EFFICIENT PERSONAL VOICE ACTIVITY DETECTION WITH WAKE WORD REFERENCE SPEECH

Bang Zeng[1,2], *Ming Cheng*[1,2], *Yao Tian*[3], *Haifeng Liu*[4], *Ming Li*[1,2†]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Duke Kunshan University, Kunshan, China
[3]Data & AI Engineering System, OPPO, Beijing, China
[4]University of Science and Technology of China, Hefei, China

ABSTRACT

Personal voice activity detection (PVAD) is gradually used in speech assistants. Traditional PVAD schemes extract the target speaker's embedding from existing query reference speech through a pre-trained speaker verification model. Consequently, the performance of the PVAD model may suffer if the quality of the extracted speaker embedding is poor, such as when only utilizing wake word speech as the reference. In this work, we introduce a novel and efficient PVAD model. In contrast to conventional approaches that rely on speaker embeddings extracted from a pre-trained speaker verification model, our proposed method directly uses the raw frame-level features of the reference speech as the target speaker's attributes. In this way, our proposed model achieves an ultra-high recall rate, which is vital for speech assistant applications. The experimental results show the effectiveness of our proposed method in both cases of using existing query speech or wake word speech as reference.

***Index Terms*—** Personal voice activity detection, Speaker verification, Wake word speech, Recall rate.

1. INTRODUCTION

Voice activity detection (VAD) [1, 2] is a task identifying speech or non-speech at frame level. In actual noisy environments, we pay more attention to whether a target speaker is speaking. Personal voice activity detection (PVAD) [3, 4, 5] is a technique to determine the target speaker's speech segments in multi-speaker scenarios. Typical PVAD methods first extract the target speaker's embedding from existing reference speech using a pre-trained speaker verification (SV) model. Subsequently, with the assistance of the speaker embedding, the PVAD models identify the segments containing the target speaker's speech. PVAD can be a robust front-end in various speech-related tasks, such as automatic speech recognition [6], speaker verification [7] and speaker diarization [8, 9].

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

However, several challenges arise when applying PVAD to speech assistants. Firstly, the reference speech used for target speaker registration, often derived from existing query speech, needs periodic updates due to the temporal variability inherent in SV [10, 11]. Mismatches between the recording environment of the reference speech and the current environment can adversely affect the PVAD system's performance. Secondly, the performance of the pre-trained SV model does not necessarily correlate positively with the PVAD system's performance, since they are optimized with different loss functions. Selecting an appropriate SV model for the PVAD task requires careful consideration. Last but not the least, the PVAD model must balance between a small number of parameters and a low rate of error, especially the miss of target speech. Low recall rate of target speaker's speech frames may fail downstream speech assistant applications.

Existing PVAD methods have not effectively tackled the issues mentioned above. [4] introduces a model training approach to mitigate the challenge of insufficient reference speech data. Nevertheless, employing online wake word speech as the reference can serve as a registration free solution, as online instant wake word speech is readily accessible through wake word detection [12, 13] whenever the speech assistant becomes active. [5] use a feature-wise transform (FiLM) [14] layer as an alternative to concatenation. However, using an utterance-level speaker embedding for fusion with frame-level acoustic features may not be optimal.

This work explores PVAD models in both cases of using existing query speech or wake word speech as reference. Inspired by [8], we propose a PVAD model, denoted as PVAD-FISV, where the front-end feature extractor is initialized with a pre-trained SV model. This approach yields a favorable correlation between the effectiveness of SV and PVAD models. However, it is crucial to highlight that initializing the front-end with the SV model significantly increases the number of parameters in the PVAD model. To address this concern, we introduce a efficient PVAD model called PVAD-FTSA, which stands for PVAD model with Frame-level Target Speaker's

