JOINT INFERENCE OF SPEAKER DIARIZATION AND ASR WITH MULTI-STAGE INFORMATION SHARING

Weiqing Wang[1], Danwei Cai[1], Ming Cheng[2], Ming Li[1,2,*]

[1]Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Duke Kunshan University, Kunshan, China

ABSTRACT

In this paper, we introduce a novel approach that unifies Automatic Speech Recognition (ASR) and speaker diarization in a cohesive framework. Utilizing the synergies between the two tasks, our method effectively extracts speaker-specific information from the lower layers of a pretrained Conformer-based ASR model while leveraging the higher layers for enhanced diarization performance. In particular, the integration of ASR contextual details into the diarization process has been demonstrated to be effective. Results on the DIHARD III dataset indicate that our approach achieves a Diarization Error Rate (DER) of 10.52%, which can be further reduced to 10.39% when integrating ASR features into the diarization model. These findings highlight the potential of our approach, suggesting competitive performance against other state-of-the-art systems. Additionally, our framework's ability to simultaneously generate text transcripts for each speaker marks a distinct advantage, which can further enhance ASR capabilities and transition towards an end-to-end multitask framework encompassing both ASR and speaker diarization.

Index Terms— Speaker diarization, automatic speech recognition, target-speaker voice activity detection

1. INTRODUCTION

Speaker diarization is the task of determining “who spoke when”, identifying both speaker identities and the timing of each speaker's presence. This task demands robust speaker representations coupled with contextual information to discriminate speaker identities at different timestamps. However, in many traditional clustering-based methods, this contextual information is often overlooked. In these approaches, speaker representations are typically evaluated using specific metrics and then clustered without any contextual cues.

Several studies have incorporated both speaker representations and contextual information into speaker diarization tasks. Landini et al. [1] introduced a diarization algorithm based on Bayesian Hidden Markov Models (HMM) to refine initial diarization results. Lin et al. [2] proposed an LSTM-based model to extract a similarity matrix, effectively incorporating contextual information. Subsequently, a target-speaker voice activity detection approach, also founded on an LSTM framework, was presented to further refine diarization results and recognize overlapping speech. Furthermore, the End-to-End Neural Diarization (EEND) [3, 4] approach, which has gained considerable popularity, uses LSTM or Transformers to directly compute speaker diarization results. Within this approach, contextual information can be implicitly learned during the training phase. However, these methodologies often overlook spoken content, which could facilitate a more contextual approach to speaker diarization.

Automatic Speech Recognition (ASR), which translates spoken language into text, has been effectively developed to enhance the performance of both speaker verification and speaker diarization. ASR's primary focus is on recognizing the linguistic content of speech, paying particular attention to frame-level details. For instance, frame-level phoneme modeling in ASR can enhance speaker verification by identifying distinctive, speaker-specific speech patterns. Previous research provides support for this collaboration, revealing that phoneme modeling improves speaker verification performance in speaker embedding networks [5, 6] as well as in the i-vector statistical model [7, 8].

Speaker diarization systems have also increasingly integrated with ASR techniques. For example, word alignment has been used to refine Speech Activity Detection (SAD) [9], while others have leveraged it to detect change points [10]. Subsequent methods incorporated lexical information for diarization, such as the text-based role recognizer [11] and segmentation using ASR outputs [12]. Furthermore, some studies have jointly optimized ASR and speaker diarization systems. In such configurations, either one system benefits from the other [13], or both systems can be improved from each other [14].

In this paper, we propose a joint inference method of speaker diarization and ASR, where we directly build a speaker diarization system from a pretrained Conformer-based ASR model to improve the diarization performance. Conformer encoders, initially designed for ASR, exhibit a natural versatility due to their layered structure, which enables them to grasp various aspects of speech. The lower layers of the ASR Conformer capture a range of speech characteristics, including speaker traits, language patterns, emotions, and phonetic nuances. In contrast, the upper layers focus more on phonetic and contextual details, aligning with ASR goals. Although their primary training focus is ASR, even the initial layers exhibit remarkable proficiency in speaker recognition [15]. This suggests that ASR trained features can be used for speaker verification effectively. However, the upper layers which conveys contextual information are ignored in transfer learning because the lack of speaker-related details may degrade the performance of speaker representation. Therefore, we follow the training protocol in [16]. Moreover, we further incorporate the output of the fixed upper layers of ASR conformer as auxiliary features for speaker diarization. In this paper, we adapt a sequence-to-sequence target-speaker voice activity detection (Seq2Seq-TSVAD) module [17] from a pretrained and fixed ASR Conformer and explore how the ASR pretraining model can improve the performance and save the parameters of speaker diarization during the joint inference.

