Enrollment-less training for personalized voice activity detection

Naoki Makishima, Mana Ihori, Tomohiro Tanaka,
Akihiko Takashima, Shota Orihashi, Ryo Masumura

NTT Media Intelligence Laboratories, NTT Corporation, Japan

Abstract

We present a novel personalized voice activity detection (PVAD) learning method that does not require enrollment data during training. PVAD is a task to detect the speech segments of a specific target speaker at the frame level using enrollment speech of the target speaker. Since PVAD must learn speakers' speech variations to clarify the boundary between speakers, studies on PVAD used large-scale datasets that contain many utterances for each speaker. However, the datasets to train a PVAD model are often limited because substantial cost is needed to prepare such a dataset. In addition, we cannot utilize the datasets used to train the standard VAD because they often lack speaker labels. To solve these problems, our key idea is to use one utterance as both a kind of enrollment speech and an input to the PVAD during training, which enables PVAD training without enrollment speech. In our proposed method, called enrollment-less training, we augment one utterance so as to create variability between the input and the enrollment speech while keeping the speaker identity, which avoids the mismatch between training and inference. Our experimental results demonstrate the efficacy of the method.

Index Terms: voice activity detection, personalized voice activity detection, data augmentation

1. Introduction

Voice activity detection (VAD) is a technique to classify an acoustic segment into speech or non-speech, which is an important frontend step in a wide range of tasks such as speaker verification [1,2], emotion estimation [3], and automatic speech recognition [4]. Although many strategies have been proposed for VAD such as time-domain-energy-based methods and likelihood-ratio-based methods [5–8], fully neural network based methods have shown promising performance even under low signal-to-noise ratio (SNR) environments [9–16].

The VAD system addressed in these studies is trained so as to detect all speech segments regardless of the speaker, which is referred to as standard VAD. On the other hand, studies have also looked at a VAD system that detects the voice activity of a specific target speaker at the frame level, which is referred to as personalized VAD (PVAD) [17–19]. Compared with standard VAD, one strength of PVAD is that it does not detect the voice included in background noise that often leads to an unexpected response or an error of the downstream tasks in a real environment. Moreover, since PVAD models can distinguish the speaker of the speech signals, they are suitable for downstream tasks that are intended to respond to a specific user, e.g., a speech recognition system that responds only to the user's voice.

In PVAD, an utterance-level target speaker embedding such as an i-vector [20] or a d-vector [21,22] is required to determine whether the input frame is the target speaker speech [18, 19]. The target speaker speech to obtain such an embedding is called enrollment speech. Note that a delay in the PVAD is almost the same as that of standard VAD because PVAD does not require the utterance-level information of the *input speech* unlike the systems that directly combine standard VAD and the extraction of detected speech features [21–23] for speaker verification.

The enrollment speech is directly set by the user in inferences. On the other hand, to use such speech during training, we have to prepare large-scale datasets that contain many utterances for each speaker, and this is referred to as enrollment-full data. Since speech characteristics such as the intonation or tone vary a little depending on the content of utterances, these utterances are used to learn such diversities, which is crucial to decide the boundaries between speakers. However, preparing enrollment-full datasets in real environments requires substantial cost. In addition, we cannot utilize real environment datasets used to train the standard VAD for training of the PVAD model because they often lack speaker labels and cannot provide enrollment speech, which is referred to as enrollment-less data. Thus, datasets to train a PVAD model are often limited, and using such easily available enrollment-less datasets are indispensable for training of a robust PVAD model.

To solve these problems, we design a PVAD model that is trained without enrollment speech, i.e., we can train our PVAD model only with abundantly available enrollment-less datasets. Our key idea is to use one utterance simultaneously for both an enrollment speech and an input to the PVAD during training. Obviously, this strategy causes a mismatch between training and inference because the target speaker speech in the input and the enrollment speech are different when inferring in a real environment. Therefore, overfitting must be avoided, and the diversity of the target speaker embeddings needs to be learned.

