Spectro-Temporal Attention-Based Voice Activity Detection

Younglo Lee, Jeongki Min, *Student Member, IEEE*, David K. Han, and Hanseok Ko, *Senior Member, IEEE*

***Abstract*—Voice Activity Detection (VAD) systems suffer from unexpected and non-stationary background noises at magnitudes sufficiently high to mask the speech signal.Although several methods of increasing the performance of VAD have been proposed, their approaches have yet to mitigate the influence of the background noise itself. This letter proposes an effective noise-robust VAD system approach. The proposed method uses spectral attention and temporal attention through applying a deep learning-based attention mechanism. The proposed method is demonstrated and compared with several other deep learning-based methods in terms of the area under the curve in experiments with either known or unknown noise-added, and real-world noisy data. The results show that the proposed method outperforms the other methods in all the scenarios considered, but moreover generalizes well in environments of unknown or unexpected noise.**

***Index Terms*—Deep neural networks, attention mechanism, voice activity detection, speech activity detection, speech detection.**

I. Introduction

VOICE Activity Detection (VAD), also known as speech activity detection or speech detection, has undergone decades of research and development. As VAD determines presence or absence of human voice, it is an important preprocessor for speech-based audio signal processing [1]. However, one of the biggest challenges for VAD is in low Signal-to-Noise Ratio (SNR) environments. In fact, presence of non-stationary and unexpected background noises may seriously degrade accuracies of VAD algorithms. Thus, several previous efforts have been made to mitigate the influence of noise on VAD algorithms.

Prior to the recent breakthroughs brought on by the deep learning paradigm [2], early VAD algorithms considered various types of features, such as energy levels, zero-crossings, spectral or cepstral features [3], as inputs to statistical signal processing or machine learning algorithms [1]. Although these conventional methods perform reasonably well, they have limitations when handling a wide variety of signal streams and diverse range of noisy real-word environments. Therefore, the results could not be generalized.

Recognizing their unprecedented effectiveness, many researchers have proposed deep learning-based methods and have demonstrated superior performance of the resultant VAD over conventional ones. To cite a few of such examples, Deep Belief Networks (DBNs) applied to VAD outperformed the conventional Support Vector Machine (SVM) VAD [4]. Recurrent Neural Networks (RNNs) [5]–[7] were successfully applied to VAD. However, it suffers from state saturation problems when the utterance is long [8]. Various Convolutional Neural Networks (CNNs) architectures [8]–[10] were also proposed to perform VAD. Zhang, *et al.* [1] proposed a boosted DNN (bDNN) by combining Multi-Resolution Stacking (MRS) and Multi-Resolution CochleaGram (MRCG) features. Although MRS achieved outstanding performance, its ensemble-style structure proved computationally intensive. More recently, multimodal VAD, which uses both audio and visual stream, was proposed and shown to be effective in environments including high levels of noise and transients [11]. In most applications, however, visual input may not always be available for VAD and video-based method may also be sensitive to occlusion when speaker's face may be blocked from the camera view. The Adaptive Context Attention Model (ACAM), proposed by Kim [12], adopted an attention mechanism to exploit contextual information. ACAM method demonstrated improved performance over the other deep learning-based VAD methods at low computational cost.

This study proposes a noise-robust attention-based VAD architecture with the ability to generalize in unknown and unexpected noise environments. The proposed VAD method expands the idea of ACAM which attempted to solve noisy environment problem by temporal attention mechanism, e.g., sharpening the attention on the most crucial frame. However, ACAM's reinforcement loss function often tends to make the model training unstable and sensitive to hyperparameters. We address these issues by applying attention mechanism to both acoustic contextual and spectral information in an efficient manner. The proposed method is compared with other state-of-the-art deep learning-based methods by taking into consideration various situations including noise-added LibriSpeech, and real-world noisy datasets.

II. Proposed VAD Method

The proposed method is a neural networks-based VAD model for solving a binary, speech or non-speech, classification

Manuscript received October 10, 2019; revised December 2, 2019; accepted December 11, 2019. Date of publication December 16, 2019; date of current version January 24, 2020. The work of Y. Lee, J. Min, and H. Ko was supported by the Korea Environment Industry & Technology Institute (KEITI) through the Public Technology Program based on Environmental Policy, funded by the Korea Ministry of Environment (MOE) under Grant 2017000210001. The work of D. K. Han was supported by the U.S. Army Research Laboratory. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Ville M. Hautamaki. *(Corresponding author: Hanseok Ko.)*

Y. Lee, J. Min, and H. Ko are with the School of Electrical Engineering, Korea University, Seoul 02841, South Korea (e-mail: yllee@ispl.korea.ac.kr; jkmin@ispl.korea.ac.kr; hsko@korea.ac.kr).

