

# TARGET-SPEAKER VOICE ACTIVITY DETECTION VIA SEQUENCE-TO-SEQUENCE PREDICTION

Ming Cheng<sup>1,2</sup>, Weiqing Wang<sup>2</sup>, Yucong Zhang<sup>2</sup>, Xiaoyi Qin<sup>1,2</sup>, Ming Li<sup>1,2†</sup>

<sup>1</sup>School of Computer Science, Wuhan University, Wuhan, China

<sup>2</sup>Data Science Research Center, Duke Kunshan University, Kunshan, China

## ABSTRACT

Target-speaker voice activity detection is currently a promising approach for speaker diarization in complex acoustic environments. This paper presents a novel Sequence-to-Sequence Target-Speaker Voice Activity Detection (Seq2Seq-TSVAD) method that can efficiently address the joint modeling of large-scale speakers and predict high-resolution voice activities. Experimental results show that larger speaker capacity and higher output resolution can significantly reduce the diarization error rate (DER), which achieves the new state-of-the-art performance of 4.55% on the VoxConverse test set and 10.77% on Track 1 of the DIHARD-III evaluation set under the widely-used evaluation metrics.

**Index Terms**— Speaker Diarization, Target-Speaker Voice Activity Detection, Sequence-to-Sequence Transformers

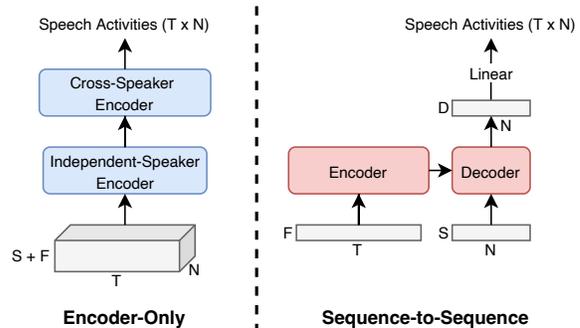
## 1. INTRODUCTION

Speaker diarization refers to the task of determining each speaker’s utterance boundaries in multi-party conversational audio [1]. As a front-end processing to segment audio into pieces with different identities, speaker diarization can effectively participate in various downstream tasks [2].

Conventional speaker diarization consists of several independent modules. First, Voice Activity Detection (VAD) [3] removes non-speech regions from the input audio. Then, the remaining is partitioned into multiple short segments, followed by speaker embedding extraction [4]. By measuring segment-wise speaker similarities with cosine distance or PLDA [5], clustering-based methods can group segmented utterances into different identities [6].

One limitation of conventional approaches is that they cannot handle overlapped speech, as each audio segment is assumed to contain only one speaker. To tackle this problem, target-speaker voice activity detection (TS-VAD) [7] has attracted much interest due to its great success in challenging tasks such as VoxSRC [8, 9], and DIHARD-III [10]. Based on speaker profiles from a clustering-based diarization, the TS-VAD system can estimate each speaker’s frame-level voice activities to refine the initial clustering-based results.

In recent years, many researchers improve the TS-VAD system by modifications on front-end preprocess [11], independent-speaker prediction [12], and cross-speaker [13] modeling. We abstract the usual TS-VAD method into the left part of Fig. 1, namely the encoder-only method. The front-end extractor is omitted in the plot, which extracts frame-level representations  $\mathbf{E}_{\text{rep}} \in \mathbb{R}^{T \times F}$  from the input audio. Let  $\mathbf{E}_{\text{spk}} \in \mathbb{R}^{N \times S}$  denote the given speaker profiles. Each target-speaker embedding in  $\mathbf{E}_{\text{spk}}$  has to be repeated  $T$  times and concatenated with the  $\mathbf{E}_{\text{rep}}$  to produce a 3-dimensional



**Fig. 1.** Diagram of encoder-only TS-VAD method (left) and our proposed sequence-to-sequence method (right).  $T$  and  $F$  denote the length and dimension of extracted frame-level features.  $N$  and  $S$  denote the number and dimension of speaker embeddings.  $D$  represents the output dimension of the decoder.