In this paper, we present a new PVAD training method that does not use enrollment speech during training, which is referred to as enrollment-less training. To learn the diversity of the target speaker embeddings from enrollment-less datasets, we introduce enrollment augmentation that combines SpecAugment [24] and dropout [25] to the target speaker embeddings. The key point of the enrollment augmentation is that we should keep the speaker identity while producing the diverse target speaker embeddings. We experimentally show that the combination of SpecAugment and dropout creates diverse target speaker embeddings resembling the ones generated from many utterances of the target speaker in terms of cosine similarity, which is often used to calculate the similarity between speaker embeddings [21, 26]. To the best of our knowledge, this is the first study that trains PVAD without using enrollment speech. Our experimental results show that the PVAD model trained with the proposed method using enrollment-less datasets achieves higher performance than the one trained with the conventional method using enrollment-full datasets. Moreover, we experimentally show that the enrollment augmentation is crucial for the proposed enrollment-less training.

2. Conventional methods

2.1. Standard voice activity detection

We denote the time-frequency representation of an input and its standard VAD state sequence as $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and $\boldsymbol{s} = (s_1, \ldots, s_T)$, respectively, where T denotes the length of the input, $\boldsymbol{x}_t \in \mathbb{R}^D$ denotes the tth frame of the input such as log mel-scale filterbank coefficients, D denotes its dimension, and $s_t \in \mathbb{R}$ represents that the tth frame is the speech state when $s_t = 1$ and that it is the non-speech state when $s_t = 0$. Standard VAD demands the estimation of the conditional probability of $\boldsymbol{s}$ given $\boldsymbol{X}$ in a frame-by-frame manner as

$$P(\boldsymbol{s}|\boldsymbol{X}, \theta) = \prod_{t=1}^{T} P(s_t|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \theta), \qquad (1)$$

where

$$P(s_t|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \theta) = [F(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \theta)]_{s_t}, \qquad (2)$$

$F(\cdot) \in \mathbb{R}^2$ denotes the neural network based function that outputs the conditional probability of each state, θ denotes its parameter, and $[\cdot]_{s_t}$ is the s_tth element of the vector. A long short-term memory (LSTM) network is often used as $F(\cdot)$ to handle long-term dependencies between input features [27]. Cross entropy (CE) is often used as a loss function to estimate the parameter θ that maximizes (1) as

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{X}, \boldsymbol{s}) \in \mathcal{D}} \sum_t \log P(s_t|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \theta), \qquad (3)$$

where $\mathcal{D}$ denotes the training minibatch for standard VAD training, and $|\mathcal{D}|$ denotes its size.

2.2. Personalized voice activity detection

We denote the time-frequency representation of an input, an enrollment speech, and the corresponding PVAD state sequence of the input as $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_T)$, $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J)$, and $\boldsymbol{q} = (q_1, \ldots, q_T)$, respectively, where $\tilde{\boldsymbol{x}}_t \in \mathbb{R}^D$ is the tth frame of the input acoustic feature for PVAD, $\boldsymbol{y}_j \in \mathbb{R}^D$ is the jth frame of the acoustic feature of the enrollment speech, J denotes the enrollment speech length, and $q_t \in \mathbb{R}$ represents that the tth frame is the target speech state when $q_t = 1$ and that it is the non-speech state or the non-target speech state when $q_t = 0$. Note that $\tilde{\boldsymbol{X}}$ contains the speech of both the target speaker and the non-target speaker to learn to avoid detecting non-target speaker speech. PVAD estimates the conditional probability of $\boldsymbol{q}$ given $\tilde{\boldsymbol{X}}$ and $\boldsymbol{Y}$ as

$$P(\boldsymbol{q}|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \theta_{\mathrm{P}}) = \prod_{t=1}^{T} P(q_t|\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_t, \boldsymbol{Y}, \theta_{\mathrm{P}}), \qquad (4)$$

where

$$
\begin{aligned}
P(q_t|\tilde{\boldsymbol{x}}_1, \ldots, &\tilde{\boldsymbol{x}}_t, \boldsymbol{Y}, \theta_{\mathrm{P}}) = \\
&[\mathrm{PVAD}(\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_t, \boldsymbol{e}_{\mathrm{target}}; \theta_{\mathrm{P}})]_{q_t}, \quad (5)
\end{aligned}
$$

$$\boldsymbol{e}_{\mathrm{target}} = \mathrm{SpeakerModel}(\boldsymbol{Y}; \theta_{\mathrm{fix}}), \qquad (6)$$

