsampled to 8 kHz (mimicking a telephone channel scenario), LibriSpeech, VoxCeleb2 and VoxPopuli. 16 kHz evaluation data were downsampled to run the inference with 8 kHz models and 8 kHz data were upsampled (with empty upper spectrum) to run the inference with 16 kHz models.

Even though some clear differences exist before fine-tuning, such as SC LibriSpeech performing the best for Maptask and both AMI sets, differences are reduced after fine-tuning. As observed in [2], differences of up to 0.5 DER can easily be attributed to chance; thus, here there are no statistically significant "winners" in most cases. Some of the patterns worth mentioning are that 8 kHz models perform the worst for Maptask and the AMI sets, suggesting that information on higher frequencies is more relevant in these sets; and that using VoxCeleb2 recordings allows for better performance in VoxConverse, probably due to the similarities between both sets, formed by diverse recordings collected from YouTube.

Unlike the effect seen in the telephony scenario where SC are clearly superior and the need for fine-tuning is reduced substantially[8] [2], when working with different wide-band datasets, fine-tuning still plays a major role and even SM allow for similar performance. Unfortunately, this shows that the main challenge in wide-band scenarios is, until now, not the realism or naturalness of the synthetic data in terms of turns but rather differences in the channel between source and test data, quality of speaker annotation, and conversationality or a combination of all these.

The mechanism proposed for generating SC makes use of statistics about pauses and overlaps observed in real conversations. We explored different sets for extracting such statistics: CTS, Maptask, Socio lab, Clinical, all the aforementioned together (in the core setup, to have them equally represented), AMI and VoxConverse generating SC using audios from LibriSpeech but there were no significant differences. While these findings do not imply that it is not possible to produce synthetic data of superior quality that will permit better performance without the need for fine-tuning, we intend to share our results with the community simply to shed light on the directions that have already been explored. There are still other aspects to consider such as improving data augmentation mechanisms like using more realistic background noises, reverberation or loudness levels that match the application scenarios. For instance, CTS, Maptask and AMI H have speech recorded with close-talk

[7]https://github.com/BUTSpeechFIT/CALLHOME_sublists

[8]Results on CTS when training with SM telephone are 4.9% and 5.2% relatively worse than when training with SC telephone before and after fine-tuning respectively.

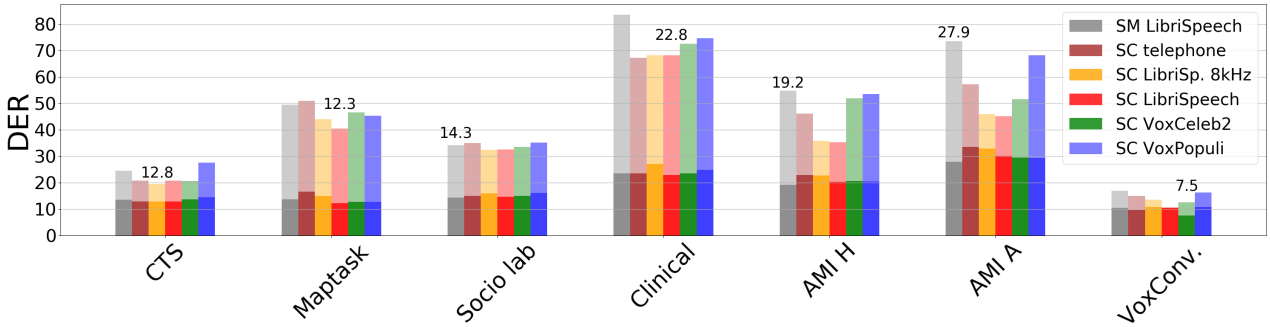

Fig. 1. Comparison of type of recordings used to generate SC (or SM): LibriSpeech, telephone (Switchboard and SRE), VoxCeleb2 and VoxPopuli. Results on the evaluation/test sets of different subsets of DIHARD3, 2-speaker AMI mix-headset (H) and beamformed array (A) and 2-speaker recordings from VoxConverse. Light shade colors show results before fine-tuning and darker colors refer to results after fine-tuning to a matching development set. Fine-tuning for AMI was for 20 epochs on the train set. Numbers mark the best result among bars.

