Data generation for speaker diarization by speaker transition information

Keigo Ichikawa, Sei Ueno, Akinobu Lee
Nagoya Institute of Technology, Japan
E-mail: k.ichikawa.667@stn.nitech.ac.jp, {sei.ueno, ri}@nitech.ac.jp

***Abstract*—A speaker-turn-aware training data generation method for end-to-end neural diarization (EEND) has been studied. To compensate the lack of real multi-talker training corpus, using artificial multi-speaker conversation data generated from several single speaker's speech has been a promising way to train EEND models. It is still expected that making the training data close to real multi-talker situation can improve EEND performance. In order to generate a data that contributes to the better performance of trained EEND model, we investigated two speaker alternation handling methods. One is a data-dependent model that introduces a probabiliry of speaker transition extracted from multi-talker data. The other one is a data-independent model that uniformly samples the speaker-alternation probability. Experimental results have showed that the data-dependent method improves the DER as compared with baseline simulated conversations (SC) method on 3-speaker and 4-speaker scenarios. Moreover, it was also confirmed that the data-independent method can achieve comparable performance on the best model on 3 and 4-speaker scenarios.**

I. Introduction

Speaker diarization is a task that detects "who spoke when" from multi-talker audio. Speaker diarization outputs are used in the preprocessing step for multi-talker tasks such as source separation and multi-talker speech recognition [1]. In order to perform speaker diarization, a clustering-based model or an end-to-end model can be used. A clustering-based model [2]–[7] is a model that performs a cascade process. In general, it performs speech activity detection, segmentation, speaker feature extraction, and clustering to distinguish the speaker of the input audio.

On the other hand, an end-to-end neural diarization (EEND) model has been investigated [8]–[10]. It directly predicts whether each speaker is talking in a single step, frame by frame. Due to the development of the EEND model, it outperforms the clustering-based model. However, for training the EEND model, a lot of annotated multi-talker data is required. Moreover, it is too costly to prepare the annotated data since we need to label when and how long each person speaks. Therefore, the data generation from the speech data of a single speaker is generally applied for training the EEND model at a lower cost [8], [11]. The simulated data are generated as follows: (1) Preparing source data containing multiple speakers' speeches with each utterance involving only a single speaker's speech. (2) Choosing 2, 3, or 4 speakers and extracting utterances of the chosen speaker from source data. (3) Generating multi-talker speech through concatenating the utterances.

The generated data needs to resemble real data to improve EEND performance. Several approaches have investigated which elements should closely match real data and how to simulate it. Simulated mixtures (SM) [8], [11], used in the original EEND model, considered silence intervals between each speaker's utterances. Simulated conversations (SC) [12], [13] extended SM method and generated multi-talker speech using statistics about distributions of pauses and overlaps estimated on real conversations.

In this work, we also investigate the elements to improve the EEND model and focus on speaker alternation in the generated data. Speaker alternation occurs when one speaker finishes and another speaker begins to speak. In a real conversation, one speaker may speak a lot, while others speak less. To model this conversation, we introduce speaker transition probabilities. Speaker transition probabilities involve not only speaker alternation but also a case where one speaker begins to speak again. The SC method implicitly models the speaker transition probability, but it depends on the source data. We first remove the dependency and then investigate two approaches. One is a data-dependent approach that uses the probabilities extracted from statistics on real conversations. The other is a data-independent approach that uses uniform probabilities.

II. Related works

A. Simulated mixtures (SM)

The SM method [8], [11] is a data generation method for creating multi-talker audio. It involves overlapping the single-speaker audio to create multi-talker audio. Single-speaker audio is created by preparing speech segments from that speaker and concatenating the prepared speech segments with a certain silent interval between them. The silent intervals are determined by random numbers that follow an exponential probability distribution.

By varying the parameters of the exponential distribution, it is possible to adjust the silent intervals, and the overlapping intervals indirectly when multi-talker audio from multiple single-speaker speech is overlapped. It is known that the way overlapping sections are created with SM method may generate data that differs from actual multi-talker audio, the overlapping sections continue for a long time or there are many sections where all speakers overlap, which is not often seen in real conversations.

B. *Simulated conversations (SC)*

SC method [12], [13] creates multi-talker audio by connecting speech segments from multiple speakers, creating overlapping and silent sections. When connecting speech segments from the same speaker, a certain interval of silence between them is made. When connecting speech segments from different speakers, a certain interval of silence or overlaps between them is made. The length of this overlapping section or silent section is determined by randomly selecting one value from the silent section lengths and overlapping section lengths collected from actual audio data. The choice of whether to create silent intervals or overlapping intervals when different speakers are speaking successively is also determined based on the distribution obtained from actual speech data. The order in which these speakers speak is randomized in simulated conversations.