$\mathrm{PVAD}(\cdot)$ denotes the PVAD model, θ_{P} denotes its parameter, $\boldsymbol{e}_{\mathrm{target}} \in \mathbb{R}^K$ denotes the target speaker embedding, K denotes its dimension, $\mathrm{SpeakerModel}(\cdot)$ denotes the pretrained speaker model, and θ_{fix} denotes its parameter that is pretrained

for speaker verification. Note that θ_{fix} is often fixed during the training of PVAD. CE is often used in PVAD as a loss function to estimate the parameter θ_{P} that maximizes (4) as

$$\mathcal{L}_{\mathrm{CEP}} = -\frac{1}{|\mathcal{D}_{\mathrm{P}}|} \sum_{(\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \boldsymbol{q}) \in \mathcal{D}_{\mathrm{P}}} \sum_t \log P(q_t|\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_t, \boldsymbol{Y}, \theta_{\mathrm{P}}), \qquad (7)$$

where $\mathcal{D}_{\mathrm{P}}$ denotes the training minibatch for the PVAD training.

3. Proposed PVAD

3.1. Strategy

In most personalized voice activity detection, a substantial amount of enrollment speech is required during training to learn the diversity of utterances and the boundary between speakers. However, as described in section 1, it limits the available dataset used for PVAD, which degenerates the PVAD performance. Our method addresses this problem using one utterance simultaneously for both an enrollment speech and an input during training, i.e., we use the standard VAD dataset $\mathcal{D}$ instead of $\mathcal{D}_{\mathrm{P}}$ to train the PVAD model.

In inferences, we use enrollment speech to estimate $\boldsymbol{q}$ by (5) and (6), as shown in Fig. 1 (a). On the other hand, since enrollment speech is not available during training, we simulate the PVAD training as follows. First, multiple utterances in the standard VAD dataset are concatenated to create input $\tilde{\boldsymbol{X}}$, which includes multiple speakers' speech, as shown in Fig. 1 (b). Next, we consider one utterance before concatenation as target speaker speech, and we use it instead of enrollment speech, i.e., the target speech and the enrollment speech are identical. Since this training causes the mismatch between training and inference, we augment the enrollment speech using enrollment augmentation, which is shown as a blue block in Fig. 1 (b).

The enrollment augmentation consists of three procedures: SpecAugment [24], the estimation of the target speaker embedding, and dropout [25]. First, we mask random frequency bands of the speech using SpecAugment, which enables the downstream speaker model to output the diverse speaker embeddings, each focusing on particular frequencies. Then, a pretrained speaker model outputs the speaker embedding of the masked speech. Finally, dropout is applied to the speaker embedding received from the speaker model. Since dropout is effective to prevent the overfitting [25], we use it to prevent PVAD from learning that the target speaker embedding is created from part of the input speech. Although this enrollment augmentation is crucial for the proposed method, we assume that it is effective even when the enrollment-full dataset $\mathcal{D}_{\mathrm{P}}$ is used during training. This is because the role of the augmentation is to create the speech variation and it can further increase the variation of enrollment speech in $\mathcal{D}_{\mathrm{P}}$. We discuss it in section 4.3.

