ADA-VAD: UNPAIRED ADVERSARIAL DOMAIN ADAPTATION FOR NOISE-ROBUST VOICE ACTIVITY DETECTION

Taesoo Kim[⋆‡], *Jiho Chang*[†] *and Jong Hwan Ko*[⋆*]

Department of Electrical and Computer Engineering, Sungkyunkwan University, Republic of Korea[⋆]
KT Corporation, Republic of Korea [‡]
Korea Research Institute of Standards and Science, Republic of Korea[†]

ABSTRACT

Voice Activity Detection (VAD) is becoming an essential front-end component in various speech processing systems. As those systems are commonly deployed in environments with diverse noise types and low signal-to-noise ratios (SNRs), an effective VAD method should perform robust detection of speech region out of noisy background signals. In this paper, we propose adversarial domain adaptive VAD (ADA-VAD), which is a deep neural network (DNN) based VAD method highly robust to audio samples with various noise types and low SNRs. The proposed method trains DNN models for a VAD task in a supervised manner. Simultaneously, to mitigate the performance degradation due to background noises, the adversarial domain adaptation method is adopted to match the domain discrepancy between noisy and clean audio stream in an unsupervised manner. The results show that ADA-VAD achieves an average of 3.6%p and 7%p higher AUC than models trained with manually extracted features on the AVA-speech dataset and a speech database synthesized with an unseen noise database, respectively.

Index Terms— Voice Activity Detection, VAD, Adversarial Domain Adaptation, Generative Adversarial Networks

1. INTRODUCTION

The purpose of voice activity detection (VAD), also known as speech activity detection, is to find speech segments in audio recordings. It has been established as an essential preprocessing stage in various applications such as auto-speech recognition and speaker verification. As those systems are commonly deployed in environments with diverse noise types and low signal-to-noise ratios (SNRs), the crucial aspect of VAD is its robustness to background noise. Recently, several DNN-based learning approaches have shown improved performance, robustness, and generality over conventional statistical methods [1, 2, 3, 4, 5]. For instance, a recent study proposed a VAD method based on the long-short term memory neural network (LSTM) [1, 2] that uses contextual information of audio. Another work proposed a boosted deep neural network (bDNN) [5] that uses multi-resolution stacking (MRS). In addition, an adaptive context attention model (ACAM) [3] has been proposed to encourage the model to focus on crucial parts of the input features. Note that all these models are trained with manually-extracted features such as multi-resolution cochlea-gram (MRCG) and mel-spectrogram. The DNN-based VAD methods generally perform well on audio steams from clean environment. However, for recordings in low-SNR environment, the performance of both approaches is drastically degraded [6]. Moreover, the performance degradation due to unseen background noises has long been a difficult task in VAD. [7, 8, 9].

One way to solve this problem is to reduce the impact of noise signals, and emphasis the clean speech signals. To address this, several studies have proposed approaches that train the DNN model using domain adaptation [8]. The most representative work is end-to-end domain-adversarial voice activity detection (DA-VAD) [9] that improved noise robustness using a gradient reversal layer. The method uses a branch of a classification network that classifies the noise type. By adding a gradient reversal layer between the feature extractor and the domain classifying branch, the feature extractor is trained to reduce the impact of the background noises in adversarial manner. However, to adopt this approach, the noise types in each audio sample of the training dataset must be labeled into a specific class in advance, which can be challenging if one audio stream includes multiple noise types.

To tackle these challenges, this paper proposes a VAD model based on adversarial domain adaptation (ADA) [10] instead of the gradient reversal layer in [9]. As a target domain to adapt, we use clean speech signals that include only speech feature with no background noise. By matching the distribution of noisy voice signals with the distribution of clean voice signals, the proposed VAD method can reduce the impact of background noise. In addition, the class of the background noises does not need to be defined in the training phase.

*This work was supported by National Research Foundation of Korea (NRF) grant (GP2022-0002-12) and Institute of Information communications Technology Planning Evaluation (IITP) grants (IITP-2021-0-00066, IITP-2019-0-00421, IITP-2020-0-01821, IITP-2021-0-02052), funded by the Korea government (MSIT).

