COIN-AT-PVAD: A Conditional Intermediate Attention PVAD

En-Lun Yu[†], Ruei-Xian Chang[‡], Jeih-Weih Hung[§], Shih-Chieh Huang[¶] and Berlin Chen[‖]
[†‖] Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
E-mail: {enlunyu, berlin}@ntnu.edu.tw
[‡§] Department of Electrical Engineering, National Chi Nan University, Taiwan
E-mail: {s110352016, jwhung}@ncnu.edu.tw
[¶] Realtek Semiconductor Corp., Taiwan

***Abstract*—Personalized voice activity detection (PVAD), compared to conventional VAD, shows more developmental potential in scenarios with multiple speaker interference. Among the various methods for integrating speaker and acoustic features, performance may be limited due to the weaker representational capability of speaker embeddings derived from external speaker verification models. This study proposes a new architecture called Conditional Intermediate Attention PVAD (COIN-AT-PVAD) to address this issue. This architecture builds upon the Attentive Score (AS) module and incorporates the Feature-wise Linear Modulation (FiLM) scheme to better integrate multimodal information. Through comparing various fusion strategies, we show that COIN-AT-PVAD significantly surpasses the baseline model, especially when external embedding features have limited representational capacity. Experimental findings also indicate that, when compared to some state-of-the-art models, COIN-AT-PVAD achieves superior average precision and accuracy while retaining a compact model size, showcasing its efficacy in real-world applications on resource-limited devices.**

I. Introduction

Voice Activity Detection (VAD) [1]–[3] is specifically developed to detect and identify speech within received audio signals. It serves as the front-end component for a variety of speech processing applications, including automatic speech recognition (ASR) [4], keyword spotting (KWS) [5], and speech enhancement (SE) [6]. VAD reduces computational load by filtering away non-speech content before further processing, which is particularly important for high-performance systems with heavy computational requirements.

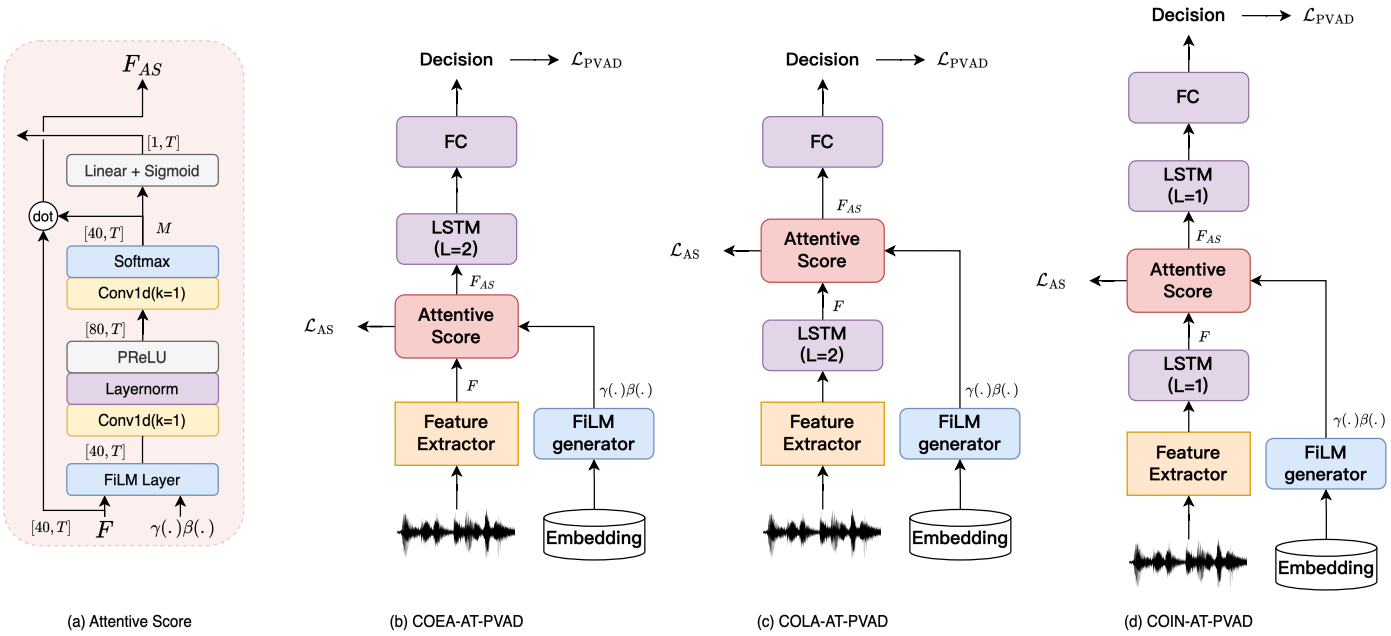
However, personalized speech applications in real-world situations often encounter voices from non-target speakers, which causes conventional VAD systems to generate disruptive false alarms, thereby impairing the performance of downstream tasks. To address this issue, recent research has developed a new variant of VAD known as Personal VAD (PVAD). In contrast to conventional VAD, PVAD is designed to recognize speech specific to a known speaker. This helps reduce interference from other speakers and mitigate false positives commonly seen in conventional VAD systems.