3.2. Enrollment-less training

In this section, we formulate the training of the proposed method. Since the enrollment speech $\boldsymbol{Y}$ is not available during training, we use one utterance $\boldsymbol{X}$ in the standard VAD dataset to obtain $\boldsymbol{e}_{\mathrm{target}}$. As described in section 3.1, we create $\tilde{\boldsymbol{X}}$ and $\boldsymbol{q}$ by concatenating multiple utterances and their state sequences w.r.t. the target speaker in the enrollment-less dataset. In our method, PVAD is trained to estimate the conditional probability
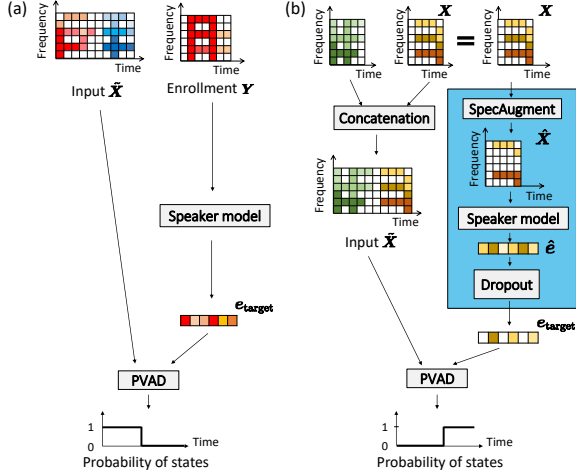
Figure 1: *Overview of proposed method in (a) inference and (b) training. Red, blue, green, and yellow slots in acoustic feature represent speech of different speakers.*

of $\boldsymbol{q}$ given $\tilde{\boldsymbol{X}}$ and $\boldsymbol{X}$ as

$$P(\boldsymbol{q}|\tilde{\boldsymbol{X}},\boldsymbol{X},\theta_{\mathrm{P}}) = \prod_{t=1}^{T} P(q_t|\tilde{\boldsymbol{x}}_1,\ldots,\tilde{\boldsymbol{x}}_t,\boldsymbol{X},\theta_{\mathrm{P}}). \quad (8)$$

First, SpecAugment [24] is applied to $\boldsymbol{X}$ to obtain the frequency masked speech $\hat{\boldsymbol{X}}$. Given $\hat{\boldsymbol{X}}$, the speaker model outputs the speaker embedding as

$$\hat{\boldsymbol{e}} = \mathrm{SpeakerModel}(\hat{\boldsymbol{X}};\theta_{\mathrm{fix}}). \quad (9)$$

We obtain $\boldsymbol{e}_{\mathrm{target}}$ by applying the dropout of ratio p to $\hat{\boldsymbol{e}}$. Given $\boldsymbol{e}_{\mathrm{target}}$, the conditional probability of $\boldsymbol{q}$ is obtained as

$$P(q_t|\tilde{\boldsymbol{x}}_1,\ldots,\tilde{\boldsymbol{x}}_t,\boldsymbol{X},\theta_{\mathrm{P}}) = \\ [\mathrm{PVAD}(\tilde{\boldsymbol{x}}_1,\ldots,\tilde{\boldsymbol{x}}_t,\boldsymbol{e}_{\mathrm{target}};\theta_{\mathrm{P}})]_{q_t}. \quad (10)$$

We denote the dataset created from $\mathcal{D}$ as $\mathcal{D}'$. The loss function to train θ_{P} is obtained as

$$\mathcal{L} = -\frac{1}{|\mathcal{D}'|} \sum_{(\tilde{\boldsymbol{X}},\boldsymbol{X},\boldsymbol{q})\in\mathcal{D}'} \sum_{t} \log P(q_t|\tilde{\boldsymbol{x}}_1,\ldots,\tilde{\boldsymbol{x}}_t,\boldsymbol{X},\theta_{\mathrm{P}}). \quad (11)$$

3.3. Implementation

This section describes the architecture of the pretrained speaker model and the PVAD model. The speaker model consists of three bidirectional LSTM (BLSTM) layers with 256 units. The forward and backward output of the BLSTM are concatenated. The third BLSTM output is averaged using an attention mechanism that calculates the weighted mean of the frame-level features [28] to get the target speaker embedding.

The input to the PVAD model is a concatenation of the target speaker embedding and the input acoustic feature in a frame-wise manner. The PVAD model consists of four LSTM layers followed by a dropout with a ratio of 0.5, one linear layer, and softmax activation. The number of units of all the LSTM layers and the linear layer is 256 and 2, respectively.