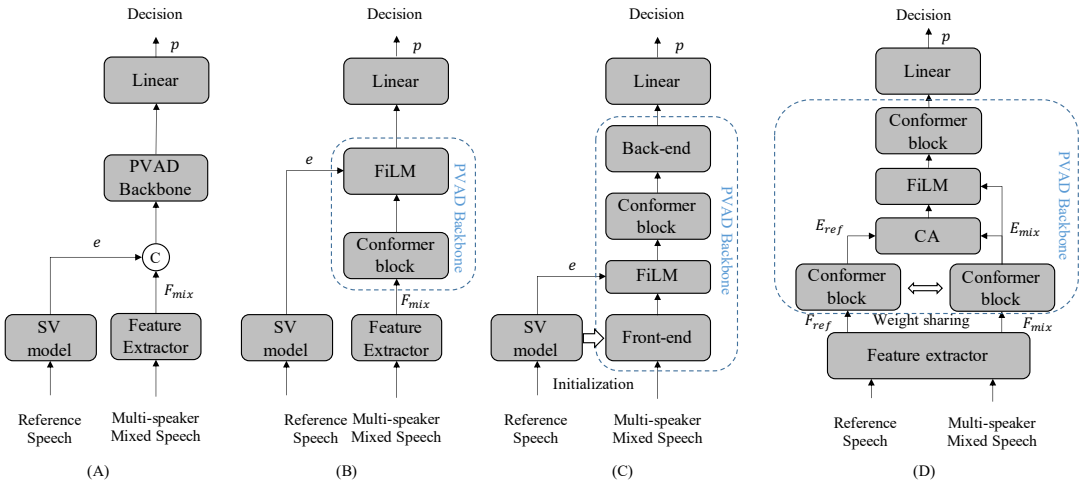

Fig. 1: (A) Typical PVAD model with concatenation as the speaker embedding fusion module. (B) PVAD model with FiLM layer as the speaker embedding fusion module. (C) The Proposed PVAD-FISV model. (D) The Proposed PVAD-FTSA model, 'CA' denotes the Cross-Attention module.

Attributes. Our proposed method applies a cross-attention mechanism to multi-speaker and reference speech acoustic features. Subsequently, we use the frame-level output as the target speaker's attributes and fuse it with the acoustic features of multi-speaker input speech through a FiLM layer. Notably, the PVAD-FTSA model operates without the need of an SV model during the training and inference stages. Moreover, in contrast to fixed utterance-level speaker embedding, this frame-level target speaker's attributes contribute to increasing the recall rate in PVAD models.

2. METHODS

In this section, we first briefly review two typical PVAD models presented in PVAD 1.0 [3] and PVAD 2.0 [5], respectively. Following that, we introduce our two proposed PVAD models, namely PVAD-FISV and PVAD-FT.

2.1. Typical PVAD models

The diagram of the typical PVAD models is shown in Figure 1 (A). The PVAD model employs a pre-trained SV model to extract the target speaker's embedding from the reference speech. The speaker embedding repeatedly concatenates with the acoustic feature of the multi-speaker mixed input speech and then input to the PVAD backbone. Finally, the output of the PVAD model is input to the linear layer and derives the decision of the target speaker's speech at the frame level:

$$\boldsymbol{p} = PVAD([\boldsymbol{F}_{mix}, \boldsymbol{e}]) \tag{1}$$

where $\boldsymbol{e} \in \mathbb{R}^{1\times D}$ and $\boldsymbol{F}_{mix} \in \mathbb{R}^{T\times N}$ denote the speaker embedding and acoustic feature, respectively. T denotes the sequence length. D and N are denote the feature dimension. $\boldsymbol{p}$ denotes the posterior of the target speaker's speech.

Fusing the speaker embedding by concatenation may not be optimal. [5] proposes to use a FiLM layer as the fusion module. The diagram of the optimized approach is shown in Figure 1 (B). The whole process can be formulated as:

$$\boldsymbol{p} = L(FiLM(Con(\boldsymbol{F}_{mix}), \boldsymbol{e})) \tag{2}$$

$$FiLM(Con(\boldsymbol{F}_{mix}), \boldsymbol{e}) = \gamma(\boldsymbol{e}) \cdot Con(\boldsymbol{F}_{mix}) + \beta(\boldsymbol{e}) \tag{3}$$

where $L(\cdot)$ and $Con(\cdot)$ denote the linear and Conformer [15] operation, respectively. $\gamma(\cdot)$ and $\beta(\cdot)$ are the scaling and shifting vectors of FiLM respectively.