D. K. Han is with the Information Science Division, US Army Research Laboratory, Adelphi, MD 20783 USA (e-mail: ctmkhan@gmail.com).

Digital Object Identifier 10.1109/LSP.2019.2959917

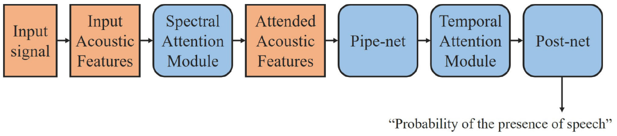

Input signal
Input Acoustic Features
Spectral Attention Module
Attended Acoustic Features
Pipe-net
Temporal Attention Module
Post-net
"Probability of the presence of speech"

Fig. 1. Overall architecture of the proposed method.

problem. After the input signal is converted into acoustic feature vectors, a set of M labeled feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m, \ldots, \mathbf{x}_M\} \in R^D$ are available for training with class labels $t_m \in \{0, 1\}$ where D is the dimension of feature vectors and m is the frame index. $t_m = 0$ or $t_m = 1$ indicates that $\mathbf{x}_m$ is a non-speech or speech frame respectively. As in [1], [12], The contextual information is incorporated by choosing the neighboring frames indexed by $\{-W, -W+u, -W+2u, \ldots, -1-u, -1, 0, 1, 1+u, \ldots, W, -2u, W-u, W\}$, which results in:

$$\mathbf{X}_m = [\mathbf{x}_{m-W}, \mathbf{x}_{m-W+u}, \ldots, \mathbf{x}_{m-1-u}, \mathbf{x}_{m-1}, \mathbf{x}_m, \mathbf{x}_{m+1}, \mathbf{x}_{m+1+u}, \ldots, \mathbf{x}_{m+W-u}, \mathbf{x}_{m+W}], \quad (1)$$

$$\mathbf{t}_m = [t_{m-W}, t_{m-W+u}, \ldots, t_{m-1-u}, t_{m-1}, t_m, t_{m+1}, t_{m+1+u}, \ldots, t_{m+W-u}, t_{m+W}]^T, \quad (2)$$

where W and u are the user-defined integer parameters described in [1], [12] and both $\mathbf{X}_m \in \mathbb{R}^{D \times L}$ and $\mathbf{t}_m \in \mathbb{R}^L$ have $L = \lfloor 2((W-1)/u) + 3 \rfloor$ neighboring frames including themselves, $\mathbf{x}_m$ and t_m. These contextual information is exploited by several attention-based modules.

The model architecture of the proposed VAD method is principally composed of four stages: 1) spectral attention, 2) pipe-net, 3) temporal attention, and 4) post-net, as depicted in Fig. 1. Each stage is described below.

A. Spectral Attention

As mentioned earlier, it is a challenge for VAD to perform well in low SNR environments. As the SNR decreases, speech patterns become heavily buried in background noise. Therefore, VAD systems must be capable of extracting the most appropriate acoustic features even in low SNR conditions. While various methods were proposed to tackle this issue [13], [14], it requires a priori SNR information obtained by noise signal estimation during non-speech periods, an approach that does not make sense given the purpose of VAD. To alleviate these difficulties, we propose a gated CNNs-based spectral attention module as depicted in Fig. 2. The gating structure has been shown to be effective in various applications [8], [15], [16]. We hypothesize that this module makes it possible to extract speech-related intrinsic features by focusing the model's attention only on relevant spectral parts in an acoustic feature space. This module acts like a preprocessor for VAD, producing several mask matrices with each element of which outputs a number between zero and one. These mask matrices are directly multiplied pointwise by another spectral feature map, which indicates how much of each spectral component should be attended by VAD.