[*] Corresponding author: Ming Li. Email: ming.li369@duke.edu

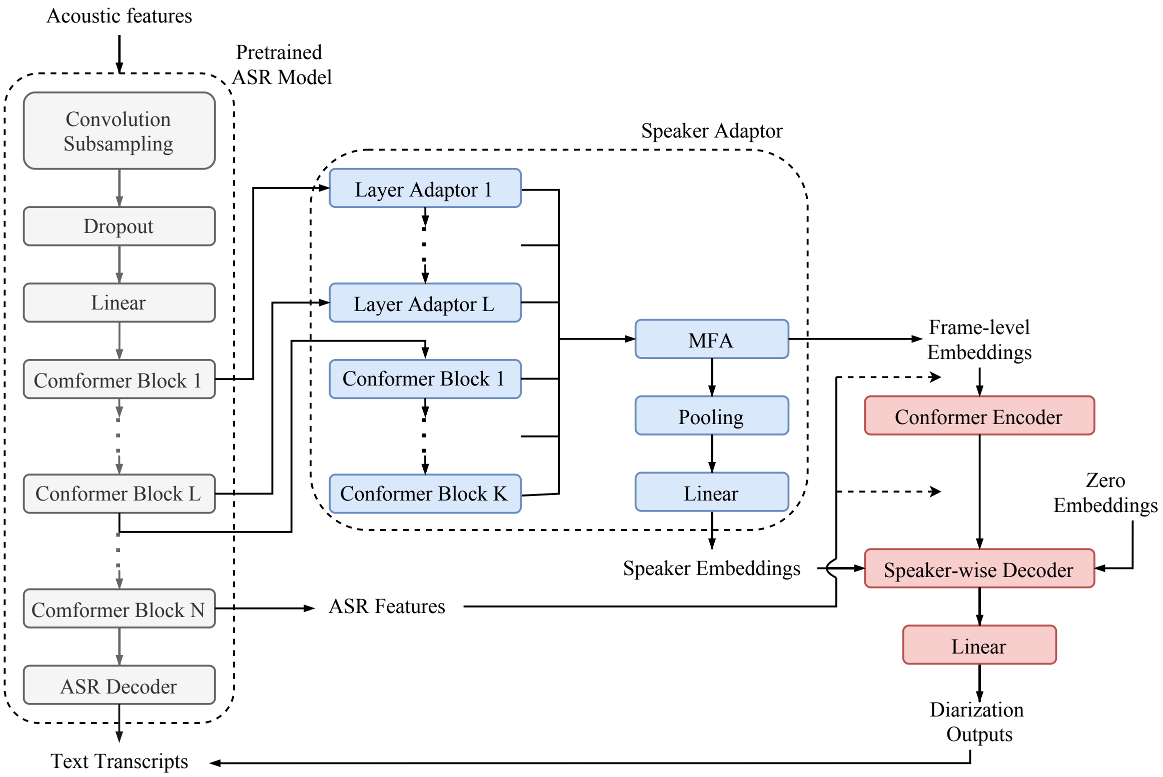

Fig. 1. The architecture of the proposed method, where the pretrained ASR model are frozen, and other parts are trainable.

2. RELATED WORKS

2.1. Unified ASR and speaker verification

In Figure 1, the architectures of the unified ASR and speaker adaptor are consistent with the design presented in [16]. On the left side, there is a frozen, pretrained Conformer-based ASR model, while on the right lies a trainable speaker adaptor (blue blocks) that interacts with the ASR model's intermediate representations. This adaptor contains three components: L layer adaptors, K trainable Conformer layers, and a fully connected layer preceded by a trainable pooling layer for extracting speaker embeddings. Each layer adaptor consists of two linear layers, interspersed with Layer Normalization and an activation function.

Outputs from the first L layers of the pretrained ASR Conformer encoder undergo dimensional reduction through the L layer adaptors and K trainable Conformer layers. Subsequently, the outputs from these adaptors and the trainable Conformer layers are concatenated using a multi-scale feature aggregation (MFA) module [18]. An attentive statistics pooling layer then generates an utterance-level speaker representation [19]. Notably, given that the parameters of the pretrained ASR remain frozen, the model can concurrently produce the text transcript.