2.2. Proposed PVAD models

2.2.1. PVAD with A Front-end Initialized by SV Model.

Inspired by [8], we propose the PVAD-FISV method. The diagram of the PVAD-FISV model is shown in Figure 1 (C). The front-end feature extractor and the pre-trained SV model share a similar structure, differing primarily in the pooling layer configuration. The front-end extractor employs segmental statistical pooling to generate a frame-level representation of the mixed speech. In contrast, SV typically produces a segment-level feature using statistical pooling. The back-end module consists of several transposed convolution layers, each with the same stride and kernel size as the corresponding convolutional layers in the front-end extractor. Notably, during the training phase, the front-end is initialized with the pre-trained SV model.

2.2.2. PVAD with Frame-level Target Speaker Attributes.

While the architecture of the PVAD-FISV model is efficient, it significantly increases the number of model parameters. To address this concern, we introduce the PVAD-FTSA model, which is depicted in 1 (D). Instead of relying on a pre-trained

Table 1: Results on simulated test set from FFSVC22 using **query** speech as reference (%). 2 speaker: mixed wave from two speaker's speech. 'Noise' denotes adding noise from Musan. REC: Recall. PRE: Precision. F1: F1 score. AUC: Accuracy.

Exp	Model	2 Speaker				2 Speaker + Noise				Parameters(M)
		REC	PRE	F1	AUC	REC	PRE	F1	AUC	PVAD/SV
E1	Seq2Seq-TSVAD	86.2	62.5	72.4	64.2	84.0	62.6	71.7	63.8	86.6/21.9
E2	LSTM+Concat	31.3	65.4	42.4	53.5	29.4	65.0	40.5	52.8	2.4/21.9
E3	Conformer+Concat	31.5	81.0	45.4	58.6	28.3	79.4	41.8	56.9	2.9/21.9
E4	Conformer+FiLM	72.6	70.5	71.6	68.5	68.8	70.3	69.5	67.1	2.9/21.9
E5	PVAD-FISV	75.5	85.8	80.3	79.8	74.8	82.4	78.4	77.6	27.8/21.9
E6	PVAD-FTSA(FFN=512)	84.5	69.0	76.0	70.9	80.6	68.2	73.9	68.9	3.3/0.0
E7	PVAD-FTSA(FFN=1024)	87.5	69.4	77.4	72.1	82.5	69.0	75.2	70.2	4.6/0.0

SV model, the PVAD-FTSA model employs two weight-sharing Conformer blocks for processing mixed and reference speech acoustic features, respectively. Subsequently, we feed these two acoustic features into the Cross-Attention (CA) module, utilizing a Transformer Encoder [16] as the CA module in this study. We use the frame-level output of the CA module as the target speaker's attribute and fuse it with $\boldsymbol{E}_{mix}$ through a FiLM layer. Follow This, $\boldsymbol{p}$ can be formulated as:

$$\boldsymbol{p} = L(Con(FiLM(\boldsymbol{E}_{mix}, CA(\boldsymbol{E}_{mix}, \boldsymbol{E}_{ref})))) \quad (4)$$

$$CA(\boldsymbol{E}_{mix}, \boldsymbol{E}_{ref}) = Tra(q = \boldsymbol{E}_{mix}; k, v = \boldsymbol{E}_{ref}) \quad (5)$$

where $CA(\cdot)$ denotes the cross-attention module. $Con(\cdot)$ and $Tra(\cdot)$ denote the Conformer and Transformer operation, respectively. $\boldsymbol{E}_{mix}$ and $\boldsymbol{E}_{ref}$ are the Conformer encoder of $\boldsymbol{F}_{mix}$ and $\boldsymbol{F}_{ref}$, respectively.