Table 2. DER comparison on CH2. SM 2 spk and SC 2 spk contain ≈2500 h, **2**-4 spk contains ≈2500 h (1250 h recordings with 2 speakers, 625 h with 3 and 4 each), 2-**4** spk contains ≈2500 h (1250 h recordings with 4 speakers, 625 h with 2 and 3 each). 2-7 spk contains ≈2500 h with recordings following the proportions of speakers seen in CH1. SM 1-4 spk contains ≈15500 h (100k mixtures of each amount of speakers) and SC contains ≈10000 h (2500 h of each amount of speakers). Sequences of 200 s were used for fine-tuning.

System	All	2-spk	3-spk	4-spk	5-spk	6-spk
SM 2 spk	28.67	16.85	27.46	40.4	52.94	50.51
+ SM 1-4 spk	26.14	16.28	24.67	34	50.15	53.24
+ CH1	17.45	8.38	16.14	24.26	36.75	46.79
SM 1-4 spk	27	16.1	25.44	37.56	46.52	54.92
+ CH1	25.78	13.92	24.77	34.62	43.77	66.37
SC 2 spk	20.86	8.48	21.07	29.56	45.61	49.2
+ SC 1-4 spk	16.18	8.95	13.78	21.22	37.35	46.32
+ CH1	16.07	10.03	14.35	19.3	30.67	46.94
+ SC **2**-4 spk	17.52	9.07	15.11	23.3	38.55	54.03
+ SC 2-**4** spk	17.09	9.55	14.69	22.5	34.96	51.46
+ SC 2-7 spk	17.49	9.16	15.43	24.18	39.17	45.41
SC 1-4 spk	19.9	10.2	17.79	26.49	42.67	58.09
+ CH1	21.24	15.45	18.89	25.17	38.16	49.54

microphones, and Socio lab, Clinical and AMI A were recorded with far-field microphones. These aspects were not particularly considered in our analysis but they might have a strong influence[9].

Finally, we hope that by sharing recipes[10] for generating training data using public datasets and the models trained on publicly available data, the community will have easier access to baselines that otherwise require expensive computations.

5.2. Multi-speaker experiments

Table 2 presents a comparison when different sets are used for training, adapting (to more speakers) and fine-tuning (to CH1). The first rows show the performance when using SM. When following the same scheme with SC, the model trained on 2 speakers is already considerably better but adapting to more speakers pushes the performance further. The fine-tuning step provides only small gains, reducing the dependence on this step. Analogously, if the model is directly trained on the 1-4 sets, SC allow for much better performance than SM. However, this set is so biased towards fewer speakers that the model cannot learn from the fine-tuning.

We explored using other SC multi-speaker sets with different proportions of recordings. Training on a set with higher proportion of recordings with 4 than 2 speakers (+SC 2-**4** spk) is beneficial since the resulting model can deal better with recordings with more speakers. Training with a set that follows the same proportion of speakers seen in CH1 (+SC 2-7 spk) does not bring considerable gains. However, it should be noted that the training set is rather small and it is possible that learning to deal with more speakers requires a larger training set. Overall, the best results are obtained when using the 1-4 set for adaptation which is considerably larger than the other ones. This suggests that it would be beneficial to produce training data on-the-fly in order to encourage larger variability in the training set.

Table 3. Error comparison on CH2 when using auxiliary VAD loss.

System	L_{VAD}	DER	Miss	FA	Conf.
SC 2 spk + SC 2-**4** spk	×	17.09	7.32	4.12	5.66
+ CH1	×	16.27	9.55	2.55	4.16
SC 2 spk + SC 2-**4** spk	✓	16.91	5.88	5.26	5.77
+ CH1	✓	16.07	9.06	2.88	4.13

Table 3 presents results when using the additional VAD loss described in Section 3 (α=0.2) when adapting the SC 2 spk model to 2-**4** spk (see Table 2) and doing fine-tuning. The use of the auxiliary loss allows for more even levels of missed and false alarm speech and this permits the fine-tuning to improve further, reaching the performance obtained when using the larger 1-4 spk set.