4. Experiments

4.1. Dataset

We used a large-scale Japanese speech dataset for an evaluation. We prepared a home-made 380-hour enrollment-less speech dataset $\mathcal{D}$ and a home-made 90-hour enrollment-full speech dataset $\mathcal{D}_{\mathrm{P}}$ for the training data. The validation data of the enrollment-less dataset and enrollment-full dataset consisted of 25 hours and 5 hours of data, respectively. The enrollment-full dataset was used to train the PVAD model with the conventional method [18]. Although the enrollment-full dataset can also be used for the proposed method, we used only the enrollment-less dataset to train the model with the proposed method for a simple comparison. The number of speakers in enrollment-full data was 256 and the number of utterances per speaker was 300 on average. We created a noisy speech dataset by synthesizing 120 manually-constructed noise types (car, shopping mall, factory, etc.) into the clean speech data. SNR levels were randomly varied between 5 to 30 dB. We used both clean speech and noisy speech data for the training.

We prepared five thousands pairs of clean input, the enrollment speech, and the PVAD state sequences for the testing. The average length of one utterance of a speaker was 4 seconds, and the input included one to three speaker utterances without overlapping. The length of the enrollment speech was 8 seconds on average, and the sentences were not included in the input. We also prepared a noisy speech dataset by corrupting the clean input with two unseen noise types (a crowd and a train station) at four noise levels: i.e., with SNR values of 5, 10, 15, and 20 dB. The sampling rate of the entire dataset was 16 kHz.

4.2. Settings

First, we compared the proposed PVAD trained using the enrollment-less dataset (enroll-less w/ aug.) and the conventional PVAD trained using the enrollment-full dataset (enroll-full, w/o aug.), where aug. means the enrollment augmentation in section 3.2. Next, we also evaluated the PVAD model trained using the enrollment-less dataset without enrollment augmentation (enroll-less w/o aug.) and the PVAD model using the enrollment-full dataset with enrollment augmentation (enroll-full w/ aug.) to investigate the effect of the enrollment augmentation. PVAD (enroll-less w/o aug.) was trained using the target speaker embedding of $\boldsymbol{X}$ without SpecAugment or dropout, and PVAD (enroll-full w/ aug.) was trained by applying SpecAugment and dropout to enrollment speech $\boldsymbol{Y}$. The standard VAD trained using the enrollment-less dataset was also evaluated just for reference.

We concatenated 1 to 3 randomly selected utterances when using the enrollment-less dataset for PVAD training. We used 40 dimensional log mel-scale filterbank coefficients appended with delta and acceleration coefficients as acoustic features; they were extracted using a 20-ms-long Hamming window with a 10-ms-long shift. The input to the speaker model and PVAD was the 840 dimensional acoustic features formed by concatenating the ± 3 left-right contexts of the 120 dimensional acoustic features. The speaker model was pretrained with a speaker classification of 5994 speakers using the VoxCeleb2 dataset [29]. It was optimized using the Adam algorithm [30] with a minibatch size of 128. We set the initial learning rate of the algorithm to 0.001. The training steps were stopped if the loss on the validation set did not decrease for firve epochs in succession. We fixed the parameter of the speaker model during the training of the PVAD model. The PVAD model was optimized in the same way as that of the speaker model except that we set the initial learning rate of the algorithm to 0.0001. We randomly masked one-third of the acoustic features in SpecAugment. We set the dropout ratio p to 0.5.

