Personal Voice Activity Detection With Ultra-Short Reference Speech

Longting Xu*, Mingjun Zhang*, Wenbin Zhang†, Tianyi Wang†, Jiawei Yin†, and YuGao†‡
* College of Information Science and Technology, Donghua University, China
E-mail: xlt@dhu.edu.cn, zmj@mail.dhu.edu.cn
† AI Research Center, Midea Group (Shanghai) Co.,Ltd., Shanghai 201702, China
E-mail: gaoyu11@midea.com
‡ Corresponding Author

***Abstract*—Personal Voice Activity Detection (PVAD) is widely used in applications such as voice assistants. To accurately detect the voice activity of the target speaker, PVAD typically requires pre-registering the target speaker's speech as a reference. However, the excessively long voice enrollment process tends to reduce user motivation. To address this problem, we explore the possibility that PVAD can maintain good performance even with short reference speech. We propose a PVAD network that supports Ultra-Short reference speech, namely US-PVAD. Unlike traditional methods that rely on pre-trained speaker verification models to extract speaker embeddings, US-PVAD allows the direct input of the original reference speech. Since RNN states can memorize historical information and use it to guide subsequent time steps, we employ a DPRNN-based network and use its RNN states as target speaker embedding. This approach eliminates the need for an external speaker embedding extractor with a large number of parameters. Additionally, the RNN states can be continuously updated during voice activity detection, allowing PVAD to obtain sufficient target speaker feature attributes from ultra-short reference speech. Experimental results show that US-PVAD exhibits better performance when using speech under 2 seconds or even as short as 0.2 seconds as the reference speech.**

***Index Terms*—personal VAD, voice activity detection, speaker verification**

I. Introduction

Voice Activity Detection (VAD) [1]–[3] is a technique that classifies audio segments as either speech or non-speech. It is typically a crucial front-end step for various tasks such as speaker verification (SV), emotion estimation, and automatic speech recognition (ASR). A well-trained VAD system can detect all speech segments regardless of the speaker, which is known as standard VAD. The benefits of using standard VAD are multifaceted. Firstly, downstream systems may be highly sensitive to noise, and filtering out these signals can improve system performance. Secondly, in terms of required computational resources, running SV and ASR systems is often quite expensive, especially when dealing with mobile personal devices. Therefore, triggering downstream systems only when necessary through VAD can help limit resource and energy consumption to some extent. However, in scenarios involving multiple users, standard VAD systems often produce false positives, causing the system to be erroneously activated when not needed. Hence, there is a need for Personal Voice Activity Detection (PVAD) systems.

PVAD [4]–[10] is a technique used to identify the speech activity of a target speaker in multi-speaker scenarios. Compared to standard VAD, PVAD models can distinguish between the speech signals of different speakers, making them more suitable for responding to specific user downstream tasks. While traditional cascaded VAD and SV models can handle PVAD tasks, PVAD systems typically require lightweight designs for widespread real-time deployment. However, SV systems have high resource demands, operational costs, and detection delays, making them unsuitable for such applications.

The PVAD system proposed in [4] represents a significant advancement, offering a lightweight solution for robust target speaker speech detection through the Embedding Conditioned Training (ET) architecture. It eliminates the need for heavy SV systems for frame scoring during runtime, simply combining the target speaker's registered embedding with the acoustic features of the input speech. In subsequent research, the authors extended the PVAD architecture, proposing PVAD 2.0 [5] and evaluating its effectiveness in downstream ASR tasks. To address the issue of missing user-specific enrollment information, [6] proposed a PVAD training method with less enrollment. Ref. [11] explored target speaker detection in diarization contexts.

However, these methods tend to favor longer durations when selecting reference speech. While longer reference speech can be used to obtain better speaker embeddings through pre-trained SV models, accurately reflecting the attributes of the target speaker, in practice, excessively long enrollment times can reduce user motivation and result in a poor user experience. Therefore, it is crucial for PVAD to maintain good performance even under shorter enrollment conditions. Ref. [9] proposed a PVAD network using wakeup words as reference speech and achieved good performance. Ref. [10] investigated the effect of reference speech longer than 1 second on their proposed networks, but neither explored how to maintain good performance in PVAD networks when using ultra-short speech as reference.