Ding *et al.* pioneered a straightforward PVAD approach in their seminal work [7]. This method involves generating speaker embeddings from enrollment speech via a speaker verification model, concatenating speaker embeddings with handcrafted acoustic features, and employing the concatenation as input features into the VAD framework. The authors further enhanced this system in PVAD 2.0 [8] by mapping speaker embeddings to the same domain as acoustic features and using a linear transformation to fuse the speaker information with acoustic features. The resulting compact input features simplify the subsequent processing stages and enhance PVAD performance.

Building upon the PVAD paradigm, which intertwines speaker information and acoustic data, subsequent studies have explored various methods to improve this integration [9]–[16]. A recent study [9] conducted a comprehensive comparison and analysis of a series of integration methods, including comparing the effects of static fusion at different points in the model pipeline and incorporating dynamic speaker estimation on both performance and complexity. Expanding on this foundation, [10] focused on integrating both pieces of information using Cross-Attention in the acoustic feature space. Furthermore, the work in [11] replaced the external speaker verification model with an internal embedding extractor and introduces an innovative Attentive Score (AS) loss function. This enables the attention score module to derive the attention weight from the concatenated features to draw attention to specific acoustic features.

Although [11] offers an intriguing solution for PVAD, there remains room for further validation and improvement. One of the concerns is that the AS module integrates acoustic and speaker information using a concatenation mechanism, which may be too simple and ineffective.

This study introduces a novel architecture dubbed Conditional Intermediate Attention PVAD (COIN-AT-PVAD) to address the above concern with the AS module. This architecture refines the AS module by conditionally fusing its multimodal inputs in order to improve its performance. In addition, we provide multiple options for locating the the presented novel AS module within the PVAD process. The experimental results indicate that the new COIN-AT-PVAD surpasses the baseline model and some state-of-the-art PVAD frameworks in performance while requiring fewer model parameters.



(a) Attentive Score (b) COEA-AT-PVAD (c) COLA-AT-PVAD (d) COIN-AT-PVAD

Fig. 1. Diagrams of (a) the advanced Attentive Score module; (b) Conditional Early Attention PVAD; (c) Conditional Latent Attention PVAD; (d) Conditional Intermediate PVAD.

II. Methods

A. Attentive PVAD with Conditioned Fusion

To begin with, we briefly review the AS module of [11]. Following that, we introduce a variant of the AS module that incorporates conditional fusion of input features.

Liu *et al.* proposed an AS module for PVAD that combines features from the target (enrollment) speaker and input utterances. This module extracts a similarity score between both features through two convolutional layers:

$$\hat{\mathbf{F}}_t = [\mathbf{F}_t; \mathbf{e}^{\text{target}}], \tag{1}$$

$$\mathbf{M}_t = \text{AS}(\hat{\mathbf{F}}_t), \tag{2}$$

where $\mathbf{F}_t$ and $\mathbf{e}^{\text{target}}$ represent the input acoustic feature at time frame t and target speaker embedding, respectively, and $\text{AS}(\cdot)$ denotes the overall function responsible for computing the similarity score $\mathbf{M}_t$ in the AS module.