We used the average precision (AP) for each class and mean

Table 1: *Evaluation results using clean test data.*

Method	AP (ns/nts)	AP (ts)	mAP
VAD	(0.926)	(0.543)	(0.691)
Conventional PVAD (enroll-full w/o aug.)	0.965	0.769	0.904
PVAD (enroll-full w/ aug.)	0.970	0.814	0.919
PVAD (enroll-less w/o aug.)	0.943	0.779	0.899
Proposed PVAD (enroll-less w/ aug.)	**0.990**	**0.937**	**0.970**

Table 2: *Evaluation results using noisy test data.*

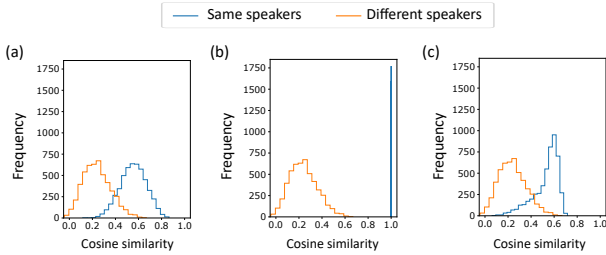| Noise	Conventional PVAD (enroll-full w/o aug.)			Proposed PVAD (enroll-less w/ aug.)		
	AP (ns/nts)	AP (ts)	mAP	AP (ns/nts)	AP (ts)	mAP
5 dB	0.932	0.696	0.840	**0.970**	**0.856**	**0.931**
10 dB	0.947	0.735	0.867	**0.979**	**0.894**	**0.946**
15 dB	0.956	0.756	0.886	**0.984**	**0.916**	**0.955**
20 dB	0.974	0.771	0.895	**0.986**	**0.928**	**0.960**

Figure 2: *(a) Cosine similarities of embeddings created using many enrollment utterances, (b) cosine similarities of embeddings without enrollment speech or enrollment augmentation, and (c) cosine similarities of embeddings created with enrollment augmentation without enrollment speech.*

AP (mAP) over all the classes to evaluate performance. We calculated the mAP using the micro-mean over all the classes as in [18].

4.3. Diversity of target speaker embedding obtained with enrollment augmentation

To examine whether enrollment augmentation is effective to create diverse target speaker embeddings while preserving the speaker characteristics, we compared the cosine similarities of speaker embeddings created using many enrollment utterances of a speaker and those created with enrollment augmentation from one utterance. We also calculated the cosine similarities between different speaker embeddings without enrollment augmentation just as reference. We used the clean test dataset.

Fig. 2 (a) shows the cosine similarities of embeddings created using many enrollment utterances, (b) shows the cosine similarities of embeddings without enrollment speech or enrollment augmentation, and (c) shows the cosine similarities of embeddings created with enrollment augmentation without enrollment speech. We can see that the cosine similarities between the speaker embeddings of the same speakers formed a distribution around 0.5 and those of different speakers formed a distribution around 0.3, as shown in Fig. 2 (a). These distributions were formed because speech characteristics varied depending on the content of the utterance, and speaker embeddings were affected by the variance [31–33]. Fig. 2 (b) shows that when no enrollment speech was available and when enrollment augmentation was not applied, speaker embedding of the same speaker did not vary, and a peak appeared at cosine similarity of 1. By applying enrollment augmentation, we can see that the variations in the

same speaker embeddings were created around cosine similarity of 0.5 without enrollment speech, as shown in Fig. 2 (c). These results demonstrated that enrollment augmentation enabled us to create diverse speaker embeddings of a speaker even when we do not have the enrollment utterances.

4.4. Detection performance

First, we compared the performance using clean test data. Table 1 shows the results. AP (ns/nts) represents the AP for non-speech or non-target speech, and AP (ts) represents the AP for target speech. We show the performance of the standard VAD just as reference and put it in parentheses. In comparing the results of the proposed PVAD (enroll-less w/ aug.) with the others, we can see that our proposed training method achieved the best performance. The results show that although using data without speaker labels could degrade the performance of PVAD when we did not use enrollment augmentation (conventional PVAD (enroll-full w/o aug.) vs. PVAD (enroll-less w/o aug.)), the data became indispensable to improve the PVAD performance when we used enrollment augmentation. We can also find that enrollment augmentation was effective even when enrollment-full dataset was used (conventional PVAD (enroll-full w/o aug.) vs. PVAD (enroll-full w/ aug.)). This is probably because enrollment augmentation further increased the variation of enrollment speech in the enrollment-full dataset.