Furthermore, as mentioned in [9], the performance of pre-trained SV models does not necessarily correlate with the performance of PVAD systems, as they are optimized using different loss functions. Additionally, the quality of embeddings extracted by different SV models for the same reference

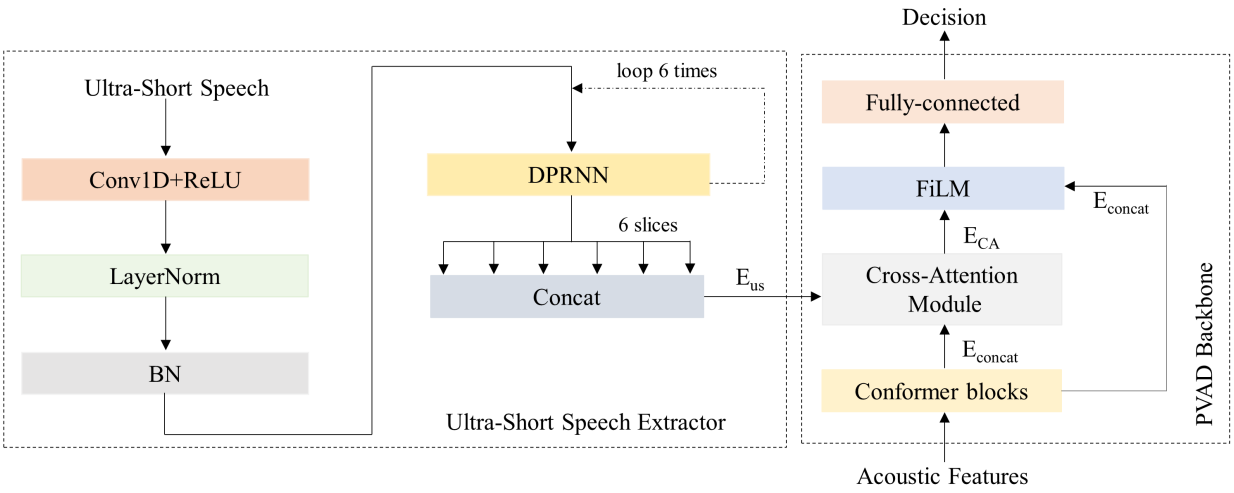


Fig. 1. The Proposed US-PVAD model.

speech can directly impact PVAD system performance [12]. Based on the above, we propose a PVAD system capable of handling ultra-short reference speech without requiring external SV models to extract target speaker embeddings, namely US-PVAD. Since ultra-short speech contains too little speaker information, directly using its acoustic features or speaker embeddings for training may not be optimal. Inspired by [13], we employ a DPRNN-based [14] network to process the raw reference speech, extracting the target speaker's attributes from the continuously updated RNN states during training. This approach avoids dependency on SV models, eliminates the impact of varying embedding quality from different SV models on PVAD performance, and ensures good performance even with very short reference speech.

II. Proposed Methods

The proposed US-PVAD takes acoustic features and the original reference speech as inputs, and provides frame-wise PVAD results as outputs. In this section, we will provide a detailed introduction to it.

A. Ultra-Short Speech Extractor

As shown in Fig. 1, US-PVAD mainly consists of ultra-short speech extractor and PVAD Backbone, which are jointly trained. Inspired by [13], we extract feature information in ultra-short speech using a network based on DPRNN, a well-known network for speech separation [14] and TSE [15]. It splits the input speech into small chunks for modeling long sequences. Its structure, shown in Fig. 2, consists of two main BiGRU layers, the local BiGRU for extracting intra-chunk features and the global BiGRU for extracting inter-chunk features. Since RNN states can memorize historical information and use it to guide the operation of the next time step, and also update it continuously while the network is running. Using RNN states as speaker features means that the feature information of the target speaker can be continuously updated in the VAD step after enrollment. Therefore, only a reference speech with a short duration is needed as a starting condition in the enrollment step, and the final RNN state can preserve its feature information and guide the network in detecting the voice activity of the target speaker.

The initial states of both local and global BiGRUs are set to zero. Since the layer output after the FC layer and Layer norm is more robust and also contains RNN states, we only use the slices of the layer output that contain RNN states, and both global BiGRU and local BiGRU states are discarded. In addition, to obtain more feature attributes from the ultra-short speech, we iteratively update the input 6 times via DPRNN, take the slices of the layer output obtained from each new iteration, and concatenate them to form the final target speaker embedding. The input and output are as follows:

$$Seq_{out}^{K\times P\times R} = DPRNN_i(Seq_{in}^{K\times P\times R}) \tag{1}$$

where Seq_{in} and Seq_{out} are the inputs and outputs of the lth $(i = 0...5)$ DPRNN iteration. The input reference speech is divided into R chunks, each of length P. Since this is a BiGRU layer with two directions, the output in each direction is K dimensional. Each iteration takes a slice of the layer output in the second dimension to get $Seq_{out}^{K\times R}$, and finally the 6 slices are stitched together to get $Seq_{out}^{K\times 6R}$, which serves as the speaker embedding of the ultra-short reference speech.