In simulated mixtures, single-speaker audio is created first, then multi-talker audio is created, but in simulated conversations, multi-talker audio is created from the beginning. This makes it possible to control silent sections and overlapping sections in multi-talker audio more directly. Simulated conversations suppress the generation of speech that is not often seen in actual conversations, such as when there are long overlapping sections or when there are many sections where all speakers overlap. By making the simulated data closer to real data, the performance of EEND is improved.

III. Speaker Transition Probability

The frequency of speaker alternation depends on the domain of a real conversation. For example, speaker alternation in telephone communication is less than in standard face-to-face conversation. We expect that controlling speaker alternation yields an improvement in the EEND model. To model the speaker alternation, we introduce the speaker transition probability. We additionally consider the last speaker and the next speaker, and the probability $P(\text{next speaker}|\text{last speaker})$ of speaker alternation occurring using a first-order Markov model. We also consider P when the next speaker is the same as the last speaker. In total, there are S^2 patterns, where S denotes the number of speakers in one audio. Fig. 1 shows an example in a 3-speaker scenario. When the last speaker is speaker A, we need to calculate $P(A|A)$, $P(B|A)$, and $P(C|A)$. We follow the SC method to model other aspects such as overlap and silence ratios.

A. *Speaker transition probability on other methods*

The speaker transition probability is not determined in the SM method because it involves concatenating speeches from a single speaker and then from multiple speakers. In the SC method, the speaker transition probability is implicitly decided by the source data. It prepares all utterances of the selected speaker and randomly samples from the candidates. Therefore, the speaker transition probability depends on the number of each speaker's candidates of source data. Specifically, in 3-speaker scenario, we consider $P_{SC}(A)$, $P_{SC}(B)$, and $P_{SC}(C)$.

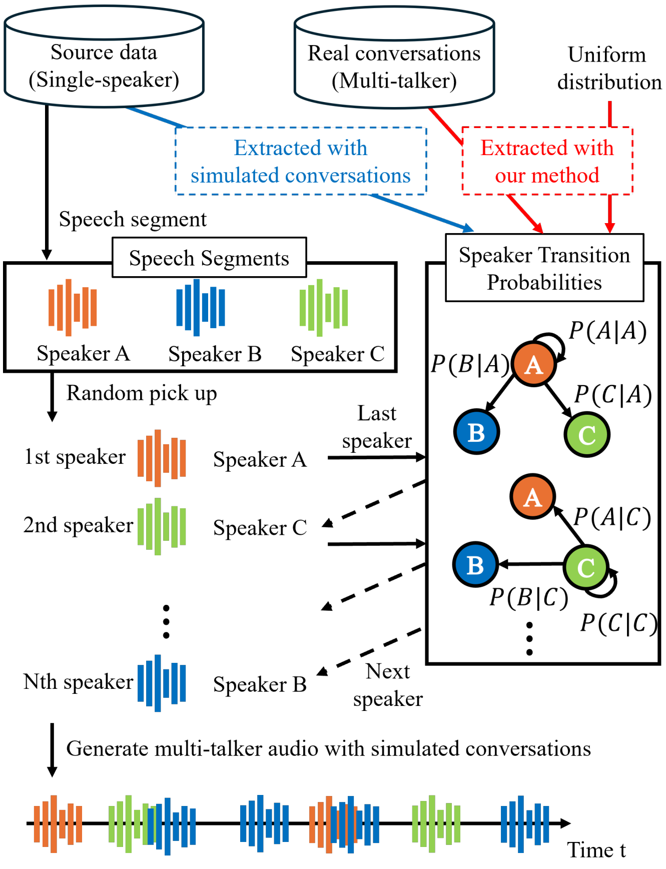

Fig. 1. Overview: Data generation of multi-talker audio using speaker transition probability. Simulated conversations extract speaker transition probabilities from source data. Our method extracts speaker transition probabilities from real conversations or specifies them as equal probabilities.

Each probability is determined as follows:

$$P_{SC}(A) = m_a/M, P_{SC}(B) = m_b/M, P_{SC}(C) = m_c/M$$

where, m_* denotes the number of candidates extracted from speaker-$*$ speeches, and $M = m_a + m_b + m_c$. Because each speech is used only once, $m_* = m_* - 1$ once speaker-$*$ is sampled. The number of candidates is based on the source data. Occasionally, P_{SC} can be biased, which is not expected.