Subsequently, by performing an element-wise multiplication of the similarity score $\mathbf{M}_t$ with $\mathbf{F}_t$, a weighted feature $\mathbf{F}_{\text{AS},t}$ is obtained:

$$\mathbf{F}_{\text{AS},t} = \hat{\mathbf{F}}_t \odot \mathbf{M}_t. \tag{3}$$

However, the use of concatenation as in (1) to combine speaker and acoustic information may have drawbacks: 1) It increases the feature dimensions, leading to a parameter-intensive model; 2) These two types of information are of different modalities, and thus concatenating them limits the model's learning capability.

Among fusion strategies, feature-wise Linear Modulation (FiLM) [17] has been demonstrated in multiple studies to effectively combine multimodal information and reduce the number of feature parameters [8]–[10]. Therefore, we propose to introduce FiLM into the AS module, as shown in Fig. 1(a), to obtain the conditioned feature $\tilde{\mathbf{F}}_t$, which replaces the concatenated feature $\hat{\mathbf{F}}$ in (1). The FiLM-wise conditioned feature $\tilde{\mathbf{F}}_t$ is obtained by:

$$\tilde{\mathbf{F}}_t = \text{FiLM}(\mathbf{F}_t) = \gamma(\mathbf{e}^{\text{target}}) \cdot \mathbf{F}_t + \beta(\mathbf{e}^{\text{target}}). \tag{4}$$

Here, γ and β represent the scaling vector and biasing vector, respectively, and they are produced by the FiLM generator using the target speaker embedding $\mathbf{e}^{\text{target}}$ as conditional information.

B. Fusion Strategies for Attention PVAD

According to subsection II-A, the advanced AS module learns the speaker-conditioned mask $\mathbf{M}_t$ and applies it to the input acoustic features $\mathbf{F}_t$. The location arrangement of this AS module in the PVAD pipeline inevitably influence the performance. Partially Inspired by the different fusion manners in [9], here we propose three variants of attentive PVAD (AT-PVAD), which mainly differ in the positioning of the advanced AS module inside the PVAD process, and their flowcharts are depicted in Figs. 1(b), 1(c), and 1(d). The underlying attention strategies are explained as follows:

- **Conditional Early Attention (COEA):** In this method, speaker and acoustic information are combined through conditional fusion at the early stage of the PVAD pipeline. The output of AS module is then used to train the subsequent classification module, which consists of a two-layer LSTM and two fully connected layers (FC). This method directly modulates the raw acoustic features and

learn a deeper classifier. Therefore, if the information from the acoustics and the speaker differs significantly, it may limit the fusion and lead to suboptimal results.

- **Conditional Latent Attention (COLA):** Similar to many other architectures that employ conditional fusion, COLA incorporates conditional fusion at a later stage in the model pipeline, specifically before the fully connected layers which serve as a simple classifier. This enables better training of the two-layer LSTM and the FiLM generator in front of the AC module to effectively integrate multimodal information across speaker and acoustics compared to COEA.
- **Conditional Intermediate Attention (COIN):** Here, the AS module is incorporated between two one-layer LSTMs. This configuration allows the front LSTM to capture improved acoustic features, while the back LSTM aids in the final classification. Clearly, COIN is intended to leverage the strengths of COEA and COLA in order to achieve superior results without increasing computational overhead.

C. Loss Function

As with other PVAD works, we employ the averaged binary cross entropy (BCE) [18] as one source of the objective function for the PVAD binary classification task, which is calculated as follows:

$$\mathcal{L}_{\text{PVAD}} = \frac{1}{T}\sum_{t=0}^{T-1} \text{BCE}(y_t, p_t). \tag{5}$$

Here, $y_t \in \{0, 1\}$ and $p_t \in [0, 1]$ respectively represent the ground-truth label and the PVAD model prediction outcome at time frame t, and T is the total number of time frames.