2.2. Seq2Seq TSVAD

In conventional TSVAD systems [20], each target-speaker embedding is concatenated with the frame-level representation to determine the probability of the speaker's presence. However, this process can be resource-intensive, particularly in terms of GPU memory. As the number of target-speaker embeddings and frame sequences (T and N respectively) increase, there's a significant surge in memory consumption. This limits the model's capability to handle longer feature sequences and accommodate a larger number of speakers simultaneously. Another limitation is that the output length of models relying solely on encoders must match the input length, constraining the temporal resolution of output and making it inflexible.

To address these challenges, Cheng et al. [17] introduced a sequence-to-sequence framework for target-speaker voice activity detection, dubbed Seq2Seq TSVAD. Within this framework, frame-level representations and speaker embeddings are channeled separately into the encoder and decoder, eliminating the need for concatenation. A key advantage is that the decoder consolidates each speaker's voice activity data into an embedding with a fixed dimension, regardless of the input features' length. Another advantage is the flexibility offered by the final linear layer, which permits voice activity predictions at a higher temporal resolution, achieved with minimal computational overhead.

3. PROPOSED METHOD

3.1. Unified ASR and speaker diarization

Given the observed benefits of speaker verification using the pretrained Conformer-based ASR model, we believe that this model can also enhance speaker diarization performance. Both the speaker identity information from the lower layers and the contextual details from the upper layers play significant roles in speaker diarization.

Initially, we utilize only the lower layers to assess the efficacy of the pretrained Conformer-based ASR model in this context.

As depicted in Figure 1, the MFA module extracts frame-level speaker representations from the acoustic features. These outputs are subsequently fed to a standard Conformer Encoder to further model the long-term dependencies among frame-level representations. The Speaker-wise Decoder (SW-D) then discerns voice activities of the target speaker, taking cross-speaker correlations into consideration. The SW-D inputs encompass both decoder embeddings and auxiliary queries. Notably, these decoder embeddings are initialized to zeros, while the target-speaker embeddings serve as the auxiliary queries. Finally, a linear layer, followed by a sigmoid activation, projects the decoder output into posterior probabilities, indicating voice activities of each respective speaker.

3.2. Integration of contextual details into diarization

To enhance the speaker diarization performance, we also leverage the upper layers of the pretrained Conformer-based ASR model. More specifically, as illustrated in Figure 1, we utilize the output from the final ASR Conformer layer as the ASR features. Subsequently, this output is combined with the intermediate features of the Seq2Seq TSVAD model and undergoes Layer Normalization. To integrate the contextual details into speaker diarization, we combine the ASR features with the input of the Conformer Encoder or Decoder to let the model learn ASR details itself.

With the integration of ASR features, the performance of speaker diarization can be further improved as the contextual details are included.

4. EXPERIMENTAL SETUP

4.1. Pretrained Conformer-based ASR

We utilized the pretrained ASR Conformer-based models available in the NEMO toolkit [21]. Our preference for this model was guided by its commendable performance and versatility across a range of benchmark datasets. The NEMO ASR Conformer is available in three variants: small, medium, and large. For our research, we opted for the 'small' variant[1], characterized by a convolution subsampling rate of 14 and a uniform kernel size of 31 within its convolution modules. This version encompasses 16 Conformer layers with an encoder dimension of 176, accommodates 4 attention heads, and comprises 704 linear hidden units.

4.2. Unified ASR and speaker verification

This model was trained using the development set from VoxCeleb 2, comprising 1,092,009 audio recordings spanning 5,994 distinct speakers.

During the training phase, the entire pretrained ASR Conformer was frozen without updates, and we only focus on training the speaker adaptor. This component contains 4 layer adaptors with output dimension of 128, 2 streamlined Conformer layers with output dimension of 176, a pooling layer, and subsequent linear layers, all of which are optimized with respect to the speaker verification objective. Consequently, only the outputs from the initial 4 Conformer layers were utilized as inputs for the speaker adaptor. For this configuration, $L = 4$ and $K = 2$. Ultimately, the outputs from both the adaptors and trainable Conformer layers are concatenated, forming a feature sequence with a dimension of $128 \times 4 + 176 \times 2 = 864$. Following layer normalization and pooling, this sequence is transformed into an utterance-level speaker embedding with a size of 256. Comprehensive training details and hyper-parameters can be referred to in [16].