3. EXPERIMENTAL SETUP

3.1. Datasets

Training Sets: We use the VoxCeleb2 [17] dataset to create a simulated multi-speaker training set. Considering that there are generally two speakers in a short speech, each simulated audio contains 1-2 speakers. Each speaker component of the simulated multi-speaker audio is randomly created online. We use the Musan [18] and RIRs [19] datastes for data augmentation. Moreover, we add the real data from the AliMeeting [20] and Aishell-4 [21] datasets into the simulated training data at a ratio of 0.2. **Evaluation Sets**: We evaluate our proposed models separately on two datasets, one is simulated from the publicly open FFSVC22 [22] dataset[1] while the other one is a vendor collected in-house dataset targeting real applications. We create 3000 2-spaker-mix utterances using the close-talking iPhone data from FFSVC22 train set of the track 1. The speech component of two speakers are set in a relative signal-noise ratio (SNR) between 0 to 5 dB and the overlap ratio has not been controlled. We randomly select query and wake word speech of the corresponding speaker as two different types of reference speech. The vendor-collected set contains 260 2-speaker mixed utterances from 13 speakers, and 3 wake word utterances for each speaker. The overlap ratio of simulated training and test sets has not been controlled in this work. Considering the usage scenario of a smartphone speech assistant, a speaker with higher SNR is selected as the target speaker.

[1]The simulated datasets from FFSVC22 are available at: `https://github.com/ZBang/pvad`

3.2. Implementation Details

We adopt the ResNet34-SE [23] architecture with statistical pooling layer as the SV model for our PVAD systems. The speaker embedding size is 256. The SV model has about 21.9 million parameters. All compared approaches use the same pre-trained SV model. There are 1 and 2 Conformer layers for the Conformer block of PVAD-FTSA and PVAD-FISV, respectively. Each Conformer layer has an attention dimension of 256, an attention head of 4 and a feed-forward dimension of 1024. The configurations for the CA module in the PVAD-FTSA model are the same as its Conformer block. We use the same model input as a 80-dim log Mel-filterbank energies with a frame length of 25 ms and a frame-shift of 10 ms. We trained our models with binary cross entropy loss and Adam optimizer. The initial learning rate is set to 1e-4.

4. RESULTS AND DISCUSSIONS

To evaluate the effectiveness of our proposed PVAD models, we conduct a series of comparisons among six PVAD models using both simulated and vendor-collected test datasets. **Seq2Seq-TSVAD** denotes one of the state-of-the-art speaker diarization (SD) models presented in [8]. **LSTM+Concat** denotes the model shown in Figure 1 (A), which is proposed in [3]. **Conformer+FiLM** denotes the model shown in the Figure 1 (B), which is proposed in [5]. The configurations of these two models are the same as that of [5]. **PVAD-FISV** and **PVAD-FTSA** are two proposed models, shown in Figure 1 (C) and Figure 1 (D), respectively.

4.1. Results on Simulated Test Set

The results on the simulated test set from FFSVC22 using query reference speech and wake word speech are shown in Table 1 and Table 2, respectively. We use the Recall (REC),

Table 2: Results on simulated test set from FFSVC22 using **wake word** speech as reference (%). 2 speaker: mixed wave from two speaker's speech. 'Noise' denotes adding noise from Musan. REC: Recall. PRE: Precision. F1: F1 score. AUC: Accuracy.

Exp	Model	2 Speaker				2 Speaker + Noise				Parameters(M)
		REC	PRE	F1	AUC	REC	PRE	F1	AUC	PVAD/SV
E8	Seq2Seq-TSVAD	85.9	62.3	72.2	63.9	83.7	62.4	71.5	63.6	86.6/21.9
E9	LSTM+Concat	39.9	65.0	49.5	55.5	38.9	64.5	48.5	54.9	2.4/21.9
E10	Conformer+Concat	36.7	81.3	50.6	60.9	33.1	80.1	46.8	59.0	2.9/21.9
E11	Conformer+FiLM	67.7	71.5	69.5	67.6	63.7	71.4	67.3	66.2	2.9/21.9
E12	PVAD-FISV	71.2	86.8	78.2	78.4	70.1	84.0	76.6	76.5	27.8/21.9
E13	PVAD-FTSA(FFN=512)	81.5	68.0	74.1	69.0	75.5	67.2	71.1	66.5	3.3/0.0
E14	PVAD-FTSA(FFN=1024)	83.7	70.2	76.4	71.7	76.0	69.9	72.9	69.1	4.6/0.0