Additionally, based on [11], we introduce the AS loss function $\mathcal{L}_{\text{AS}}$ to specifically learn the AS module. The AS loss encourages the AS module to focus on learning the similarities across multimodal information by calculating the Mean Squared Error (MSE) loss between the ground-truth label y_t and the estimated weighted score $\tilde{m}_t$:

$$\tilde{m}_t = \text{Sigmoid}(\text{Linear}(\mathbf{M}_t)), \tag{6}$$

$$\mathcal{L}_{\text{AS}} = \frac{1}{T}\sum_{t=0}^{T-1} (y_t - \tilde{m}_t)^2, \tag{7}$$

where $\mathbf{M}_t$ is the similarity score shown in Eq. (2), and Sigmoid and Linear denote the sigmoid-layer and linear-layer operations, respectively. Finally, the above two losses are added together to be the total loss used to train the entire PVAD network:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PVAD}} + \mathcal{L}_{\text{AS}}. \tag{8}$$

III. Experimental Setup

A. Dataset

The LibriSpeech corpus [19] is employed as the source data to evaluate the presented PVAD systems. The training set consists of three subsets, totaling 960 hours of speech data from 2,338 different speakers: the training subsets train-clean-100 and train-clean-360 provide a total of 460 hours of clean speech, while train-other-500 provides 500 hours of noisy speech. Similarly, the LibriSpeech test set includes both clean and noisy speech, totaling 10 hours of speech from 73 speakers. To proceed with the experiments for PVAD, each individual utterance in the LibriSpeech corpus cannot be directly used since it just corresponds to a single speaker. Therefore, we prepare a dataset that contains concatenated utterances from multiple speakers using this corpus. We choose utterances from one to three speakers at random following a uniform distribution and concatenate them. For each concatenated utterance, one of the speakers is randomly chosen as the target speaker. In addition, we randomly select the concatenated sentences with a probability of 0.2 and replace their corresponding target speaker embeddings with the embeddings from non-target speakers that are absent from those sentences. This arrangement aims to prevent the model from detecting any target speaker activity for these sentences in training. It also improves the model's generalization capabilities and prevents biased outputs, which could result in high false-positive rates.

Regarding speaker embedding, utterances from each speaker are randomly selected and fed into a pre-trained speaker verification model to generate window-level 256-dim d-vectors. These d-vectors are then L2-normalized and averaged to produce the utterance-level d-vector, which serves as the target speaker embedding $\mathbf{e}^{\text{target}}$. Additionally, to prevent overfitting and improve the model's robustness, we employ the MTR data augmentation technique [20]. This technique introduces random noise sources with different room impulse responses, effectively improving the model's simulation of various noise and reverberation conditions. Consequently, the model is expected to perform well in a broader range of real-world environments.

B. Implementation details

We extract 40-dimensional log Mel-filterbank energies as raw acoustic features from utterances with a frame size of 25 ms and a step of 10 ms. For three AT-PVAD variants presented, the model details are as follows:

1) COEA: The 2-layer (front) LSTM contains 64 cells for each layer.
2) COLA: The 2-layer (back) LSTM contains 40 cells for each layer.
3) COIN: The one-layer front LSTM contains 40 cells, and the one-layer back LSTM contains 64 cells.

Additionally, we prepare an **AT-PVAD baseline model**, which is nearly identical to the COEA-AT-PVAD structure, with the exception that it employs **the concatenation of raw acoustic features (40-dim Mel-filterbank energies) and target speaker embedding as the AS module input**. This baseline model allows us to examine whether conditional fusion outperforms direct concatenation for the two sources of information. Finally, all AT-PVAD structures end with two fully connected layers that serve as the classifier.

TABLE I
AVERAGE PRECISION (AP) FOR THE TARGET-SPEAKER SPEECH (TSS), ACCURACY AND MODEL SIZE OF THE AT-PVAD VARIANTS.