B. PVAD Backbone

The PVAD backbone first processes the acoustic features of the concatenated speech through a four-layer stacked Conformer block to generate E_{concat}. Next, we input the speaker feature E_{us}, obtained from the ultra-short speech extractor, as the key and value, and E_{concat} as the query, into the cross-attention (CA) module [16]. The output of the CA module is then used to perform a feature-wise affine transformation (FiLM) [17] on E_{concat} to fuse these two sets of features. FiLM includes scaling and shifting operations, where E_{CA} is the output of CA, $\gamma()$ and $\beta()$ are the scaling and shifting

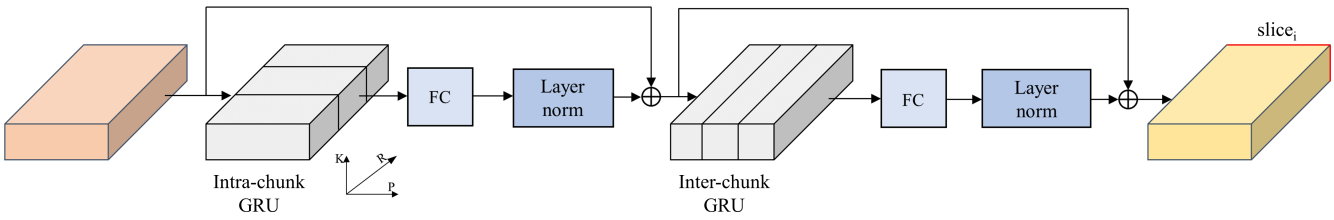


Fig. 2. The DPRNN block structure in US-PVAD. The red line in the output represent the extracted slice, with the slice being concatenated in the P direction.

vectors of FiLM, respectively. The FiLM process can be represented as:

$$FiLM(E_{concat}, E_{CA}) = \gamma(E_{CA}) \cdot E_{concat} + \beta(E_{CA}) \quad (2)$$

Finally, the fused features are fed into a fully connected layer to generate the decision p for the PVAD model. The entire process can be formulated as follows:

$$p = FC(FiLM(Con(F_{concat}), CA(Con(F_{concat}), E_{us})) \quad (3)$$

where Con denotes the Conformer blocks and F_{concat} is the acoustic feature of the concatenated speech.

III. Experimental Setup

A. Datasets

Referring to the setup described in [4] for simulated conversational speech, we created a multi-speaker training set using three main subsets of LibriSpeech [18]: the 100-hour, the 360-hour, and the 500-hour subsets totaling 960 hours and encompassing 2,338 different speakers. This resulted in 312,020 concatenated utterances. Each audio file contains three speakers, with one randomly selected as the target speaker in the concatenated utterance. We used the Montreal Forced Aligner [19] to create VAD ground truth annotations for the concatenated utterances. The PVAD labels for each frame were modified according to the target speaker: "non-speech" frames remained unchanged, while "speech" frames were relabeled as either "target speaker speech" or "non-target speaker speech" depending on whether the speech originated from the target speaker. Additionally, we generated 5,350 concatenated utterances as a test set using LibriSpeech's test-clean and test-other subsets, created in the same manner as the training set. We added noise and reverberation to the created training and test sets using the Musan [20] and RIRs [21] datasets to simulate real-world environments.

B. Implementation details

In the DPRNN block of the ultra-short speech extractor, the chunk size is set to 250 and the hop size is 125. The 1-D convolutional layer has 1 input channel and 256 output channels, with a kernel size of 2 and a stride of 1. The input and output sizes of the batch normalization (BN) are both 256. For the PVAD backbone, the LSTM-based model consists of 2 layers of LSTMs, each layer having 256 units, and the last layer is a fully connected layer that acts as a classifier. The Conformer-based model contains 4 Conformer layers in the block, each having a dimension of 256, an attention head count of 8, causal 7 × 7 convolution kernel, and 31 left context.