Precision (PRE), F1 score (F1) and Accuracy (AUC) as metrics. The Seq2Seq-TSVAD is a sota SD model with multi-speaker outputs. Consequently, the PRE of the Seq2Seq-TSVAD (E1, E8) tends to be low when the target output is only one target speaker. The results indicate that the PVAD-FISV model (E5, E12) excels in overall PVAD performance. However, the number of parameters (27.8 M) and REC (E5:75.5%, E12:71.2%) constrains the utilization of the PVAD-FISV model for speech assistant applications. Despite the PVAD-FTSA model (E7, E14) performing less effectively in terms of PRE compared to the baseline models (E2-E4, E9-E11), its overall performance (F1 and AUC) surpasses that of the baseline models. This is attributed to the PVAD-FTSA model achieving an ultra-high REC (E7:87.5%, E14:83.7%), a crucial factor for speech assistant applications. Although we adjust the dimension of the feed-forword network (FFN) in PVAD-FTSA to 512 (E6, E13) to closely match the parameter count size of the baseline models (3.3 M vs. 2.9 M), the overall performance of the PVAD-FTSA model, particularly in terms of the REC, remains superior to that of the baseline model. Furthermore, in contrast to the baseline model, the PVAD-FTSA model eliminates the need for an additional SV model during both the training and inference stages.

Comparing Table 1 and 2, the performance of all PVAD systems with FiLM degrades a little bit when using wake word speech as the reference, as opposed to using query speech as the reference. Nevertheless, when it comes to the REC, the decrease in performance of the PVAD-FTSA model (E7:87.5% to E14:83.7%) is less pronounced compared to the baseline models (E4:72.6% to E11:67.7%). Based on the aforementioned results, the PVAD-FTSA model demonstrates effectiveness in both scenarios, whether utilizing existing query speech or wake word speech as reference. It is worth noting that we have observed a notable impact of noise on the REC of the PVAD-FTSA model when using wake word reference speech (E14: 83.7% to 76.0%). We intend to delve deeper into this issue in our future research endeavors.

4.2. Results on Vendor-Collected Test Set

The results on the vendor-collected test set are shown in Table 3. We use the Word Error Rate (WER) and Deletion Error Rate (DEL) as metrics. We use WeNet [24] model as the speech recognition model in this experiments. The PVAD-FISV model performs best in terms of WER (E19:30.4%) but has a relatively high DEL (E19:12.3%). While the WER of the PVAD-FTSA model is slightly higher than that of PVAD-FISV (E21:35.6% vs. E19:30.4%), it is important to note that the PVAD-FTSA model attains an ultra-low DEL (E21:1.0%). Furthermore, the DEL of the PVAD-FTSA model closely approaches that of the original mixture (E21:1.0% vs. E15:0.7%), which is considered as the upper bound. In summary, the conclusions drawn from the Table 3 align with Tables 1 and Table 2. It highlights our proposed model's efficacy when employing query and wake word speech as the reference for speech assistant applications.

Table 3: Results on vendor-collected test set.

Exp	Model	DEL(%)	WER(%)
E15	Mixture	0.7	81.2
E16	LSTM+Concat	24.1	57.63
E17	Conformer+Concat	21.5	57.51
E18	Conformer+FiLM	19.4	55.77
E19	PVAD-FISV	12.3	30.4
E20	PVAD-FTSA (FFN=512)	1.3	39.6
E21	PVAD-FTSA (FFN=1024)	1.0	35.6

5. CONCLUSIONS

This work explores PVAD models utilizing wake word reference speech. Specifically, we introduce the PVAD-FTSA model, a PVAD model that operates independently of a pre-trained SV model. The PVAD-FTSA model directly employs the reference speech's frame-level feature as the target speaker's attributes. Experimental results highlight the superior performance of our proposed approach over other baseline. Furthermore, our model accomplishes an ultra-high recall rate, a critical aspect for smartphone speech assistant.