Model	AP (**tss**)	Accuracy (%)	Parameters (k)
AT-PVAD baseline	0.868	83.79	84.949
COIN-AT-PVAD	**0.912**	**86.74**	71.869
COLA-AT-PVAD	0.901	86.27	**55.285**
COEA-AT-PVAD	0.903	86.08	92.029

TABLE II
AVERAGE PRECISION (AP) FOR THE TARGET-SPEAKER SPEECH (TSS), ACCURACY AND MODEL SIZE OF THE ORIGINAL PVAD, PVAD 2.0, THE AT-PVAD BASELINE AND COIN-AT-PVAD

Model	AP (**tss**)	Accuracy (%)	Parameters (k)
PVAD	0.884	84.34	130.242
PVAD 2.0	0.908	86.56	97.602
AT-PVAD baseline	0.868	83.79	84.949
COIN-AT-PVAD	**0.912**	**86.74**	**71.869**

We implemented all models using PyTorch [21]. For model training, we initially employed the Adam optimizer [22] with a learning rate of 1×10^{-3} for the first epoch, and subsequently reduced the learning rate to 1×10^{-5} for the following epochs.

IV. RESULTS AND DISCUSSIONS

We evaluate PVAD models with a variety of metrics. Accuracy (%) is calculated as the ratio of the number of correctly detected frames to the total number of detected frames. It serves as an evaluation metric to determine how well the PVAD system identifies target speaker speech versus non-target speaker speech frames. Average Precision (AP) is the area under the Precision-Recall Curve with respect to the target-speaker speech (tss). The metric AP is important because the PVAD dataset contains fewer positive samples than negative ones. In addition, we assess the suitability of models for resource-limited devices by taking into account the number of model parameters.

A. Comparison of the AT-PVAD methods

Table I displays the results of the AT-PVAD baseline and three proposed AT-PVAD variants for comparison and analysis. From this table, we have the following observations:

1) The three AT-PVAD variants behave better than the AT-PVAD baseline model, revealing that the conditioned fusion is a better way than the concatenation to merge speaker embeddings and acoustic features as the input to the AS module.
2) COIN-AT-PVAD exhibits the optimal performance among the three AT-PVAD variants in the two metrics, AP and Accuracy, and it has fewer model parameters compared to the AT-PVAD baseline. The central placement of the AS module in the PVAD pipeline provides COIN-AT-PVAD with potential advantages:
   - The front LSTM in the model acts like an acoustic encoder, enhancing the model's non-linear capability to achieve better feature representation. This allows the speaker information to be more effectively integrated with the enhanced acoustic features through the collaborative operation of the FiLM generator.
   - As the AS module highlights the target speaker's characteristics in the deep acoustic features, the back LSTM captures complex patterns and temporal relationships of features before final classification, further enhancing feature representation.
3) COLA-AT-PVAD has the fewest model parameters while it behaves closely to COEA-AT-PVAD. The smaller size of the LSTM used in COLA-AT-PVAD makes it more suitable for deployment on resource-limited devices.

B. Comparison of AT-PVAD and the SOTA PVAD

Table II includes the evaluation results of the original PVAD [7], PVAD 2.0 [8], the AT-PVAD baseline and the newly presented COIN-AT-PVAD for a more comprehensive comparison. As a note, PVAD 2.0 mainly revises the original PVAD by introducing an FiLM layer to make the acoustic feature conditioned with speaker embedding. We have some findings from this table:

1) Equipped with the FiLM layer, PVAD 2.0 and the novel COIN-AT-PVAD behaves better than the original PVAD and AT-PVAD baseline. In particular, COIN-AT-PVAD outperforms PVAD 2.0 in all of the three metrics by providing moderate better AP and Accuracy scores but requiring much fewer model parameters (71.869k versus 97.602k). Thus COIN-AT-PVAD offers substantial benefits for practical applications.
2) The AT-PVAD baseline, which grabs the idea of the AS module in AS-pVAD [11], does not perform as well as expected. It can be attributed to the fact that, compared to the AT-PVAD baseline, AS-pVAD utilizes a more advanced speaker embedding extractor (ECAPA-TDNN) and acoustic feature encoder (TCNN) to offer superior input representation for the AS module.