B. *Data-dependent approach*

As a data-dependent approach, we extract the probability from the real multi-talker data. We calculate the speaker transition probability for each conversation using the multi-talker data. We prepare lists of speaker transition probabilities. In the data generation step, we randomly choose one of the lists and generate a multi-talker speech following the extracted speaker transition probability. To identify speakers A, B, ..., and N, we follow the order in which each speaker speaks in the real conversation.

C. *Data-independent approach*

As a data-independent approach, we use a uniform probability. All probabilities are $1/N_{spk}$. In the data-dependent

TABLE I
COMPARISON OF THE NUMBER OF AUDIO FILES FROM REAL CONVERSATIONS DATASET.

Dataset	2-speaker	3-speaker	4-speaker
DIHARD 3	157	16	16
CALLHOME 2	148	74	20
AMI	0	4	129
CALLHOME 1 (oracle)	155	61	23

approach, we need to prepare real conversations to extract the speaker transition probability. However, the public dataset provides a limited number of utterances (detailed in TABLE I), which may cause over-fitting. Therefore, we remove the dependency on speaker alternation and analyze how this affects the performance of the EEND model. This data-independent method helps remove the dependency on source data in SC methods.

IV. EXPERIMENTAL EVALUATION

A. Dataset and speaker transition probabilities

For data generation, we used LibriSpeech [14]. LibriSpeech is a speech corpus consisting of 961 hours of single-speaker recordings from 2,338 speakers. Statistical information to control overlap and silence ratio was obtained from the DIHARD 3 dev set [15]. The DIHARD 3 dev set is a speech corpus consisting of 34 hours of multi-talker conversations involving 548 speakers. For the evaluation, we used the 2, 3, and 4-speaker audio from CALLHOME 1, a subset of the 2000 NIST Speaker Recognition Evaluation (CALLHOME) corpus [16]. We divided CALLHOME into CALLHOME 1 and CALLHOME 2 using the Kaldi toolkit [17]. The duration of CALLHOME 1 and CALLHOME 2 are as follows: 2-speaker audio has 3.2 hours and 3.0 hours, 3-speaker audio has 2.1 hours and 3.0 hours, and 4-speaker audio has 2.2 hours and 1.8 hours, respectively. All audio was used at 8kHz.

In the data-dependent approach, we extracted the speaker transition probability from the DIHARD 3 dev set, CALLHOME 1/2, and AMI train set [18]. CALLHOME 1 was used as an oracle. TABLE I shows the number of audio files for each dataset. For the 2-speaker, 3-speaker, and 4-speaker scenarios, DIHARD 3 and CALLHOME 1 and 2 were used. To prevent overfitting, we only used datasets containing more than 30 audio files, and we mixed the datasets when the dataset provided a limited number of audio files. In the data-independent setting, the value of each speaker transition probability was set to 0.25 for the 4-speaker scenario, 0.33 for the 3-speaker scenario, and 0.50 for the 2-speaker scenario as equal probabilities.

B. End-to-end neural diarization

EEND [8], [11] is a model approach that achieves end-to-end speaker diarization by framing the task as a multi-label classification problem, which classifies the presence of multiple speakers within each frame. This model usually uses a transformer encoder, and given acoustic features of length T with D dimensions $(\boldsymbol{x}_t \mid \boldsymbol{x}_t \in \mathbb{R}^D)_{t=1}^T$, the EEND model generates outputs as described in Equation (1).

$$\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T = \mathrm{f}_{\mathrm{EEND}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) \tag{1}$$

Here, $\boldsymbol{y}_t = [y_{t,1}, \cdots, y_{t,S}]^T \in (0,1)^S$ denotes the probability of speech presence for S speakers in frame t.

C. Experimental settings

In all settings, we generated 2,480.6 hours of audio. We generated audio according to the number of speakers in the evaluation set. Impulse responses were not used, but background noises obtained from the MUSAN corpus [19] were used, and the SNR was set to one of 5, 10, 15, or 20 dB. The input features for the EEND model were 23-dimensional log Mel-filterbanks with a frame length of 25 milliseconds and a frameshift of 10 milliseconds. The EEND model consists of four transformer encoder blocks, each with 4 heads and 256 attention units. The dropout rate was also set to 0.1, and the position-wise feedforward layer had 1024 internal units. The optimizer used was Adam with a noam scheduler, and the warmup period was set to 25,000 steps. The batch size during training was 64, and the model was trained for 100 epochs. After 100 epochs of training, averaging was performed on the models from epochs 91 to 100, and this averaged model was used for evaluation.