input tensor with the shape of  $T \times N \times (S + F)$ . Then, various encoders (e.g., Bi-LSTM, Self-Attention) predict speech activities by processing the input data along the time-axis ( $T$ ) and speaker-axis ( $N$ ). However, the encoder-only method leads to high demand for GPU memory since the space consumption of the input tensor is proportional to  $T \times N \times (S + F)$ . With the growth of  $T$  and  $N$ , the dramatic boost of memory usage limits the model to process longer feature sequences and more speakers at once. Furthermore, the output length of encoder-only models must be equal to  $T$ , which means the temporal resolution of voice activity detection (VAD) cannot be changed arbitrarily.

Considering the limitations above, we propose a sequence-to-sequence framework for target-speaker voice activity detection, shown in the right part of Fig. 1. Our main contributions include:

1. Frame-level representations and speaker embeddings are separately fed into the encoder and decoder sides. Thus, the memory consumption of the inputs becomes proportional to  $T \times F + N \times S$ , which reduces redundant tensors.
2. The decoder compresses each speaker’s voice activity information into a fixed-dim ( $D$ ) embedding regardless of the length of input features. The last linear layer can adjust the model to predict voice activities in higher temporal resolution with a small computational cost.
3. Experimental results show that joint modeling of more speakers with higher VAD resolution can reduce the diarization error rate (DER) significantly, which achieves the new state-of-the-art DERs of 4.55% on the VoxConverse [14] test set and 10.77% on Track 1 of the DIHARD-III [15] evaluation set.

† Corresponding Author, E-mail: ming.li369@dukekunshan.edu.cn

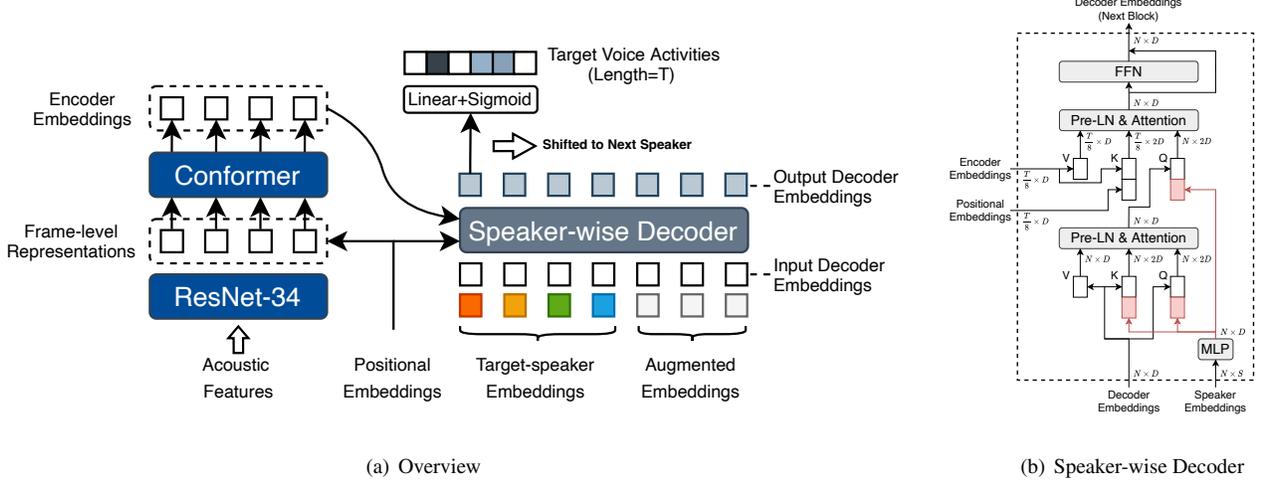


Fig. 2. The Sequence-to-Sequence Target-Speaker Voice Activity Detection (Seq2Seq-TSVAD) Framework.