V. CONCLUSIONS

In this study, we propose a novel PVAD framework which adopts an advanced attentive score module. This module uses acoustic features conditioned on the target speaker embedding to generate the attention weight. We have conducted comparative analyses on the PVAD variants that incorporate this conditional attention mechanism at the early, mid, and late stages of the PVAD pipeline. The experimental results show that all the PVAD variants equipped with conditional attention perform well. Specifically, the PVAD with mid-stage conditional attention (COIN-AT-PVAD) shows excellent performance and has a compact model size, making it suitable for deployment on resource-constrained devices.

ACKNOWLEDGMENT

This work was supported by Realtek Semiconductor Corporation. Any findings and implications in the paper do not

necessarily reflect those of the sponsor.

REFERENCES

[1] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[2] S.-Y. Chang, B. Li, T. N. Sainath, G. Simko, and C. Parada, "Endpoint Detection Using Grid Long Short-Term Memory Networks for Streaming Speech Recognition," *Interspeech 2017*, 2017.

[3] M. Shannon, G. Simko, S.-Y. Chang, and C. Parada, "Improved End-of-Query Detection for Streaming Speech Recognition," *Interspeech 2017*, 2017.

[4] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-End Automatic Speech Recognition Integrated With CTC-Based Voice Activity Detection," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[5] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[6] Y. G. Thimmaraja, B. G. Nagaraja, and H. S. Jayanna, "Speech enhancement and encoding by combining SS-VAD and LPC," *Int J Speech Technol*, 2021.

[7] S. Ding, Q. Wang, S.-y. Chang, L. Wan, and I. L. Moreno, "Personal VAD: Speaker-Conditioned Voice Activity Detection," *Speaker Odyssey*, 2020.

[8] S. Ding, R. Rikhye, Q. Liang, *et al.*, "Personal VAD 2.0: Optimizing Personal Voice Activity Detection for On-Device Speech Recognition," *Proc. Inrespeech*, 2022.

[9] S. Kumar, S. S. Buddi, U. O. Sarawgi, *et al.*, "Comparative Analysis of Personalized Voice Activity Detection Systems: Assessing Real-World Effectiveness," *Proc. Interspeech*, 2024.

[10] B. Zeng, M. Cheng, Y. Tian, H. Liu, and M. Li, "Efficient Personal Voice Activity Detection with Wake Word Reference Speech," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[11] 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," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[12] I. Medennikov, M. Korenevsky, T. Prisyach, *et al.*, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," *Proc. Interspeech*, pp. 274–278, 2020.

[13] A. Jayasimha and P. Paramasivam, "Personalizing Speech Start Point and End Point Detection in ASR Systems from Speaker Embeddings," *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 771–777, 2021.

[14] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe, "Target-speaker Voice Activity Detection with Improved I-Vector Estimation for Unknown Number of Speaker," *arXiv preprint arXiv:2108.03342*, 2021.

[15] N. Makishima, M. Ihori, T. Tanaka, A. Takashima, S. Orihashi, and R. Masumura, "Enrollment-less training for personalized voice activity detection," *Interspeech 2021*, 2021.

[16] E.-L. Yu, K.-H. Ho, J.-W. Hung, S.-C. Huang, and B. Chen, "Speaker Conditional Sinc-Extractor for personal VAD," *Proc. Inrespeech*, 2024.

[17] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," *AAAI*, 2018.

[18] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2018.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[20] C. Kim, A. Misra, K. Chin, *et al.*, "Generation of Large-Scale Simulated Utterances in Virtual Rooms to Train Deep-Neural Networks for Far-Field Speech Recognition in Google Home," *Proc. Interspeech 2017*, pp. 379–383, 2017.

[21] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2019.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.