The spectral attention module consists of T blocks with each block composed of a pair of convolutional layers and

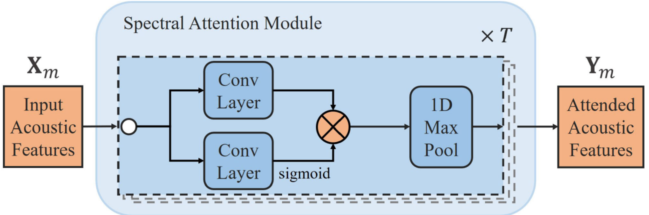


Fig. 2. Spectral attention module.

one-dimensional max pooling layer. First, the input acoustic features $\mathbf{X}_m$ are first convolved with a pair of N_c $f_c \times f_c$ convolutional filters. One of the output feature maps is then fed into a sigmoid function, which acts like a gate as it is multiplied to another convolved feature map. In order to make this module deformation invariant while maintaining temporal information, one-dimensional max pooling is applied along the frequency axis. The number of filter N_c is doubled after each block, which is repeated T times. For example, the first two block output shapes are $\frac{D}{2} \times L \times N_c$ and $\frac{D}{4} \times L \times 2N_c$ repectively. Note that the output of the spectral attention module, namely attended acoustic features $\mathbf{Y}_m$, is reshaped by merging frequency and channel dimension.

B. Pipe-Net

The pipe-net contains two fully connected layers, each with N_p units, which acts as an information bridge between the spectral attention module and the temporal attention module [17]. Note that the weights are shared across all the frames. In order to prevent the vanishing gradient problem and aid convergence of the spectral attention module, another hidden layer with a single unit including a sigmoid activation $\mathbf{z}_m \in \mathbb{R}^L$, and cross-entropy loss with target $\mathbf{t}_m$ are added after the pipe-net as follows:

$$L_{pipe} = -\sum_{m=1}^{M} \sum_{l=1}^{L} \left(t_m^l \log z_m^l + (1 - t_m^l) \log(1 - z_m^l) \right), \quad (3)$$

where t_m^l and z_m^l are the l^{th} component of $\mathbf{t}_m$ and $\mathbf{z}_m$ respectively. Note that this loss is for training only.

C. Temporal Attention

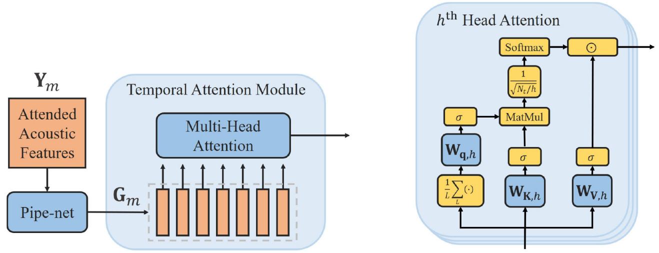
The temporal attention module, which originates from attention models [18], [19], allows the VAD to attend to the most important positions from several neighboring input features. As the expected number of ones of the target $\mathbf{t}_m$ is more than one, e.g., $[0, 1, 1, 0, 1, 0, 0]^T$, multi-head self-attention is adopted, allowing the model to simultaneously attend to information at different positions [20]. As far as we know, this is the first attempt to use multi-head attention mechanism for VAD. Fig. 3 depicts the temporal attention module, which is described below.

1) Attention: Let $\mathbf{G}_m \in \mathbb{R}^{N_p \times L}$ be the pipe-net output. First, query $\mathbf{q}_m$, key $\mathbf{K}_m$ and value $\mathbf{V}_m$ are calculated using $\mathbf{G}_m$:

$$\mathbf{q}_m = \sigma(\mathbf{W}_\mathbf{q} \mathbf{g}_m) \in \mathbb{R}^{N_t}, \quad (4)$$

$$\mathbf{K}_m = \sigma(\mathbf{W}_\mathbf{K} \mathbf{G}_m) \in \mathbb{R}^{N_t \times L}, \quad (5)$$

$$\mathbf{V}_m = \sigma(\mathbf{W}_\mathbf{q} \mathbf{G}_m) \in \mathbb{R}^{N_t \times L}, \quad (6)$$


$\mathbf{Y}_m$
Attended Acoustic Features
Pipe-net
$\mathbf{G}_m$
Temporal Attention Module
Multi-Head Attention
h^{th} Head Attention
Softmax
$\odot$
$\frac{1}{\sqrt{N_t/h}}$
σ
MatMul
$\mathbf{W}_{\mathbf{q},h}$
σ
σ
$\frac{1}{L}\sum_{L}(\cdot)$
$\mathbf{W}_{\mathbf{K},h}$
$\mathbf{W}_{\mathbf{V},h}$

(a) (b)

Fig. 3. (a) Temporal attention module, (b) h^{th} head attention function.

where N_t denotes attention dimension, σ is an activation function, and $\mathbf{g}_m \in \mathbb{R}^{N_p}$ is obtained by averaging $\mathbf{G}_m$ along the frame axis. $\mathbf{W}_* \in \mathbb{R}^{N_t \times N_p}$ are the network parameters to be trained.