6. CONCLUSIONS

Recently proposed simulated conversations as training data for EEND have shown remarkable performance with respect to the original simulated mixtures in 2-speaker telephone conversations. In this work, we extended the approach for conversations with more speakers and have shown that the same trend holds. We have also generated simulated conversations with different wide-band datasets in order to have models suited to non-telephone scenarios. However, the results were not conclusive showing that many challenges remain in order to generate adequate wide-band training data.

[9]We generated SC from specific domains, i.e., from Maptask recordings only, and evaluated the performance on the same domain but the performance was considerably inferior; probably due to the limited amount of speakers.

[10]https://github.com/BUTSpeechFIT/EEND_dataprep

7. REFERENCES

[1] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, et al., "End-to-end neural speaker diarization with permutation-free objectives," *Interspeech*, 2019.

[2] Federico Landini, Alicia Lozano-Diez, Mireia Diez, and Lukáš Burget, "From Simulated Mixtures to Simulated Conversations as Training Data for End-to-End Neural Diarization," *Interspeech*, 2022.

[3] Natsuo Yamashita, Shota Horiguchi, and Takeshi Homma, "Improving the Naturalness of Simulated Conversations for End-to-End Neural Diarization," *Odyssey*, 2022.

[4] Tsun-Yat Leung and Lahiru Samarakoon, "End-to-End Speaker Diarization System for the Third DIHARD Challenge System Description," 2021.

[5] Yi Chieh Liu, Eunjung Han, Chul Lee, and Andreas Stolcke, "End-to-end neural diarization: From transformer to conformer," *Interspeech*, 2021.

[6] Keisuke Kinoshita, Marc Delcroix, and Naohiro Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *ICASSP*. IEEE, 2021.

[7] Soumi Maiti, Hakan Erdogan, Kevin Wilson, et al., "End-to-end diarization for variable number of speakers with local-global networks and discriminative speaker embeddings," in *ICASSP*. IEEE, 2021.

[8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015.

[9] Guoguo Chen, Shuzhou Chai, Guanbo Wang, et al., "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *Interspeech*, 2021.

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech*, 2018.

[11] Changhan Wang, Morgane Riviere, Ann Lee, et al., "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," in *ACL — IJCNLP*. 2021, Association for Computational Linguistics.

[12] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[13] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, et al., "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," *Interspeech*, Oct 2020.

[14] Qingjian Lin, Tingle Li, Lin Yang, et al., "Optimal Mapping Loss: A Faster Loss for End-to-End Speaker Diarization.," in *Odyssey*, 2020.

[15] Naijun Zheng, Na Li, Xixin Wu, et al., "The CUHK-TENCENT speaker diarization system for the ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP*. IEEE, 2022.

[16] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] "NIST Rich Transcription Evaluations," `https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation`, version: md-eval-v22.pl.

[19] Neville Ryant, Prachi Singh, Venkat Krishnamohan, et al., "The third DIHARD diarization challenge," *Interspeech*, 2021.

[20] Jean Carletta, Simone Ashby, Sebastien Bourban, et al., "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2006.

[21] Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, 2022.

[22] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, 2007.

[23] Joon Son Chung, Jaesung Huh, Arsha Nagrani, et al., "Spot the conversation: speaker diarisation in the wild," *Interspeech*, 2020.

[24] Mark Przybocki and Alvin Martin, "NIST Speaker Recognition Evaluation LDC2001S97," *Philadelphia, New Jersey: Linguistic Data Consortium*, 2001.

[25] "NIST SRE 2000 Evaluation Plan," `https://www.nist.gov/sites/default/files/documents/2017/09/26/spk-2000-plan-v1.0.htm_.pdf`.