## 2. METHODS

The framework of our proposed Seq2Seq-TSVAD method is shown in Fig. 2(a). The details are introduced as follows.

### 2.1. ResNet-34

The front-end feature extractor is ResNet-34 [16], where the widths (number of channels) of the residual blocks are  $\{64, 128, 256, 512\}$ . Given  $\mathbf{X} \in \mathbb{R}^{T \times H}$  denoting the acoustic feature sequence with the length of  $T$  and dimension of  $H$ , the model outputs a CNN feature map  $\mathbf{M} \in \mathbb{R}^{C \times \frac{T}{8} \times \frac{H}{8}}$ , where  $C$  is the number of channels. Then, we adopt the segmental statistical pooling [12] (SSP) method to aggregate the channel-wise feature map into frame-level representations  $\mathbf{E}_{\text{rep}} \in \mathbb{R}^{\frac{T}{8} \times F}$ , where  $F$  denotes the output dimension of the SSP layer. This process can be viewed as the downsampling from the input acoustic feature sequence.

### 2.2. Conformer

We introduce the Conformer [17] to further model long-term dependencies between frame-level representations. The input layer of the Conformer module firstly maps  $\mathbf{E}_{\text{rep}}$  to  $\mathbf{E}_{\text{rep}}' \in \mathbb{R}^{\frac{T}{8} \times D}$  with sinusoidal positional encodings, where  $D$  is equal to the attention size. Then, the input sequence  $\mathbf{E}_{\text{rep}}'$  is encoded into  $\mathbf{E}_{\text{enc}} \in \mathbb{R}^{\frac{T}{8} \times D}$  by the following Conformer blocks.

### 2.3. Speaker-wise Decoder

The Speaker-wise Decoder (SW-D) estimates target-speaker voice activities by considering cross-speaker correlations. The input on the decoder side can be broken into two parts: decoder embeddings and auxiliary queries. In this case, the input decoder embeddings are initialized by zeros, and target-speaker embeddings are fed as auxiliary queries.

Let  $\mathbf{E}_{\text{spk}} = [e_1 \dots e_N]^T \in \mathbb{R}^{N \times S}$  denote the given target-speaker embeddings, where  $N$  is the number of speakers and  $S$  is the embedding dimension. The SW-D module incorporates  $\mathbf{E}_{\text{enc}}$  and  $\mathbf{E}_{\text{spk}}$  to output processed decoder embeddings which contain information about each speaker's voice activities. Let

$\mathbf{E}_{\text{dec}} = [v_1 \dots v_N]^T \in \mathbb{R}^{N \times D}$  denote the predicted decoder embeddings, where  $D$  is equal to the attention size used in Conformer blocks. After a linear projection with sigmoid activation, each  $v_i$  can be transformed into posterior probabilities of voice activities belonging to the  $i$ -th speaker. The output dimension of the last linear layer controls the temporal resolution of detected voice activities, which means a longer output length can provide more precise timestamps.

Fig. 2(b) demonstrates the structure of the SW-D module. Inspired by DAB-DETR [18], we modify the vanilla Transformer decoder so that speaker embeddings can be introduced into the attention mechanism. In each decoder block, target-speaker embeddings go through a multi-layer perceptron (MLP) to generate the within-block representations. Before the first attention layer, the within-block embeddings are concatenated with the original keys and queries to calculate attention scores. Since decoder embeddings are designed to learn positional (temporal) information about voice activities, we also concatenate keys from encoders with corresponding positional embeddings. This way, all key-query calculations can incorporate positional and speaker-related information. The decoding process effectively exchanges information between all speakers by self-attention mechanism.

The adopted MLP module consists of two linear layers with in-between Layer Normalization and ReLU activation. The Pre-LayerNorm method [19] is employed before each attention layer. For clarity, we omit to plot residual connections of attention layers.