We use the pre-trained ResNet34 model provided by SpeechBrain [22] as the SV model in all baseline systems, which has about 15.5 million parameters, and the extracted speaker embedding vectors have a size of 256. In all models, we input acoustic features with the same settings, i.e., 40-dim log Mel-filterbank energy with a frame length of 25 ms and a frame shift of 10 ms. The loss function uses Weighted Pairwise Loss [4], where the weights are set to $w\langle \text{tss}, \text{ns} \rangle = w\langle \text{tss}, \text{ntss} \rangle = 1$ and $w\langle \text{ns}, \text{ntss} \rangle = 0.5$. During the training process, we use the Adam optimizer [23] with an initial learning rate of 5e-4. We use Recall, Precision, F1-score, and average precision (AP) for each class as metrics.

IV. Results

To illustrate that US-PVAD can better handle ultra-short reference speech, we compare our proposed US-PVAD model with several baseline models on reference speech of different lengths in Table 1. **LSTM+Concat** denotes the ET structure proposed in [4], while **Conformer+Concat** and **Conformer+FiLM** are the structures proposed in [5]. The configuration of all models follows the previous description.

A. Comparison of results

From the results, it can be seen that US-PVAD outperforms the baseline systems in all metrics for a reference speech duration of 0.2 seconds. In particular, in the extreme case of additional noise, the REC of the baseline systems are all around 50% (E13-E15), whereas US-PVAD still achieves a level of 60% (E16: 60.91%), even higher than the baseline system with a 2-second speech reference (E1: 57.01%). In terms of overall performance (F1 and AP), US-PVAD not only outperforms the baseline systems but also maintains high detection efficacy for NON-SPEECH (NS) even after adding noise (E16: 96.29%).

Moreover, US-PVAD demonstrates great detection efficacy for TARGET SPEAKER SPEECH (TSS), with a 0.2-second reference speech (E16: 77.65%) outperforming one baseline using more than 1 second of reference speech (E1: 69.75%,

TABLE I

RESULTS COMPARISON WITH SHORT REFERENCE SPEECH (%). REF DENOTES REFERENCE SPEECH OF VARYING DURATIONS. 3 SPEAKERS INDICATE CONCATENATED UTTERANCES FROM THREE DIFFERENT SPEAKERS. NOISE DENOTES ADDING NOISE FROM MUSAN AND RIRS. REC: RECALL. PRE: PRECISION. F1: F1 SCORE. AP REPRESENTS THE AVERAGE PRECISION ACROSS THREE CLASSES: NON-SPEECH (NS), NON-TARGET SPEAKER SPEECH (NTSS), AND TARGET SPEAKER SPEECH (TSS).