The attention function calculates the attention vector for multiple frames. It can be denoted as:

$$\text{Attention}(\mathbf{q}_m, \mathbf{K}_m, \mathbf{V}_m) = \text{softmax}\left(\frac{\mathbf{q}_m^T \mathbf{K}_m}{\sqrt{N_t}}\right) \odot \mathbf{V}_m \quad (7)$$

where $\sqrt{N_t}$ is the scaling factor to prevent the magnitude of attention vector from growing too large.

2) Multi-Head Attention: In order for the model to attend to multiple frames instead of single frame, multi-head attention network can be achieved by simply modifying the attention network as follows:

$$\text{Multihead}(\mathbf{q}_m, \mathbf{K}_m, \mathbf{V}_m) = \text{concat}(\text{head}_1, \ldots, \text{head}_H), \quad (8)$$

$$\text{head}_h = \text{Attention}(\mathbf{q}_{m,h}, \mathbf{K}_{m,h}, \mathbf{V}_{m,h}), \quad (9)$$

where H is the number of heads, and $\mathbf{q}_{m,h}, \mathbf{K}_{m,h}, \mathbf{V}_{m,h}$ are the hth slice of $\mathbf{q}_m, \mathbf{K}_m, \mathbf{V}_m$ repectively. For example, $\mathbf{q}_m$, being composed of H vectors, can be written as $\mathbf{q}_m = [\mathbf{q}_{m,1}; \cdots; \mathbf{q}_{m,H} \in \mathbb{R}^{N_t/H}]$.

After the multi-head attention output, which can be denoted as $\text{Multihead}(\mathbf{q}_{m,h}, \mathbf{K}_{m,h}, \mathbf{V}_{m,h})$, is acquired, then it is passed through the post-net which is the final module. As the contextual information for the target is already known, the additional cross-entropy loss between the target $\mathbf{t}_m$ and the multi-head attention vector is added to help the temporal attention focus on the appropriate frames:

$$\text{softmax}\left(\frac{\mathbf{q}_m^T \mathbf{K}_m}{\sqrt{N_t}}\right) = \left[\alpha_m^1, \ldots, \alpha_m^L\right], \quad (10)$$

$$L_{att} = -\sum_{m=1}^{M}\sum_{l=1}^{L}\left(t_m^l \text{log}\alpha_m^l + (1 - t_m^l)\text{log}(1 - \alpha_m^l)\right). \quad (11)$$

D. Post-Net

The post-net is comprised of 2 fully connected layers, each with N_p hidden units except the final layer to produce a single scalar value. The scalar value is passed through a sigmoid activation to predict the probability $\hat{\mathbf{t}}_m$ of the presence of speech, which can be expressed as:

$$\hat{\mathbf{t}}_m = f_{\text{post}}(\text{Multihead}(\mathbf{q}_m, \mathbf{K}_m, \mathbf{V}_m) \mid \theta_{\text{post}}) \in \mathbb{R}^L, \quad (12)$$

and the final prediction corresponding to the m^{th} frame $\mathbf{x}_m$ can be obtained by combining all the predictions relative to frame m across l as in [12]. θ_{post} is the postnet parameters to be trained. Finally, post-net loss function is

$$L_{post} = -\sum_{m=1}^{M}\sum_{l=1}^{L}\left(t_m^l \text{log}\hat{t}_m^l + (1 - t_m^l)\text{log}(1 - \hat{t}_m^l)\right), \quad (13)$$

and the final combined loss function of our model is

$$L = L_{post} + L_{pipe} + \lambda L_{att}, \quad (14)$$

where λ is the mixed weight, which is used to regulate the impact of attention information.

III. Experiments

A. Dataset

The TIMIT training corpus is used in the training phase [21]. In order to mitigate the speech/non-speech class imbalance problem of TIMIT utterances, 1 second long silence is added before and after each utterance as in [12]. In addition, the TIMIT training corpus is augmented with eight types of additive noise of NOISEX-92 [22] with SNR values set at: −10, −5, 0, 5, 10 dB. This augmentation resulted in 189,420 training segments and added up to about 267 hours of audio stream in total. The model was trained with 95% of the training data and the remaining 5% was left as a validation set.