During inference, the threshold for outputting speech segments was set to 0.5, and a median filter of size 12 frames was applied before determining the speech segments. The diarization error rate (DER) defined by NIST [20] was used as the evaluation metric. DER is broken down into missed speech (Miss), false alarm (False), and speaker confusion (Conf.). During evaluation, errors of less than 250 milliseconds were tolerated.

V. RESULTS AND DISCUSSION

A. Comparison of end-to-end neural diarization performance with 2 and 3-speaker

TABLE II shows the DER for 2-speaker and 3-speaker settings and TABLE III also shows speaker alternation occurrence rates. In the 2-speaker setting, the data-dependent approach using CALLHOME 2 achieved comparable performance compared with the SC method. While all methods improved missed speech, they degraded false alarm and speaker confusion compared with the SC method. All methods caused more speaker alternations than the SC method as shown in TABLE III. We consider that it positively affects missed speech. However, it worsens false alarm and speaker confusion because the 2-speaker scenario is a relatively easy task.

In the 3-speaker setting, we observed that our methods yielded improvement in CALLHOME 2 + DIHARD 3. Different from the 2-speaker scenario, our methods improved speaker confusion. Because the number of speakers increased, detecting speakers in the 3-speaker scenario is a harder task than the 2-speaker scenario. However, the data-dependent approach using CALLHOME 2 worsened the DER. It drastically improved the false alarm, but degraded missed speech.

TABLE II
DER (%) FOR 2-SPEAKER AND 3-SPEAKER SETTING. DER IS DECOMPOSED INTO MISS (MISSED SPEECH), FALSE (FALSE ALARM), AND CONF. (SPEAKER CONFUSION).

#spk	Speaker transition	Miss	False	Conf.	DER
2	SC [baseline]	8.2	3.6	4.4	16.2
	Data-dependent				
	DIHARD 3	7.4	5.4	6.3	19.0
	CALLHOME 2	6.5	5.0	4.6	**16.1**
	CALLHOME 2 + DIHARD 3	7.5	5.3	5.2	17.9
	CALLHOME 1 (oracle)	7.7	3.4	5.8	16.9
	Data-independent (uniform)	7.1	5.0	5.6	17.7
3	SC [baseline]	11.2	6.1	12.5	29.7
	Data-dependent				
	CALLHOME 2	17.4	2.9	12.2	32.6
	CALLHOME 2 + DIHARD 3	10.9	6.5	11.6	**29.1**
	CALLHOME 1 (oracle)	11.8	5.8	11.0	**28.6**
	Data-independent (uniform)	10.8	6.8	10.8	**28.5**

TABLE III
COMPARISON OF SPEAKER ALTERNATION OCCURRENCE RATES (%) IN GENERATED DATA USING DIFFERENT SPEAKER TRANSITION PROBABILITIES FOR EACH NUMBER OF SPEAKERS

#spk Speaker transition	2	3	4
SC [baseline]	45.9	63.0	71.7
Data-dependent			
DIHARD 3	58.8	-	-
CALLHOME 2	75.5	81.4	-
CALLHOME 2 + DIHARD 3	66.9	77.3	74.7
AMI	-	-	76.4
AMI + CALLHOME 2	-	-	77.3
AMI + DIHARD 3	-	-	75.5
AMI + CALLHOME 2 + DIHARD 3	-	-	76.3
CALLHOME 1 (oracle)	78.9	81.8	83.1
Data-independent (uniform)	50.1	66.8	75.1
CALLHOME 1 (eval)	81.6	83.1	83.3

The speaker alternation occurrence rate was similar to the evaluation data, but the limited number of samples may have led to overfitting. In the CALLHOME 2 + DIHARD 3 setting, the overfitting is alleviated. The data-independent approach also improved the performance compared with the SC method and achieved comparable performance to the data-dependent approach using CALLHOME 1 (oracle). This result implies that the balance of the number of the speaker's utterances is one of the keys to the EEND model to train the speaker transition.

B. Comparison of end-to-end neural diarization performance with 4-sepaker

TABLE IV shows the DER for 4-speaker setting. We observed that our methods yielded improvement in AMI and AMI + CALLHOME 2 + DIHARD 3. Similarly to the 3-speaker setting, our methods improved speaker confusion. We also observed that our method yielded improvement in CALLHOME 2 + DIHARD 3. Specifically, it did not improve speaker confusion, but it improved false alarm. Using AMI + CALLHOME 2 and AMI + DIHARD 3 improved false alarm compared to using AMI only. Similar to this, we consider that using CALLHOME 2 + DIHARD 3 improved false alarm. But using AMI + CALLHOME 2 and AMI + DIHARD 3 degraded DER. Because the number of AMI is larger, the dataset ratio bias in AMI + CALLHOME 2 and AMI + DIHARD 3 is bigger than in CALLHOME 2 + DIHARD 3 (2-speaker and 3-speaker). We consider that the dataset ratio bias negatively impacts performance.