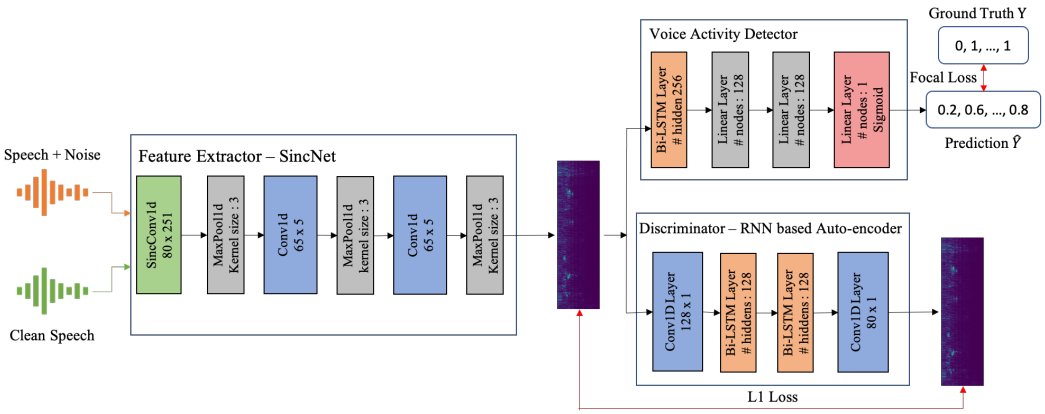

Fig. 1: The architecture of ADA-VAD. Both raw waveforms of clean speech and speech corrupted by noises are processed by the feature extractor. The voice activity detector is trained in features extracted from the feature extractor of the damaged speech waveform. On the other hand, the discriminator uses both of the clean speech waveforms and the damaged speech waveforms.

2. PROPOSED METHOD

For noise-robust VAD, it is necessary to reduce the impact of background noise and extract features appropriate for distinguishing speech from noise. To address this, the proposed method trains a DNN in a supervised manner and matches the distribution between noise and clean voice signals via ADA. The proposed method consists of three components : feature extractor, discriminator and voice activity detector. First, the purpose of the feature extractor is to extract suitable feature vectors for VAD and to mitigate domain discrepancy between clean speech signals and noisy speech signals. Therefore, it is designed to process both clean speech signals x_T^{clean} and speech signals corrupted by noisy signals x_T^{noisy}, where T indicates time. Then, the discriminator determines whether the input is from clean speech signals or noisy speech signals. It obtains the output vectors z from the feature extractor and determines whether it is z_T^{clean} or z_T^{noisy}. Meanwhile, the voice activity detector outputs the VAD predictions. It only obtains z_T^{noisy} as an input, and outputs an expected label sequence $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_T\}$ corresponding to a given audio stream, where $\hat{y}_t = 1$ if speech is detected at time t and $\hat{y}_t = 0$ if not. The proposed method accepts the raw waveform as an input and performs processing in an end-to-end manner. The overall structure of our system is illustrated in **Fig. 1**.

2.1. Adversarial Domain Adaptation

The adversarial domain adaptation technique utilizes GANs to alleviate the problem of distribution mismatch between two different datasets [10]. GANs is a generative model that generates realistic samples by training the generator G and the discriminator D adversarially. Meanwhile, Tzeng et al. [10] proposed a domain adaptation technique that maps source data x_s drawn from a one domain distribution p_s to a target distribution p_t by utilizing GANs. Based on this approach, we use a feature extractor that extracts feature vectors z_T^{clean} and z_T^{noisy} and match the distribution discrepancy between them. To stabilize training and improve the quality of the feature extractor and the discriminator, we adopt a boundary equilibrium GAN (BG) [11] approach. By training the feature extractor to deceive the discriminator D, it maps both z_T^{clean} and z_T^{noisy} on the same distribution. On the other hand, the discriminator D is trained to determine whether its input is z_T^{clean} or z_T^{noisy}. To train the discriminator perform this role, **Eq. 1** is given as the objective function. l_D indicates the L1 loss and E indicates the expectation. k_t is a constant value that balances the power of the feature extractor and the discriminator. Since the feature extractor must be trained to deceive the discriminator, the objective function **Eq 2.** is given as its objective function. The importance of loss on the feature vector z_T^{clean}, which is from clean speech signal x_T^{clean}, is controlled by the proportional control theory given by **Eq 3.**, where λ_k indicates the proportional gain for k_t. This theory encourages the discriminator to be trained by a loss with balanced ratio as a hyper-parameter $\gamma \in [0,1] = E[l_D(z_T^{noisy})]\backslash E[l_D(z_T^{clean}))]$. By applying the BG method, the feature extractor F and the discriminator D are trained with the following objectives V:

$$\max_D V_{BG}(D) = E[l_D(z_T^{clean})] - k_t * E[l_D(z_T^{noisy})] \quad (1)$$

$$\min_F V_{BG}(F) = E[l_D(z_T^{noisy})] \quad (2)$$

$$k_{t+1} = k_t + \lambda_k \gamma E[l_D(z_T^{clean})] - E[l_D(z_T^{noisy})], \quad where\ k_t \in [0,1], k_0 = 0 \quad (3)$$