For test phase, the LibriSpeech [23] development and test corpus were employed. As there is no ground-truth for LibriSpeech dataset, we created pseudo ground-truth targets with the help of rVAD toolkit [24]. AURORA dataset [25] with eight types of noise was added to the test utterances with SNR levels at −10, −5, and 0 dB. In addition, TED-LIUM 3 [26], real-world noisy speech dataset provided by [12] and our in-vehicle noisy speech dataset were utilized. The test dataset can be categorized into four groups as follows:

1) Test 1: LibriSpeech dev/test + AURORA (≈510 hours)
Unseen speech utterances with unseen background noise.

2) Test 2: TED-LIUM 3 (≈452 hours)
TED talk speech utterances with transient applause, laughter, music, reverberation and general noise.

3) Test 3: [12] (≈2 hours)
Real-world recorded noisy speech utterances with various types of non-stationary and unexpected background noise.

4) Test 4: Ours (≈1 hour)
Real-world recorded noisy speech utterances spoken to car navigation with stationary in-vehicle background noise.

B. Acoustic Feature

First, entire audio segments were resampled to get the fixed sampling rate at 16 kHz and framed by applying a 25 ms Hann window with 10 ms window shifts, followed by a Fast Fourier Transform (FFT) with 1024 points. Finally, 80 Mel

TABLE I
AUC (%) FOR TEST 1, 2, 3 AND 4

Dataset (# param.)	DNN (552K)	LSTM (1.40M)	bDNN (556K)	ACAM (639K)	Ours-1 (426K)	Ours-2 (559K)
Test 1	85.57	82.38	87.04	84.44	88.95	**89.67**
Test 2	71.35	71.60	71.29	69.73	71.00	**72.04**
Test 3	96.49	95.43	96.89	95.94	96.92	**96.96**
Test 4	93.02	92.50	96.10	92.18	96.17	**96.20**

For Test 1, AUCs were averaged over both the SNRs (dB) $\{-10, -5, 0\}$ and the noise types due to space limitations.

TABLE II
SOME DETAILED AUC (%) RESULTS FOR TEST 1

Noise	SNR	DNN	LSTM	bDNN	ACAM	Ours-1	Ours-2
Airport	-10 dB	82.36	78.66	82.10	80.11	86.20	**87.25**
	-5 dB	84.89	81.66	86.68	84.69	89.48	**90.15**
	0 dB	87.75	86.12	91.05	89.53	92.17	**92.56**
Babble	-10 dB	84.12	79.41	83.43	78.60	86.11	**87.28**
	-5 dB	85.91	81.73	86.92	82.85	88.83	**89.69**
	0 dB	88.22	85.45	90.69	88.03	91.63	**92.14**
Car	-10 dB	83.91	79.70	84.60	83.03	86.79	**87.50**
	-5 dB	85.99	82.46	88.25	87.38	89.65	**89.99**
	0 dB	87.96	85.80	91.42	91.17	92.00	**92.20**
Exhibition	-10 dB	82.95	79.50	83.20	79.94	85.09	**85.87**
	-5 dB	85.10	81.13	86.66	83.78	87.96	**88.47**
	0 dB	87.51	84.07	90.43	88.39	90.75	**91.06**
Restaurant	-10 dB	80.75	77.05	79.66	75.80	83.48	**85.35**
	-5 dB	82.66	79.15	83.51	79.68	86.56	**88.02**
	0 dB	86.34	83.92	88.78	85.68	90.17	**91.03**
Street	-10 dB	82.65	78.65	80.71	76.14	85.07	**86.31**
	-5 dB	85.91	83.20	86.80	82.42	89.45	**90.25**
	0 dB	88.77	87.77	91.70	88.87	92.50	**92.97**

filter bank coefficients ($D = 80$) were obtained for each frame and then logarithm operation was applied. The extracted 80-dimensional log-mel spectrograms were adopted as the acoustic feature vectors $\mathbf{x}_m$ of our model to summarize the frequency content. It is also shown that high quality original raw audio stream can be reasonably regenerated from the corresponding log-melspectrograms [17], which implies that such acoustic features are sufficient to represent the original speech. Each log-mel spectrogram was normalized by its minimum and maximum value to have a value between 0 and 1.