TABLE IV
DER (%) FOR 4-SPEAKER SETTING. DER IS DECOMPOSED INTO MISS (MISSED SPEECH), FALSE (FALSE ALARM), AND CONF. (SPEAKER CONFUSION).

Speaker transition	Miss	False	Conf.	DER
SC [baseline]	8.2	7.2	17.8	33.2
Data-dependent				
AMI	8.5	7.6	16.4	**32.4**
AMI + CALLHOME 2	8.7	6.4	18.9	34.0
AMI + DIHARD 3	9.2	7.0	17.5	33.7
CALLHOME 2 + DIHARD 3	8.8	6.3	17.5	**32.7**
AMI + CALLHOME 2 + DIHARD 3	9.0	7.4	16.1	**32.4**
CALLHOME 1 (oracle)	7.7	8.2	17.6	33.5
Data-independent (uniform)	8.3	7.0	17.0	**32.3**

The data-dependent approach using CALLHOME1 (oracle) worsened the DER. The speaker alternation occurrence rate was most similar to the evaluation data in the 4-speaker scenario dataset, but the limited number of samples may lead to overfitting as well as using CALLHOME2 in the 3-speaker scenario.

The data-independent approach improved the performance compared SC method and achieved comparable performance to the data-dependent approach using AMI + CALLHOME 2 + DIHARD 3.

C. Discussion of data-dependency in speaker transition

In any speaker scenario, while the best DER of the data-dependent approach was better, the DERs of some data-dependent approaches were worse than the baseline method. This is because the speaker transition highly depends on the dataset. On the other hand, data-independent approaches are more robust than data-dependent methods and achieve the best performance in 3 and 4-speaker settings. The data-dependent approach may improve further, but we need to prepare more data of multi-talker data. We consider that the speaker transition itself can affect the DER performance and can be seen as a hyperparameter of the data generation.

VI. CONCLUSIONS

In this study, we extended the method of determining speaker alternation in simulated conversations, which was originally based on the speaker transition probabilities from source data, to use speaker transition probabilities from real conversational audio and uniform distribution. In the 3-speaker and 4-speaker settings, we achieved DER improvements through significant reductions in speaker confusion. Furthermore, in scenarios where our method is effective, it suggests that the data-independent approach can achieve performance improvements comparable to the data-dependent approach.

Future work will involve investigating whether similar trends can be observed using the same method on evaluation corpora with a larger number of speakers and different tasks. Additionally, we will continue to explore the effects of introducing deeper speaker transition probabilities, such as those modeled by a second-order Markov process.

References

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101 317, 2022.

[2] 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, 2013.

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

[4] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 4930–4934.

[5] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6301–6305.

[6] Q. Li, F. L. Kreyssig, C. Zhang, and P. C. Woodland, "Discriminative neural clustering for speaker diarisation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 574–581.

[7] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, p. 101 254, 2022.

[8] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[9] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022.

[10] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7198–7202.

[11] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 296–303.

[12] F. Landini, A. Lozano-Diez, M. Diez, and L. Burget, "From simulated mixtures to simulated conversations as training data for end-to-end neural diarization," *arXiv preprint arXiv:2204.00890*, 2022.

[13] F. Landini, M. Diez, A. Lozano-Diez, and L. Burget, "Multi-speaker and wide-band simulated conversations as training data for end-to-end neural diarization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[15] N. Ryant, P. Singh, V. Krishnamohan, *et al.*, "The third dihard diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[16] *2000 nist speaker recognition evaluation*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2001S97.

[17] *Kaldi callhome diarization*, Accessed: 2024-07-28, 2022. [Online]. Available: https://github.com/kaldi-asr/kaldi/tree/master/egs/callhome_diarization/v1.

[18] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The ami meeting corpus," in *Proc. International Conference on Methods and Techniques in Behavioral Research*, 2005, pp. 1–4.

[19] D. Snyder, G. Chen, and D. Povey, *MUSAN: A Music, Speech, and Noise Corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[20] NIST, *Rich transcription evaluation*, https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation, version: md-eval-v22.pl.