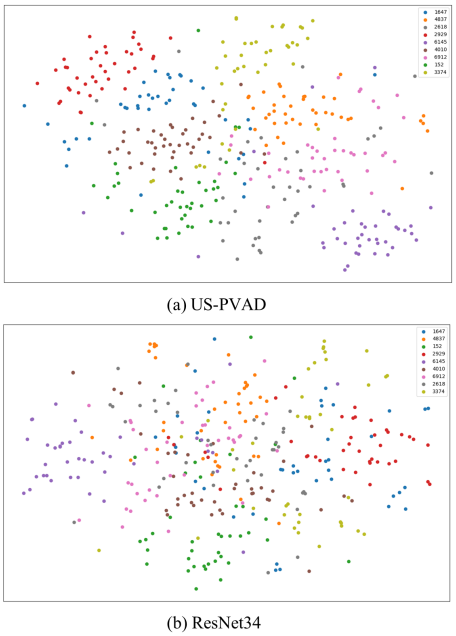
Ref	Exp	Model	Params(M) (PVAD/SV)	3 Speakers				3 Speakers + Noise			
				REC	PRE	F1	AP [NS NTSS TSS]	REC	PRE	F1	AP [NS NTSS TSS]
2s	E1	LSTM+Concat	1.09/15.53	60.45	72.89	78.04	[89.08 88.98 72.62]	57.01	71.27	75.29	[85.59 87.95 69.75]
	E2	Conformer+concat	1.16/15.53	**77.05**	78.65	83.48	[89.55 94.49 86.51]	74.22	78.81	83.08	[92.51 94.02 85.56]
	E3	Conformer+FiLM	1.16/15.53	74.94	81.99	84.26	[90.47 95.33 87.78]	**74.61**	79.48	84.28	[94.34 94.93 86.50]
	E4	US-PVAD	1.93/0.00	76.72	**83.19**	**86.98**	**[97.17 96.19 90.64]**	74.33	**82.64**	**85.75**	**[95.54 96.03 88.81]**
1s	E5	LSTM+Concat	1.09/15.53	52.64	72.07	76.22	[89.80 87.48 68.56]	45.70	70.41	72.55	[86.72 85.53 64.43]
	E6	Conformer+concat	1.16/15.53	73.90	76.01	83.65	[94.83 93.33 83.06]	68.49	73.11	80.58	[93.16 91.45 79.21]
	E7	Conformer+FiLM	1.16/15.53	73.99	80.58	85.38	[95.81 95.37 85.11]	65.95	78.45	81.60	[93.49 93.45 82.49]
	E8	US-PVAD	1.93/0.00	**74.41**	**81.48**	**85.48**	**[96.86 95.39 88.39]**	**70.45**	**78.65**	**83.52**	**[95.54 94.39 84.06]**
0.5s	E9	LSTM+Concat	1.09/15.53	48.01	66.62	73.48	[88.07 84.15 62.73]	44.31	67.52	72.31	[87.04 83.45 61.32]
	E10	Conformer+concat	1.16/15.53	63.29	67.40	77.22	[88.94 88.47 71.12]	59.87	64.14	75.29	[91.13 86.73 67.68]
	E11	Conformer+FiLM	1.16/15.53	67.14	77.02	81.14	[90.43 92.87 81.39]	63.95	75.31	79.42	[91.38 92.23 79.47]
	E12	US-PVAD	1.93/0.00	**73.01**	**79.74**	**84.09**	**[96.75 93.67 85.26]**	**70.27**	**77.59**	**81.88**	**[95.35 92.99 83.23]**
0.2s	E13	LSTM+Concat	1.09/15.53	40.98	58.22	69.20	[86.93 81.73 53.93]	40.77	52.46	67.01	[85.36 76.90 48.51]
	E14	Conformer+Concat	1.16/15.53	52.60	59.95	73.40	[95.53 85.62 65.17]	48.98	54.90	69.67	[85.77 80.34 55.69]
	E15	Conformer+FiLM	1.16/15.53	52.22	70.83	78.60	[96.38 89.16 72.20]	50.36	69.17	76.28	[93.72 88.24 70.57]
	E16	US-PVAD	1.93/0.00	**62.74**	**77.42**	**82.31**	**[97.05 92.24 79.30]**	**60.91**	**76.91**	**80.11**	**[96.29 91.97 77.65]**

E5: 64.43%) and closely matching two other baselines using 1 second of reference speech (E6: 79.21%, E7: 82.49%).

Furthermore, we evaluated the performance of US-PVAD with 1-second and 0.5-second reference speech. Especially for 0.5 seconds, which is a very short duration and can be used as a reference speech in practical applications, US-PVAD performs significantly better than the baseline systems. The results show that the baseline systems underperform US-PVAD in all metrics, especially in the AP metric of TSS, where US-PVAD (E12) using a 0.5-second reference speech outperforms all the baseline systems (E5-E7) using a 1-second reference speech. This suggests that US-PVAD can support ultra-short reference speech well due to the constant updating of the target speaker embedding during the ultra-short speech extraction step.

Additionally, the impact of changing the enrollment duration from 2 seconds of longer speech to 0.2 seconds of ultra-short speech varies across different models. The performance metrics of US-PVAD exhibit a smaller change compared to the baseline model, particularly in noise-free conditions. Specifically, the PRE of US-PVAD only decreases by 5.77% (E4: 83.19% to E16: 77.42%), whereas the baseline model experiences a maximum drop of 18.70% (E2: 78.65% to E14: 59.95%) and a minimum drop of 11.16% (E3: 81.99% to E15: 70.83%). This indicates that US-PVAD maintains more consistent performance when faced with shorter enrollment duration compared to the baseline models.

It is worth noting that US-PVAD (E4) outperforms the baseline models in terms of PRE and overall performance when using a 2-second reference speech, but its efficacy in terms of REC is lower compared to the two baselines (E2, E3). This may be because the baseline systems rely on an external SV model to extract speaker embeddings, and the embeddings extracted from the 2-second reference speech



(a) US-PVAD

(b) ResNet34

Fig. 3. t-SNE results for aggregated embedding vectors of 0.2s speech from 9 speakers. (a) US-PVAD's ultra-short speech extractor; (b) ResNet34.

contain more speaker information, resulting in a slightly higher REC than US-PVAD. Since US-PVAD does not rely on an SV model, and the ultra-short speech extractor and the PVAD backbone network are co-trained, the observed performance difference aligns with our study's expectations. We intend to further investigate this issue in our future research. In addition, compared to the baseline model that relies on the SV model, US-PVAD has significantly fewer total parameters (1.93/0.00), making it more suitable for lightweight deployments. Overall, US-PVAD is able to better handle ultra-short speech while

maintaining good performance for normal-length reference speech.