4.3. Unified ASR and speaker diarization

For speaker diarization, we utilize the pretrained ASR model and speaker adaptor as the front-end module, with the speaker adaptor pretrained on VoxCeleb 2. We adopt the feature sequence preceding the pooling layer as the frame-level speaker representation, which possesses a dimension of 864 for each frame. Subsequently, a Conformer Encoder refines this feature sequence, and the outputs with target-speaker embeddings and zero-initialized embeddings are fed to the Decoder. The Decoder's output is then converted into posterior probabilities representing voice activities for all speakers. All encoder-decoder components consist of 6 layers and maintain identical configurations: 512-dimensional attentions with 8 heads, and 1024-dimensional feed-forward layers with a dropout rate set at 0.1.

The training audio signals are segmented into 8-second chunks. These segments serve as input to the model, which utilizes 80-dimensional log Mel-filterbank energies with a frame length of 25 ms and a frame shift of 10 ms for acoustic features. Additionally, data augmentation is performed using background noise from Musan [22] and reverberation from RIRs [23].

During the training phase, we employ the BCE loss and the Adam optimizer with a linear learning rate warm-up strategy. Initially, the model with a frozen ASR model and frozen speaker adaptor is trained using simulated data created from Voxceleb 2 dataset [24] until convergence. Subsequently, the parameters of the speaker adaptor are unfrozen for training. Real data from the DIHARD III dataset [25] is then incorporated with the simulated data at a ratio of 0.2. Ultimately, the model undergoes fine-tuning exclusively with real data without any simulation. The initial two training stages encompass roughly 200 epochs with a learning rate set at 1e-4, while the final stage adjusts the learning rate to 5e-6.

During the inference stage, we utilize spectral clustering [2] to obtain the initial clustering-based results. Note that this step can be replaced by adopting an EEND model for initial clustering [26]. For evaluation purposes, each test recording is divided into 8-second segments. These are then input into the Seq2Seq TS-VAD model, accompanied by the target-speaker embeddings. We cap the number of speaker profiles at 10, padding with zero values for any additional speaker embeddings as required. To enhance the results, oracle voice activity detection (VAD) is employed during post-processing to retrieve missing frames. Comprehensive details on training and inference can be referenced in [17].

4.4. Integration of ASR features

We leverage the output from the final ASR Conformer layers to acquire contextual information. This output corresponds to the feature sequence preceding the ultimate linear layer responsible for token conversion, with a feature dimension of 176. The backbone architecture of the Seq2Seq TSVAD model remains unchanged, and the ASR feature is only utilized to assist the model in learning enhanced representations. This ASR feature sequence can be integrated either with the Encoder or the Decoder input. The combined input will be layer normalized before sending to Encoder or Decoder.

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_small

5. EXPERIMENTAL RESULTS AND DISCUSSION

For simplicity and efficiency, we utilize the small version of the unified ASR and speaker verification ($L = 4, K = 2$). This version achieves an EER of 0.98 on VoxCeleb1-O with a model size of 6.6M [16].

The unified ASR and speaker diarization models are evaluated on Track 1 of the DIHARD III dataset [25]. The initialization for clustering, which is used for target speaker embedding extraction, is based on the LSTM-SC method [2]. This method has a DER of 15.4% with Oracle VAD on the evaluation set.

Table 1 presents the results of the unified ASR and speaker diarization in comparison with other approaches. Without the aid of the ASR features, it achieves a DER of 10.52%, which is competitive with other state-of-the-art (SOTA) systems. It also offers the capability to produce text transcripts based on diarization results for each speaker. Moreover, by integrating the ASR features with either Encoder or Decoder input, the DER can be further lowered to 10.39% and 10.42%, respectively.

Table 1. DERs (%) on the DIHARD III Dataset.

Method	DER (%)	JER (%)
VBx [27]	16.54	37.82
Hitachi-JHU [28]	12.74	34.08
USTC-NELSLP [29]	12.41	-
ANSD-MA-MSE [30]	11.12	-
Seq2Seq TSVAD [17]	10.77	28.46
LSTM-SC	15.40	33.27
+ Unified ASR & SD	10.52	28.12
+ ASR feat before SD Encoder	**10.39**	27.97
+ ASR feat before SD Decoder	10.42	**27.85**

Actually, as Table 1 demonstrates, there isn't a significant overall improvement when utilizing ASR features from the last Conformer layer. We hypothesize that this is because certain domains are too complex for the model to effectively learn the contextual details. For instance, audio might contain excessive overlapping speech, or the signal-to-noise ratio may be too low. As indicated by Table 2, we assess performance across different domains. We can find out that there are performance improvement (sys2 vs sys1) in all domains expect the Restaurant and Socio field domains, which might be because the inaccurate ASR predictions.