### 2.4. Embedding Augmentation

To align the different speaker numbers in training mini-batch data, a set of augmented embeddings for each audio sample is supplied to ensure that the SW-D module has a fixed decoding length. Each augmented embedding may be directly set to zeros with a probability of 0.5. Otherwise, it will be padded by another one not appearing in the current audio. Then, there is a probability of 0.2 to replace all input speaker embeddings with non-existent speakers. Furthermore, the input embeddings and corresponding ground-truth labels should be shuffled to keep the model invariant to speaker order. This augmentation method is vital because it forces the network to learn to distinguish valid or invalid speakers in given audio.

### 3. EXPERIMENTAL RESULTS

#### 3.1. Speaker Embedding

We adopt the ResNet-34 model as the pattern extractor. After the statistical pooling [4] layer, a linear projection outputs the 256-dim speaker embedding. The ArcFace ( $s=32, m=0.2$ ) [20] is used as the classifier. The detailed configuration of the neural network is the same as [21]. The model trained on the VoxCeleb2 [22] dataset finally achieves an equal error rate (EER) of 0.814% on the Vox-O [23] trial. The pre-trained model is used for speaker embedding extraction in the subsequent Seq2Seq-TSVAD system.

#### 3.2. Seq2Seq-TSVAD Configuration

##### 3.2.1. Model Parameters

The front-end ResNet-34 is initialized by the pre-trained speaker embedding model. We replace the original statistical pooling (SP) with a segmental statistical pooling (SSP) [12] layer to extract frame-level features. The back-end model consists of Conformer encoders [17] and Transformer decoders [24]. All encoder-decoder modules have 6 blocks sharing the same settings: 512-dim attentions with 8 heads and 1024-dim feed-forward layers with a dropout rate of 0.1. The kernel size of convolutions in Conformer blocks is 15, and the other implementation details are the same as [17].

##### 3.2.2. Data Simulation

As neural network-based methods can usually benefit from large-scale training data, we create a simulated corpus by the VoxCeleb2 [22] and LibriSpeech [25] datasets. The WebRTC<sup>1</sup> toolkit is used to remove non-speech regions from raw audio. Then, each simulated signal is generated by mixing source utterances of 1-4 speakers from the processed corpus. The overlap ratio of simulated data has not been controlled because each single-speaker component is produced randomly and individually in an online process.

##### 3.2.3. Training Process

All training audio signals are split into 16-sec chunks. The model input takes 80-dim log Mel-filterbank energies with a frame length of 25 ms and a frameshift of 10 ms as acoustic features. Furthermore, background noise from Musan [26] and reverberation from RIRs [27] are applied as the data augmentation.

We implement the BCE loss and Adam optimizer to train the neural network with a linear learning rate warm-up. Firstly, the model with frozen ResNet-34 is trained by simulated data until back-end convergence. Then, all model parameters are unfrozen to keep them trainable. The real data from the specific dataset is added to the simulated data at a ratio of 0.2. Finally, we finetune the model on real data without any simulation. The first two stages take around 200 epochs with a learning rate of  $1e-4$ , and the last stage decreases the learning rate to  $1e-5$ .

##### 3.2.4. Inference Settings

A clustering-based diarization first provides an initial result in the inference stage. Based on that, the pre-trained speaker embedding model extracts target embeddings from speech segments. The speakers with non-overlapped speech shorter than 2 seconds are discarded.

<sup>1</sup><https://github.com/wiseman/py-webrtcvad>

**Table 1.** DERs (%) of different Seq2Seq-TSVAD models on the VoxConverse test set (collar = 250 ms). \* denotes the model training without embedding augmentation.