Next, we compared the performance using noisy test data. Table 2 shows the result of the conventional method and the proposed one. We averaged the results of two types of noises. We can see that the proposed method detected the target speaker speech and avoided detecting the non-target speaker speech or noise under noisy environments.

5. Conclusions

In this paper, we proposed a new PVAD learning method that does not need enrollment speech during training. We achieved it using one utterance for both an input and an enrollment utterance. Since the PVAD model must learn speakers' speech variations to clarify the boundary between speakers, we combined SpecAugment and dropout to create diverse target speaker embeddings while preserving the speaker identity from one utterance. Our experimental results show that the PVAD model trained with the proposed method using enrollment-less datasets achieves higher performance than the one trained with the conventional method using enrollment-full datasets.

6. References

[1] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. ICASSP*, 2013, pp. 7229–7233.

[2] Y. Jung, Y. Choi, and H. Kim, "Self-adaptive soft voice activity detection using deep neural networks for robust speaker verification," in *Proc. ASRU*, 2019, pp. 365–372.

[3] F. Ringeval, S. Amiriparian, F. Eyben, K. R. Scherer, and B. W. Schuller, "Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion," in *Proc. ICMI*, 2014, pp. 473–480.

[4] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-end automatic speech recognition integrated with CTC-based voice activity detection," in *Proc. ICASSP*, 2020, pp. 6999–7003.

[5] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE SP letters*, vol. 6, no. 1, pp. 1–3, 1999.

[7] A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari, and K. Shikano, "Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs," in *Proc. Interspeech*, 2004.

[8] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE SP letters*, vol. 12, no. 10, pp. 689–692, 2005.

[9] X.-L.Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. ASLP*, vol. 21, no. 4, pp. 697–710, 2013.

[10] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. ICASSP*, 2013, pp. 7378–7382.

[11] 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," in *Proc. ICASSP*, 2013, pp. 483–487.

[12] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks." in *Proc. Interspeech*, 2013, pp. 728–731.

[13] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Proc. ICASSP*, 2014, pp. 2519–2523.

[14] R. Masumura, K. Matsui, Y. Koizumi, T. Fukutomi, T. Oba, and Y. Aono, "Context-aware neural voice activity detection using auxiliary networks for phoneme recognition, speech enhancement and acoustic scene classification," in *Proc. EUSIPCO*, 2019, pp. 1–5.

[15] R. Lin, C. Costello, C. Jankowski, and V. Mruthyunjaya, "Optimizing voice activity detection for noisy conditions," in *Proc. Interspeech*, 2019, pp. 2030–2034.

[16] J. Lee, Y. Jung, and H. Kim, "Dual attention in time and frequency domain for voice activity detection," in *Proc. Interspeech*, 2020, pp. 3670–3674.

[17] D. Gerber, S. Meier, and W. Kellermann, "Efficient target activity detection based on recurrent neural networks," in *Proc. HSCMA*, 2017, pp. 46–50.

[18] S. Ding, Q. Wang, S. Chang, L. Wan, and I. L. Moreno, "Personal VAD: Speaker-conditioned voice activity detection," *arXiv preprint arXiv:1908.04284*, 2018.

[19] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. 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.

[20] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. ASLP*, vol. 19, no. 4, pp. 788–798, 2010.

[21] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*. IEEE, 2014, pp. 4052–4056.

[22] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*, 2018, pp. 4879–4883.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[24] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *Proc. ICASSP*, 2020, pp. 6829–6833.

[27] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for VAD," in *Proc. ICASSP*, 2016, pp. 5695–5699.

[28] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[29] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[31] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Proc. Interspeech*, 2017, pp. 1497–1501.

[32] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Speaker embedding extraction with phonetic information," in *Proc. Interspeech*, 2018, pp. 2247–2251.

[33] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, and J. Cernocký, "On the usage of phonetic information for text-independent speaker embedding extraction," in *Proc. Interspeech*, 2019, pp. 1148–1152.