It is worth noted that in some specific domains (e.g. Clinical), where the predicted text has more role information, the ASR feature can bring in more improvement.

Table 3 shows the number of the trainable parameters in the front-end model of Seq2Seq TSVAD. The unified ASR and SD system with a small front-end, with a mere tenth of the parameters of the one in original Seq2Seq TSVAD system, can produce competitive diarization results. This efficiency demonstrates the potential of the Speaker Adaptor in harnessing vital speaker information with fewer trainable parameters.

6. CONCLUSION

In this paper, we introduced a unified ASR and speaker diarization framework that can concurrently perform ASR and speaker diarization inference. Within the proposed framework, features from the lower layers of a pretrained ASR model are utilized to extract speaker-related information, while the upper layers provide ASR-related features that can enhance diarization performance. On the DIHARD III dataset, this framework demonstrates competitive performance when compared to other state-of-the-art systems. However, the DIHARD dataset, being a challenging compilation that contains substantial background noise and overlapping speech, can severely ruin the contextual information. As such, the benefits of integrating ASR features might be moderate on this dataset. Future research will involve evaluating our model on some domain specific datasets.

It's also noteworthy that the parameters of the ASR model remain frozen, complicating the joint training of ASR and speaker diarization. In the future, we plan to unfreeze the ASR model, leveraging diarization results to refine ASR performance, such as in multi-talker ASR or target speaker ASR scenarios. Furthermore, we believe that an end-to-end framework would be more apt for addressing the multitasking needs of ASR and speaker diarization, which will be a focus of our subsequent research.

Table 2. DERs (%) over 11 domains on the DIHARD III dataset. Sys 1, 2 and 3 refer to the last three rows in Table 1.

Domain	LSTM-SC	Sys 1	Sys 2	Sys 3
Audiobook	0.00	0.00	0.00	0.00
Broadcast	5.06	3.82	3.78	3.85
Clinical	7.59	5.01	4.63	4.61
Courtroom	4.10	2.06	2.04	2.18
CTS	14.46	6.01	5.89	5.85
Maptask	3.85	1.21	1.12	1.22
Meeting	28.70	22.95	22.54	23.16
Restaurant	42.30	38.97	39.36	39.25
Socio field	8.42	5.12	5.45	5.39
Socio lab	7.08	2.82	2.79	2.71
Webvideo	38.13	34.38	33.31	33.38
All	15.40	10.52	10.39	10.42

Table 3. The number of parameters for the speaker embedding extractor front-end in the Seq2Seq TSVAD and unified ASR and speaker diarization systems. The difference is that the trainable front-end of these two systems are ResNet34 and speaker adaptor, respectively. The number of the parameters of the first 4 layers of ASR model is 3.92M, which is borrowed from ASR and maintains unchanged.

System	Front-end	#Parameters
Seq2Seq TSVAD	ResNet34	20.56M
Unified ASR & SD	Speaker Adaptor	2.68M (+3.92M)

7. REFERENCES

[1] 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, pp. 101254, 2022.

[2] Qingjian Lin, Ruiqing Yin, Ming Li, Hervé Bredin, and Claude Barras, "LSTM Based Similarity Measurement with Spectral

Clustering for Speaker Diarization," in *Proc. of Interspeech*, 2019, pp. 366–370.

[3] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with self-attention," in *Proc. of ASRU*, 2019, pp. 296–303.

[4] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe, "End-to-End Neural Speaker Diarization with Permutation-Free Objectives," in *Proc. of Interspeech*, 2019, pp. 4300–4304.

[5] Tianyan Zhou, Yong Zhao, Jinyu Li, Yifan Gong, and Jian Wu, "Cnn with phonetic attention for text-independent speaker verification," in *Proc. of ASRU*, 2019, pp. 718–725.

[6] Shuai Wang, Johan Rohdin, Lukáš Burget, Oldřich Plchot, Yanmin Qian, Kai Yu, and Jan Černocký, "On the Usage of Phonetic Information for Text-Independent Speaker Embedding Extraction," in *Proc. of Interspeech*, 2019, pp. 1148–1152.