VAD Resolution	Decoding Length	DER (%)		
		1-10 SPKs	10+ SPKs	Total
R = 80 ms	L = 10	3.94	9.16	5.01
	L = 20	4.03	7.08	4.66
	L = 30	3.96	7.03	<b>4.59</b>
	L = 30 *	9.07	21.66	11.65
R = 10 ms	L = 10	4.03	10.29	5.32
	L = 20	3.99	7.10	4.63
	L = 30	3.94	6.93	<b>4.55</b>
	L = 30 *	8.45	16.52	10.10

**Table 2.** Comparisons of our proposed Seq2Seq-TSVAD models with others on the VoxConverse test set (collar = 250 ms).

Method	DER (%)	JER (%)
ByteDance [28] †	5.17	29.56
DKU-DukeECE [9] ‡	4.94	28.79
Wang et al. [13] *	4.57	-
AHC [9]	5.35	27.99
+ Seq2Seq-TSVAD (R=80ms, L=30)	4.59	29.39
+ Seq2Seq-TSVAD (R=10ms, L=30)	<b>4.55</b>	29.13

† VoxSRC-21 2nd-ranked result with 3-system fusion. The top-ranked team did not report the DER/JER for this set.

‡ VoxSRC-22 1st-ranked result with 4-system fusion.

\* The authors update this result from 4.74% to 4.57% in their latest version.

We split each test recording into 16-sec chunks and feed them into the TS-VAD model with extracted speaker embeddings. Then, the final predictions are stitched chunk by chunk. The maximum number of speaker profiles (decoding length) is set to  $L$ . If the detected speakers are less than  $L$ , the rest of the speaker embeddings will be padded by zeros. Otherwise, the extra part will be inferred in the next group.

Lastly, we adopt the general voice activity detection (VAD) to revise the TS-VAD predictions as post-processing. According to specific evaluation requirements, Oracle or Estimated VAD information can be obtained from the initial diarization result. The timestamps marked as active speech will directly assign a positive label to the speaker with the highest predicted score. Predictions at the timestamps marked as non-speech will be zeroed.

### 3.3. Evaluation

#### 3.3.1. VoxConverse Dataset

We directly copy our AHC-based diarization model used in VoxSRC-22 [9] to be the initial system for speaker embedding extraction, which obtains a DER of 5.35% on the VoxConverse [14] test set with a collar of 250 ms.

According to the instructions described in Section 3.2, we train and evaluate our proposed Seq2Seq-TSVAD models using different VAD resolutions (duration per frame-level prediction) and decoding lengths (maximum capacity for speaker embeddings). Two types of VAD resolutions are provided as coarse (80 ms) and precise (10 ms) options, which can be implemented easily by adjusting the dimension of the output layer in our model. As the VoxConverse test set

**Table 3.** DERs (%) of different Seq2Seq-TSVAD models on Track 1 of the DIHARD-III evaluation set (Oracle VAD). \* denotes the model training without embedding augmentation.

VAD Resolution	Decoding Length	DER (%)		
		1-5 SPKs	5+ SPKs	Total
R = 80 ms	L = 5	10.46	24.57	12.44
	L = 10	10.56	23.00	12.30
	L = 20	10.42	22.69	<b>12.14</b>
	L = 20 *	10.73	25.40	12.79
R = 10 ms	L = 5	8.91	23.96	11.02
	L = 10	8.97	22.02	10.80
	L = 20	8.94	21.99	<b>10.77</b>
	L = 20 *	9.29	24.25	11.39

**Table 4.** Comparisons of our proposed Seq2Seq-TSVAD models with others on Track 1 of the DIHARD-III evaluation set (Oracle VAD).

Method	DER (%)	JER (%)
USTC-NELSLIP [10] <sup>†</sup>	11.30	29.94
Hitachi-JHU [29] <sup>‡</sup>	11.58	32.37
Wang et al. [11]	11.30	-
LSTM-SC [12]	15.40	33.27
+ Seq2Seq-TSVAD (R=80ms, L=20)	12.14	29.49
+ Seq2Seq-TSVAD (R=10ms, L=20)	<b>10.77</b>	28.46

<sup>†</sup> DIHARD3-III 1st-ranked result with 5-system fusion.