As in [9], we use the raw waveform as an input, which is processed in an end-to-end manner. Also, as a feature extrac-

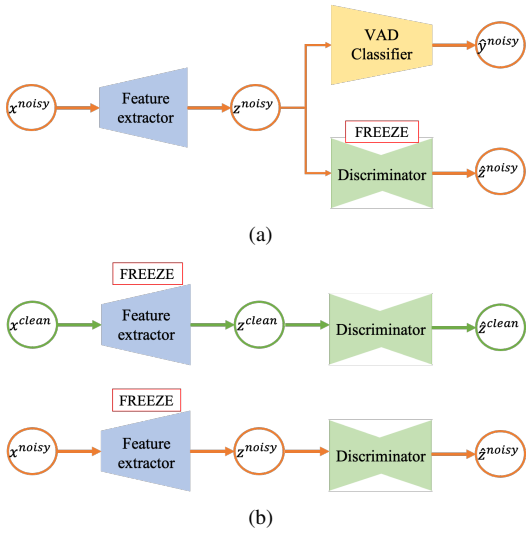

Fig. 2: (a) Procedure for updating weights of the feature extractor and the VAD classifier (b) Procedure for updating weights of the discriminator

tion module, we use SincNet [12] that showed improved performance in a speaker verification task. For the discriminators, three layers of bi-directional Long Short-Term Memory (Bi-LSTM) [13] are used, along with two 1-D Convolution layers before and after the Bi-LSTM layers.

2.2. Voice Activity Detector

The task of the voice activity detector V is to detect speech/non-speech area of a specific audio stream using the feature vector z_T^{noisy}. It is trained in a supervised manner with the corresponding ground truths $y = \{y_1, y_2, ..., y_T\}$, where $y \in [0, 1]$. Because each frame of a given audio stream is predicted $\hat{y}_1 = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_T\}$, it can be treated as multiple classification. We utilize the focal loss function (FL) [14] as an objective function to alleviate the class imbalance problem of speech/non-speech classes.

$$FL(p_t) = (1 - p_t)^{\beta} \log(p_t) \tag{4}$$

where p_t is probability of model estimation at time t. The term $(1 - p_t)^{\beta}$ is called a modulating factor with a tunable focusing parameter β that makes the model focus on hard examples while it makes the model focus less on easy samples. The feature extractor and the VAD classifier are updated using noisy speech signal as illustrated in **Fig 2(a)**. Meanwhile, both the clean speech signals and noisy speech signals are used for updating the weights of the discriminator as illustrated in **Fig 2(b)**.

3. EXPERIMENTS

3.1. Experimental setup

The TIMIT [15] corpus is designed for the development and evaluation of automatic speech recognition systems. It is used as the speech database for training and test. Among 6,300 utterances in development and evaluation sets of the TIMIT corpus, we used 4,620 utterance for training, 630 utterances for validation, and 1,000 utterances for test. In addition, 2-seconds long silence segments were added to each of the utterance to mitigate the class imbalance problem of speech and silence of the TIMIT corpus. The ground truth labels of TIMIT corpus were used. The sound effect library [16] was used for noise database for training. Among approximately 20,000 sound effects, 5,000 effects were randomly selected and used. For the training datasets, the TIMIT corpus was synthesized with the sound effect library assuming two different scenarios. The SNR level was randomly selected from -12 to 10 dB.

The datasets for training and testing consists of the following. **Train Dataset 1 (Train D_1)** In the real-world scenarios, multiple background noise types can be present in one audio stream. To incorporate this, the speech corpus was synthesized with the noise database that was randomly sorted and concatenated in a single long noise stream. **Train Dataset 2 (Train D_2)** As aforementioned, the types of background noise must be classifiable to adapt the DA-VAD [9] method. To satisfy this condition, we categorized the randomly selected noise data of the sound effect library into 18 classes.