C. Model Details

The parameters for the contextual information $\mathbf{X}_m$ and $\mathbf{t}_m$ were set to be $W = 19$ and $u = 9$, which provides three neighboring frames before and after the current frame $\mathbf{X}_m$ and $\mathbf{t}_m$ ($L = 7$). The parameters for our model T, N_c, f_c, N_p, N_t, H and λ are empirically set to be 4, 16, 3, 256, 128, 4 and 0.1. A Rectified Linear Unit (ReLU) is used for nonlinear activation function. In order to prevent our model from overfitting, batch normalization was applied to all the layers before the activation [27] and dropout at a rate of 0.5 was only used in all the fully connected layers after the activation, except the temporal attention module [28]. The model was optimized with the Adam optimizer with a learning rate decay starting at 0.001 [29] and the early stopping strategy based on validation loss was used after a minimum of ten epochs. The batch size was set to be 512.

For performance comparison, four deep learning-based approaches were used: DNN, bDNN, LSTM and ACAM [30]. We followed the model architecture and parameters as described in [12], e.g., two layers with 512 nodes for DNN and bDNN, three layers with 256 units for LSTM. All the other parameters were the same as the above. The total number of parameters for each model is shown in Table I. The Area Under the Curve (AUC) was used to characterize VAD performance.

D. Experimental Results

In order to explore the influence of attention modules, pipe-net output, denoted by Ours-1, was also evaluated in addition to post-net output, denoted by Ours-2. Note that Ours-1 exploits spectral attention only while Ours-2 takes both spectral and temporal attention into consideration. The results are summarized in Table I. For all the test scenarios considered, the proposed VAD outperformed all the baseline methods. It was also observed that using spectral and temporal attention jointly, Ours-2, showed monotonic increase in AUC performance comparing with only-spectral attention method, Ours-1, while Ours-1 achieved near the best performance results among the baseline methods. These results imply that several attention modules help the model focus on the appropriate parts for the performance improvement. ACAM-based method, which depends entirely on temporal attention, showed the worst performance possibly due to its brittleness in unknown noise conditions. Table II shows some detailed results for Test 1 according to noise types and different SNR values. As Test 1 included speech utterances and noise types unseen during the training phase, these results showed that proposed VAD demonstrated its ability to generalize in unknown noisy environments robustly in extremely low SNR conditions. We found that each attention module can effectively process both spectral and temporal information for robust VAD performance.

IV. CONCLUSIONS

In this letter, a noise-robust attention-based VAD architecture was proposed and validated. Our proposed approach tackles the issue of generalization by simultaneously applying spectral and temporal attention modules. The several attention modules can learn to attend the appropriate speech-related parts from spectral and temporal information jointly. Although the samples in NOISEX-92 do not have wide variety of noise types, the proposed method performed well in unknown real-world noise situations and demonstrated its ability to generalize robustly in extermely low SNR conditions. The proposed architecture can be applied to a variety of audio classification frameworks such as speaker recognition and acoustic scence classification, which is left for future works.

REFERENCES

[1] X.-L. Zhang and D. Wang, “Boosting contextual information for deep neural network based voice activity detection,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 2, pp. 252–264, Feb. 2016.

[2] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[3] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.

[4] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, Nov. 2012.

[5] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7378–7382.

[6] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 483–487.

[7] G. Gelly and J.-L. Gauvain, "Optimization of RNN-based speech activity detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 646–656, Nov. 2017.

[8] S.-Y. Chang *et al.*, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5549–5553.

[9] D. A. Silva, J. A. Stuchi, R. P. V. Violato, and L. G. D. Cuozzo, "Exploring convolutional neural networks for voice activity detection," in *Cognitive Technologies*. Cham, Switzerland: Springer, 2017, pp. 37–47.

[10] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.

[11] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using wavenet encoder and residual networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 265–274, Feb. 2019.

[12] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018.

[13] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7229–7233.

[14] A. Sholokhov, M. Sahidullah, and T. Kinnunen, "Semi-supervised speech activity detection with an application to automatic speaker verification," *Comput. Speech Lang.*, vol. 47, pp. 132–156, Jan. 2018.

[15] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[16] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 933–941.

[17] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4779–4783.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Repres.*, May 2015.

[19] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[20] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon, USA, Tech. Rep. NISTIR 4930, vol. 93, 1993.

[22] A. Varga and H. J. 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 Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[24] Z.-H. Tan, A. k. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Comput. Speech Lang.*, vol. 59, pp. 1–21, Jan. 2020.

[25] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Autom. Speech Recognit.: Challenges Millenium ISCA Tut. Res. Workshop*, 2000, pp. 181–188.

[26] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Proc. Int. Conf. Speech Comput.*, 2018, pp. 198–208.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Repres.*, May 2015, pp. 1–15.

[30] J. Kim, "VAD-toolkit," GitHub Repository, 2017. [Online] Available: https://github.com/jtkim-kaist/VAD