<sup>‡</sup> DIHARD3-III 2nd-ranked result with 5-system fusion.

has a maximum of 21 speakers in a single recording, we choose the decoding lengths of 10, 20, and 30 to cover situations of insufficient, suitable, and sufficient speaker capacities, respectively.

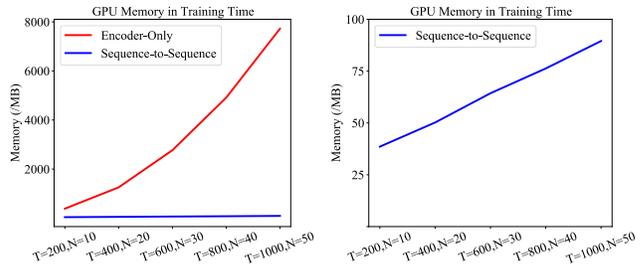
Table 1 shows that the total DER reduces obviously when the decoding length increases from 10 to 20. And the performance gain mainly comes from the subset of recordings with over 10 speakers. In addition, a longer decoding length of 30 still brings a slight improvement. It can be seen that the precise VAD resolution (10 ms) achieves the lowest DER of 4.55%. Table 2 compares our proposed method with the current state-of-the-art results. Our best performance significantly outperforms previous ones, especially for some multi-system fusion results in past VoxSRC challenges.

### 3.3.2. DIHARD-III Dataset

To evaluate our proposed method on highly overlapped speech with complex background noises, we also test it on Track 1 of DIHARD-III Diarization Challenge [15]. The LSTM-SC method in [12] is adopted as the initial system for speaker embedding extraction, which has a DER of 15.4% on the evaluation set with Oracle VAD.

Since there are up to 9 speakers can appear in the DIHARD-III recordings, we explore decoding lengths of 5, 10, and 20. Table 3 illustrates that the longer decoding length and more precise VAD resolution can bring a better result, which has the same conclusion as Table 1. Since the DIHARD-III dataset is annotated at 10-ms levels and evaluated without a tolerance collar, it is sensitive to utterance boundaries. Thus, the benefits of precise VAD resolution (10 ms) on it seem more significant than VoxConverse Dataset.

Table 4 shows the comparisons of our proposed method with others. Our best single system obtains a DER of 10.77% on Track 1 of the DIHARD-III evaluation set, demonstrating a competitive



**Fig. 3.** Visualization of overall GPU memory usage in designed training configurations.

performance over previous state-of-the-art results.

### 3.4. GPU Memory Visualization

There are usually many variable components in a TS-VAD system (e.g., front-end extractors, encoders, decoders), and it is not easy to precisely compare the memory usage of our proposed method with others. To simplify the ablation test, we only count the encoder-decoder modules shown in Fig. 1. The encoders in the two frameworks are Conformer blocks, and the decoder in the Seq2Seq-TSVAD is our modified speaker-wise decoder. All encoder-decoder modules are set to contain a single block.

By fixing the  $S$ ,  $F$ , and  $D$  equal to 256 (dimension of speaker embedding), Fig. 3 visualizes the memory usage of two frameworks in different combinations of  $T$  and  $N$ . With the increase of  $T$  and  $N$ , the memory usage of the encoder-only TS-VAD method shows exponential growth, whereas the Seq2Seq-TSVAD method has a nearly linear trend. It should be noted that the numerical results are not absolutely precise because there are also a series of extra memory costs during model training (e.g., GPU context, optimizer, gradient checkpoints). In practice, the Seq2Seq-TSVAD does not need to adopt longer  $T$  for higher VAD resolution but adjusts the last linear layer, which can save more GPU memory.