For the test scenarios, two different datasets were used: the synthetic dataset and Atomic visual action speech (AVA-speech) [17]. **Synthetic dataset (Test D_1)** TIMIT corpus is synthesized with unseen noise database called NOISEX-92 [18]. NOISEX-92 is a database containing recordings of 15 types of noise. The all types of noise are independently selected from NOISEX-92 and synthesized with silence-added TIMIT corpus in five SNR levels: -10, -5, 0, 5, 10. **AVA-speech dataset (Test D_2)** AVA-speech dataset contains audio streams extracted from 46-hour-long videos and manually annotated ground-truth labels for VAD.

Evaluating SNR levels, the FaNT [19] noise adding tool was used for synthesizing all speech and noise databases. All of the audio databases were down-sampled to 16 kHz.

3.2. Implementation Details

For the baseline models, bDNN [5], DNN [4], LSTM [1] and DA-VAD [9] are used. All of these models are implemented followed by [3] and [20]. For DNN and bDNN, two layers with 512 nodes are used with dropout at a rate of 0.5, and the rectified linear unit (ReLU) is used as an activation function. For LSTM, two long-short term layers with 256 hidden units are used with the $arctangent$ activation function. As the hand-crafted acoustic features for the input, 80-

SNRs	DNN	bDNN	LSTM	LSTM-F	ADA-VAD
-10	67.46	70.44	73.12	83.64	**86.89**
-5	67.46	80.46	81.44	91.85	**94.36**
0	85.4	88.6	87.73	95.46	**97.01**
5	91.73	93.62	91.44	97.04	**98.01**
10	95.69	96.32	93.11	97.8	**98.48**
AVG	83.4	85.89	85.37	93.16	**94.95**

Table 1: Train D_1 as the training set. AUC(%) on the Test D_1

Model	DNN	bDNN	LSTM	LSTM-F	ADA-VAD
AUC(%)	63.46	81.64	67.29	83.79	**85.31**

Table 2: Train D_1 as the training set. AUC(%) on the Test D_2

dimensional log-mel spectrograms were adopted. The binary cross-entropy loss (BCE) is used as a loss function. Extracting log-mel spectrograms, a 25 ms Hann window with 10 ms window shift, followed by the fast Fourier transform (FFT) with 1,024 points, was adopted. DA-VAD contains a branch that classifies the type of background noise in given audio streams. Therefore, the background noise type of each audio stream in the training dataset should be defined. To handle these conditions, we used the Train D_2 dataset as a training set for the comparative experiment with DA-VAD. The DA-VAD model the implemented followed by [9]. The area under the curve (AUC) [21] is used as a metric, which is commonly used for evaluation of binary classification tasks.

The Bi-LSTM layer of the discriminator consists of three layers with 128 hidden units each, which is equal to the number of filters in the convolution layer. The voice activity detector also has three layers of Bi-LSTM, each consisting of 128 hidden units. The $arctangent$ activation function is used for all Bi-LSTM layers. Similarly, the rectified linear units (ReLU) are used as the activation functions after all the convolution layers. The τ and β are set to 0.2 and 2, respectively. The Adam gradient-based optimization method is used to update the weights of each model. The initial learning rate is set at $1e-4$ and the cyclic learning rate scheduler is adopted. The batch size is set to 1,024. All of the models are trainded and tested on an Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz with an NVIDIA TITAN RTX GPU.

4. RESULT

To investigate the effect of the proposed method, we compared the VAD performance of the LSTM models with and without the proposed method. **TABLE 1** shows the AUCs for each models that learned D_1. ADA-VAD represents the proposed method. LSTM-F represents the model composed of the same architecture with the proposed method, but without the proposed method applied. TABLE 1 shows that ADA-VAD shows $1.79\%p$ higher AUC than LSTM-F. The difference in the AUC score increases as the SNR level decreases. The result shows that the proposed method achieves higher AUC than other baseline methods that learned hand-crafted feature. Also, the adversarial domain adaptation technique helps mitigating degradation of VAD model performance in extremely low SNR scenarios. We also conducted a comparative experiment with the DNN-based models trained with the acoustic features. **TABLE 1** also shows AUC comparison of our proposed models with VAD models that learned hand-crafted feature. In particular, the performance difference between the proposed model and other DNN-based models increases as the SNR level goes down. ADA-VAD achieves 9%p higher AUC than other DNN-based models that learned hand-crafted features. **TABLE 2** shows the performance of ADA-VAD and the DNN-based VAD models on the Test D_2. ADA-VAD achieves $3.6\%p$ higher AUC than other VAD methods. It can be inferred from the result that the proposed method is robust and generalized compared to the models trained with the hand-crafted features.