[7] Ming Li, Lun Liu, Weicheng Cai, and Wenbo Liu, "Generalized i-vector representation with phonetic tokenizations and tandem features for both text independent and text dependent speaker verification," *Journal of Signal Processing Systems*, vol. 82, pp. 207–215, 2016.

[8] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. of ICASSP*, 2014, pp. 1695–1699.

[9] Jing Huang, Etienne Marcheret, Karthik Visweswariah, and Gerasimos Potamianos, "The ibm rt07 evaluation systems for speaker diarization on lecture meetings," in *International Evaluation Workshop on Rich Transcription*, 2007, pp. 497–508.

[10] Wei Xia, Han Lu, Quan Wang, Anshuman Tripathi, Yiling Huang, Ignacio Lopez Moreno, and Hasim Sak, "Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection," in *Proc. of ICASSP*, 2022, pp. 8077–8081.

[11] Nikolaos Flemotomos, Panayiotis Georgiou, and Shrikanth Narayanan, "Linguistically aided speaker diarization using speaker role information," in *Proc. of Odyssey*, 2020, pp. 117–124.

[12] Tae Jin Park and Panayiotis Georgiou, "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks," in *Proc. of Interspeech*, 2018, pp. 1373–1377.

[13] Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison W. Cottrell, "Speech Recognition and Multi-Speaker Diarization of Long Conversations," in *Proc. of Interspeech*, 2020, pp. 691–695.

[14] Naoyuki Kanda, Zhong Meng, Liang Lu, Yashesh Gaur, Xiaofei Wang, Zhuo Chen, and Takuya Yoshioka, "Minimum bayes risk training for end-to-end speaker-attributed asr," in *Proc. of ICASSP*, 2021, pp. 6503–6507.

[15] Danwei Cai, Weiqing Wang, Ming Li, Rui Xia, and Chuanzeng Huang, "Pretraining conformer with asr for speaker verification," in *Proc. of ICASSP*, 2023, pp. 1–5.

[16] Danwei Cai and Ming Li, "Leveraging asr pretrained conformers for speaker verification through transfer learning and knowledge distillation," *arXiv preprint arXiv:2309.03019*, 2023.

[17] Ming Cheng, Weiqing Wang, Yucong Zhang, Xiaoyi Qin, and Ming Li, "Target-speaker voice activity detection via sequence-to-sequence prediction," in *Proc. of ICASSP*, 2023, pp. 1–5.

[18] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. of Interspeech*, 2020, pp. 3830–3834.

[19] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. of Interspeech*, 2018, pp. 2252–2256.

[20] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *Proc. of Interspeech*, 2020, pp. 274–278.

[21] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al., "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

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

[23] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. of ICASSP*, 2017, pp. 5220–5224.

[24] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc. of Interspeech*, 2017.

[25] Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman, "The third dihard diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[26] Weiqing Wang and Ming Li, "Incorporating end-to-end framework into target-speaker voice activity detection," in *Proc. of ICASSP*, 2022, pp. 8362–8366.

[27] Federico Landini, Shuai Wang, Mireia Diez, Lukáš Burget, Pavel Matějka, Kateřina Žmolíková, Ladislav Mošner, Oldřich Plchot, Ondřej Novotnỳ, Hossein Zeinali, et al., "But system description for dihard speech diarization challenge 2019," *arXiv preprint arXiv:1910.08847*, 2019.

[28] Shota Horiguchi, Nelson Yalta, Paola Garcia, Yuki Takashima, Yawen Xue, Desh Raj, Zili Huang, Yusuke Fujita, Shinji Watanabe, and Sanjeev Khudanpur, "The hitachi-jhu dihard iii system: Competitive end-to-end neural diarization and x-vector clustering systems combined by dover-lap," *arXiv preprint arXiv:2102.01363*, 2021.

[29] Yuxuan Wang, Maokui He, Shutong Niu, Lei Sun, Tian Gao, Xin Fang, Jia Pan, Jun Du, and Chin-Hui Lee, "Ustc-nelslip system description for dihard-iii challenge," *arXiv preprint arXiv:2103.10661*, 2021.

[30] Mao-Kui He, Jun Du, Qing-Feng Liu, and Chin-Hui Lee, "Ansd-ma-mse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.