B. Comparison of t-SNE

The quality of target speaker embeddings extracted from ultra-short speech is critical for the performance of the PVAD backbone. We compared the t-SNE visualizations of embedding vectors extracted by ultra-short speech extractor of US-PVAD and ResNet34. As illustrated in Fig. 3, for 0.2-second speech, the speaker features produced by US-PVAD exhibit clear clustering and distribution according to different speakers, achieving better separation of embedding vectors from different speakers compared to ResNet. This is a key factor contributing to US-PVAD's great performance in handling ultra-short speech.

V. Conclusions

This work explores the PVAD model that maintains good performance even with ultra-short reference speech. Specifically, we propose a PVAD model named US-PVAD, which can handle ultra-short reference speech and operates independently of a pre-trained SV model. US-PVAD utilizes RNN states generated by DPRNN, which can be continuously updated during the PVAD process, as features of the target speaker. Experimental results show that under conditions of short reference speech with varying lengths of less than 2 seconds, our proposed method outperforms other baseline methods and demonstrates greater robustness to variations in enrollment duration.

References

[1] S.-Y. Chang, B. Li, G. Simko, *et al.*, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5549–5553. DOI: 10.1109/ICASSP.2018.8461921.

[2] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4466–4469. DOI: 10.1109/ICASSP.2010.5495611.

[3] J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," *Robust speech recognition and understanding*, vol. 6, no. 9, pp. 1–22, 2007.

[4] S. Ding, Q. Wang, S.-y. Chang, L. Wan, and I. L. Moreno, "Personal vad: Speaker-conditioned voice activity detection," *arXiv preprint arXiv:1908.04284*, 2019.

[5] S. Ding, R. Rikhye, Q. Liang, *et al.*, "Personal vad 2.0: Optimizing personal voice activity detection for on-device speech recognition," *arXiv preprint arXiv:2204.03793*, 2022.

[6] N. Makishima, M. Ihori, T. Tanaka, A. Takashima, S. Orihashi, and R. Masumura, *Enrollment-less training for personalized voice activity detection*, 2021. arXiv: 2106.12132 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2106.12132.

[7] W. Wang, X. Qin, and M. Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for the m2met challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9171–9175.

[8] Z. Kang, J. Wang, J. Peng, and J. Xiao, "Svvad: Personal voice activity detection for speaker verification," *arXiv preprint arXiv:2305.19581*, 2023.

[9] B. Zeng, M. Cheng, Y. Tian, H. Liu, and M. Li, "Efficient personal voice activity detection with wake word reference speech," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 241–12 245.

[10] F. Liu, F. Xiong, Y. Hao, K. Zhou, C. Zhang, and J. Feng, "As-pvad: A frame-wise personalized voice activity detection network with attentive score loss," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 476–11 480. DOI: 10.1109/ICASSP48485.2024.10446581.

[11] M. Cheng, W. Wang, Y. Zhang, X. Qin, and M. Li, "Target-speaker voice activity detection via sequence-to-sequence prediction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[12] A. Aloradi, M. Elminshawi, S. R. Chetupalli, and E. A. Habets, "Target-speaker voice activity detection in multi-talker scenarios: An empirical study," in *Speech Communication; 15th ITG Conference*, VDE, 2023, pp. 250–254.

[13] L. Yang, W. Liu, L. Tan, J. Yang, and H.-G. Moon, "Target speaker extraction with ultra-short reference speech by ve-ve framework," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[14] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50. DOI: 10.1109/ICASSP40776.2020.9054266.

[15] Y. Hao, J. Xu, J. Shi, P. Zhang, L. Qin, and B. Xu, "A unified framework for low-latency speaker extraction in cocktail party environments," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226202888.

[16] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[18] 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.

[19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:12418404.

[20] D. Snyder, G. Chen, and D. Povey, *Musan: A music, speech, and noise corpus*, 2015. arXiv: 1510.08484 [cs.SD]. [Online]. Available: https://arxiv.org/abs/1510.08484.

[21] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224. DOI: 10.1109/ICASSP.2017.7953152.

[22] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, *Speechbrain: A general-purpose speech toolkit*, 2021. arXiv: 2106.04624 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2106.04624.

[23] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1412.6980.