Test Set	SNRs	DA-VAD	ADA-VAD
$TestD_1$	-10	85.02	**87.42**
	-5	93.53	**94.95**
	0	96.8	**97.74**
	5	98.17	**98.73**
	10	98.8	**99.6**
	AVG	94.47	**95.6**
$TestD_2$	-	71.58	**79.1**

Table 3: Train D_2 as the training set. AUC(%) of DA-VAD and ADA-VAD for each SNR levels on the Test D_1 and the Test D_2

To investigate the impact of ADA, we conducted a comparative experiment with DA-VAD [9]. **TABLE 3** shows performances of ADA-VAD and DA-VAD for each test dataset. The result indicates that ADA-VAD achieves $1.2\%p$ higher AUC than DA-VAD for the Test D_1. In addition, ADA-VAD achieves $7.5\%p$ higher AUC than DA-VAD for the Test D_2. From the result, It can be seen that even if the background noise of the input audio stream can be defined, the proposed method is more beneficial than the DA-VAD.

5. CONCLUSION

In this paper, we proposed a novel VAD method called ADA-VAD, which is trained with the adversarial domain adaptation method. In contrast to DA-VAD, the proposed method allows using a training dataset with one or more noise types per audio sample. Training with the proposed method, the feature extractor encourages the voice activity detector model to be robust to noise and low SNR scenarios by mitigating discrepancy between noisy data and clean speech data. In addition, the proposed method can be applied to extract characteristics that are better than the manually extracted acoustic characteristics for the task. Especially, the effectiveness of the proposed method is more pronounced in low SNR scenarios.

6. REFERENCES

[1] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 483–487.

[2] Juntae Kim, Jaeseok Kim, Seunghyung Lee, Jinuk Park, and Minsoo Hahn, "Vowel based voice activity detection with lstm recurrent neural network," in *Proceedings of the 8th International Conference on Signal Processing Systems*, 2016, pp. 134–137.

[3] Juntae Kim and Minsoo Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1181–1185, 2018.

[4] Xiao-Lei Zhang and Ji Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2012.

[5] Xiao-Lei Zhang and DeLiang Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Fifteenth annual conference of the international speech communication association*, 2014.

[6] Sibo Tong, Hao Gu, and Kai Yu, "A comparative study of robustness of deep learning approaches for vad," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5695–5699.

[7] Kentaro Ishizuka, Tomohiro Nakatani, Masakiyo Fujimoto, and Noboru Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, 2010.

[8] Xiao-Lei Zhang, "Unsupervised domain adaptation for deep neural network based voice activity detection," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6864–6868.

[9] Marvin Lavechin, Marie-Philippe Gill, Ruben Bousbib, Hervé Bredin, and Leibny Paola Garcia-Perera, "End-to-end domain-adversarial voice activity detection," *arXiv preprint arXiv:1910.10655*, 2019.

[10] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[11] David Berthelot, Thomas Schumm, and Luke Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[12] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[13] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[15] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at mit: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[16] Sound-ideas.com, "General series 6000 combo," 2012, [Online] Available: https://www.sound-ideas.com/Product/51/General-Series-6000-Combo.

[17] Sourish Chaudhuri, Joseph Roth, Daniel PW Ellis, Andrew Gallagher, Liat Kaver, Radhika Marvin, Caroline Pantofaru, Nathan Reale, Loretta Guarino Reid, Kevin Wilson, et al., "Ava-speech: A densely labeled dataset of speech activity in movies," *arXiv preprint arXiv:1808.00606*, 2018.

[18] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] H Guenter Hirsch, "Fant-filtering and noise adding tool," *Niederrhein University of Applied Sciences, http://dnt.-kr. hsnr. de/download. html*, 2005.

[20] Younglo Lee, Jeongki Min, David K Han, and Hanseok Ko, "Spectro-temporal attention-based voice activity detection," *IEEE Signal Processing Letters*, vol. 27, pp. 131–135, 2019.

[21] James A Hanley and Barbara J McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.