6. ACKNOWLEDGEMENT

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

7. REFERENCES

[1] Lee Ngee Tan, Bengt J. Borgstrom, and Abeer Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *Proc. of ICASSP*, 2010, pp. 4466–4469.

[2] Shuo-Yiin Chang, Bo Li, Gabor Simko, Tara N. Sainath, Anshuman Tripathi, Aäron van den Oord, and Oriol Vinyals, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proc. of ICASSP*, 2018, pp. 5549–5553.

[3] Shaojin Ding, Quan Wang, Shuo-Yiin Chang, Li Wan, and Ignacio Lopez Moreno, "Personal VAD: Speaker-Conditioned Voice Activity Detection," in *Proc. of Odyssey*, 2020, pp. 433–439.

[4] Naoki Makishima, Mana Ihori, Tomohiro Tanaka, Akihiko Takashima, Shota Orihashi, and Ryo Masumura, "Enrollment-less training for personalized voice activity detection," in *Proc. of Interspeech*, 2021.

[5] Shaojin Ding, Rajeev Vijay Rikhye, Qiao Liang, Yanzhang He, Quan Wang, Arun Narayanan, Tom O'Malley, and Ian McGraw, "Personal vad 2.0: Optimizing personal voice activity detection for on-device speech recognition," in *Proc. of Interspeech*, 2022.

[6] Aditya Jayasimha and Periyasamy Paramasivam, "Personalizing speech start point and end point detection in asr systems from speaker embeddings," in *Proc. of SLT*, 2021, pp. 771–777.

[7] Zuheng Kang, Jianzong Wang, Junqing Peng, and Jing Xiao, "Svvad: Personal voice activity detection for speaker verification," *Proc. of Interspeech*, 2023.

[8] 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*. IEEE, 2023, pp. 1–5.

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

[10] Finnian Kelly, Andrzej Drygajlo, and Naomi Harte, "Speaker verification in score-ageing-quality classification space," *Computer Speech & Language*, vol. 27, no. 5, pp. 1068–1084, 2013.

[11] Xiaoyi Qin, Na Li, Weng Chao, Dan Su, and Ming Li, "Cross-Age Speaker Verification: Learning Age-Invariant Speaker Embeddings," in *Proc. of Interspeech*, 2022, pp. 1436–1440.

[12] Yiming Wang, Hang Lv, Daniel Povey, Lei Xie, and Sanjeev Khudanpur, "Wake word detection with streaming transformers," in *Proc. of ICASSP*. IEEE, 2021, pp. 5864–5868.

[13] Haoxu Wang, Ming Cheng, Qiang Fu, and Ming Li, "The dku post-challenge audio-visual wake word spotting system for the 2021 misp challenge: Deep analysis," in *Proc. of ICASSP*. IEEE, 2023, pp. 1–5.

[14] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. of AAAI*, 2018, vol. 32.

[15] 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.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. of Interspeech*, pp. 1086–1090, 2018.

[18] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *ArXiv*, vol. abs/1510.08484, 2015.

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

[20] Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, et al., "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *Proc. of ICASSP*, 2022.

[21] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, et al., "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," *arXiv*, 2021.

[22] Xiaoyi Qin, Ming Li, Hui Bu, Shrikanth S. Narayanan, and Haizhou Li, "The 2022 far-field speaker verification challenge: Exploring domain mismatch and semi-supervised learning under the far-field scenario," *Proc. of FFSVC2022*, 2022.

[23] Jianfeng Zhou, Tao Jiang, Zheng Li, Lin Li, and Qingyang Hong, "Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function.," in *Proc. of Interspeech*, 2019, pp. 2883–2887.

[24] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. of Interspeech*, 2021.