## 4. CONCLUSIONS

This paper presents a novel Sequence-to-Sequence Target-Speaker Voice Activity Detection (Seq2Seq-TSVAD) method to be scalable in terms of speaker capacity and temporal resolution. By factorizing the time-axis ( $T$ ) and speaker-axis ( $N$ ) of the original TS-VAD input, our proposed method reduces its space complexity from  $\mathcal{O}(T \times N)$  into  $\mathcal{O}(T + N)$ , where  $S$  and  $F$  are ignored as constant values. Consequently, the novel framework can efficiently deal with 30 speakers to predict 10ms-level voice activities in a single 12-GB GPU device, which still does not reach the upper limit. Meanwhile, it obtains new state-of-the-art DERs of 4.55% on the VoxConverse [14] test set and 10.77% on Track 1 of the DIHARD-III [15] evaluation set, respectively.

## 5. ACKNOWLEDGMENTS

This research is funded in part by the National Natural Science Foundation of China (62171207), Science and Technology Program of Guangzhou City (202007030011), Science and Technology Program of Suzhou City (SYC2022051). Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

## 6. REFERENCES

- [1] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.
- [2] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, pp. 101317, 2022.
- [3] Shuo-Yiin Chang, Bo Li, Gabor Simko, Tara N. Sainath, Anshuman Tripathi, Aron van den Oord, and Oriol Vinyals, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proc. ICASSP*, 2018, pp. 5549–5553.
- [4] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.
- [5] Simon J.D. Prince and James H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.
- [6] Qingjian Lin, Ruiqing Yin, Ming Li, Herv Bredin, and Claude Barras, "Lstm based similarity measurement with spectral clustering for speaker diarization," in *Proc. Interspeech*, 2019, pp. 366–370.
- [7] 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. Interspeech*, 2020, pp. 274–278.
- [8] Weiqing Wang, Danwei Cai, Qingjian Lin, Lin Yang, Junjie Wang, Jin Wang, and Ming Li, "The dku-dukeeece-lenovo system for the diarization task of the 2021 voxceleb speaker recognition challenge," *arXiv preprint arXiv:2109.02002*, 2021.
- [9] Weiqing Wang, Xiaoyi Qin, Ming Cheng, Yucong Zhang, Kangyue Wang, and Ming Li, "The dku-dukeeece diarization system for the voxceleb speaker recognition challenge 2022," *arXiv preprint arXiv:2210.01677*, 2022.
- [10] 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.
- [11] Yu-Xuan Wang, Jun Du, Maokui He, Shu-Tong Niu, Lei Sun, and Chin-Hui Lee, "Scenario-dependent speaker diarization for dihard-iii challenge," in *Proc. Interspeech*, 2021, pp. 3106–3110.
- [12] Weiqing Wang, Qingjian Lin, Danwei Cai, and Ming Li, "Similarity measurement of segment-level speaker embeddings in speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2645–2658, 2022.
- [13] Dongmei Wang, Xiong Xiao, Naoyuki Kanda, Takuya Yoshioka, and Jian Wu, "Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization," *arXiv preprint arXiv:2208.13085*, 2022.
- [14] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman, "Spot the conversation: Speaker diarisation in the wild," in *Proc. Interspeech*, 2020, pp. 299–303.
- [15] Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman, "The third dihard diarization challenge," in *Proc. Interspeech*, 2021, pp. 3570–3574.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
- [17] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.
- [18] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," in *Proc. ICLR*, 2022.
- [19] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Teyan Liu, "On layer normalization in the transformer architecture," in *Proc. ICML*, 2020, pp. 10524–10533.
- [20] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4685–4694.
- [21] Weiqing Wang, Danwei Cai, Xiaoyi Qin, and Ming Li, "The dku-dukeeece systems for voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2010.12731*, 2020.
- [22] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
- [23] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.
- [25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [26] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [27] 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. ICASSP*, 2017, pp. 5220–5224.
- [28] Keke Wang, Xudong Mao, Hao Wu, Chen Ding, Chuxiang Shang, Rui Xia, and Yuxuan Wang, "The bytedance speaker diarization system for the voxceleb speaker recognition challenge 2021," *arXiv preprint arXiv:2109.02047*, 2021.
- [29] 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.