



# Voice Activity Detection for Live Speech of Baseball Game Based on Tandem Connection with Speech/Noise Separation Model

Yuto Nonaka<sup>1</sup>, Chee Siang Leow<sup>1</sup>, Akio Kobayashi<sup>2</sup>, Takehito Utsuro<sup>3</sup>, Hiromitsu Nishizaki<sup>1</sup>

<sup>1</sup>Graduate School of Medicine, Engineering, and Agricultural Sciences,  
University of Yamanashi, Japan

<sup>2</sup>Faculty of Industrial Technology, Tsukuba University of Technology, Japan

<sup>3</sup>Faculty of Engineering, Information and Systems, University of Tsukuba, Japan

{nonaka0422, cheesiang\_leow}@alps-lab.org, a-kobayashi@a.tsukuba-tech.ac.jp,  
utsuro@iit.tsukuba.ac.jp, hnishi@yamanashi.ac.jp

## Abstract

When applying voice activity detection (VAD) to a noisy sound, in general, noise reduction (speech separation) and VAD are performed separately. In this case, the noise reduction may suppress the speech, and the VAD may not work well for the speech after the noise reduction. This study proposes a VAD model through the tandem connection of neural network-based noise separation and a VAD model. By training the two models simultaneously, the noise separation model is expected to be trained to consider the VAD results, and thus effective noise separation can be achieved. Moreover, the improved speech/noise separation model will improve the accuracy of the VAD model. In this research, we deal with real-live speeches from baseball games, which have a very poor signal-to-noise ratio. The VAD experiments showed that the VAD performance at the frame level achieved 4.2 points improvement in F1-score by tandemly connecting the speech/noise separation model and the VAD model.

**Index Terms:** multi-task learning, speech/noise separation, tandem connection, voice activity detection

## 1. Introduction

Voice activity detection (VAD) [1–3] is a technique for improving speech processing efficiency and accuracy. VAD is a detection technique that identifies human speech and non-speech segments in audio data. Using a VAD technique to detect speech segments, it is no longer necessary to perform automatic speech recognition (ASR) for segments that do not include voice. As a result, this technique is expected to improve speech recognition accuracy because ASR for a non-speech segment is unnecessary. However, for speech in a noisy environment, VAD becomes difficult because the difference between the volume of the speech segment and the volume of the other noisy segments is very small.

In this paper, we aim to improve the accuracy of VAD for speech containing loud noises by simultaneously using VAD and speech/noise separation techniques. The target speech data dealt with in this study come from the live broadcast sounds of baseball games. Various kinds of loud noises are included in the sounds of a baseball game. In particular, like soccer games, Japanese baseball games are characterized by cheering using musical instruments, etc. At the moment of excitement, the signal-to-noise (S/N) ratio between the live speech and the other noises becomes almost zero. Although speech processing research for clean speech has reached a certain point, there are still many problems to be solved for speech containing loud noises, such as live sports speech. One of them is VAD.

There are many studies on VAD for speech data [2–6]. Recently, VAD research using deep learning techniques has become popular. For example, Sainath et al. [2] proposed a convolutional long short-term memory deep neural network (CLDNN)-based VAD, which accepts raw waveform. In addition,

Bai et al. [3] also studied a DNN-based VAD with a Viterbi algorithm.

To improve the performance of VAD with noisy speech, it is effective to perform denoising before VAD. First, the speech-containing noise is separated into speech and noise, and VAD is then applied to the separated speech to detect the speech segment. However, in this case, the speech part is also suppressed by noise removal, which may adversely affect the subsequent VAD processing. Therefore, in this study, we propose a simultaneous training approach with speech/noise separation and VAD models using a tandem connection. Using the proposed method, the separation model is trained while taking into account the results of VAD, which allows the model to perform effective denoising against VAD. In addition, by improving the performance of the noise reduction, the accuracy of the VAD is expected to be improved. Furthermore, since there is no model training corpus for VADs in live speech from baseball games, we adopted a semi-supervised approach to dynamically generate training data from an existing speech corpus and baseball game sounds. We also show that the tandem model trained from sounds of simulated noise environments performs well on real-world data.

There is a similar study [7] to our proposed approach. Lee et al. proposed to use a U-Net model [8] to estimate speech segments from noisy speech. This U-Net model directly estimated the spectrogram of the noise and the spectrogram of the clean speech from the synthesized speech with noise through multi-task learning. At the same time, it estimated an ideal ratio mask (IRM) [9] or an ideal binary mask (IBM) [10] of speech and noise and calculated the estimated speech segment to achieve VAD. On the other hand, our research is an end-to-end approach that estimates the noise suppression mask from speech with noise and uses a neural network for VAD. In addition, Mirsamadi et al. [11, 12] have proposed a model for estimating the noise mask from the outputs from the DNN-based VAD and the clean speech estimator; however, they did not optimize multiple models simultaneously. Furthermore, Wang et al. [13] showed that estimating speech presence probabilities from outputs of a multi-task model was effective for speech enhancement tasks. However, in this study, they did not use a neural network like VAD to estimate the speech segment. On the other hand, there are also many studies on speech separation, singing voice separation, and noise reduction from speech [14]. For example, Hermans et al. [15] improved speech/singing voice/noise separation performance by jointly optimizing time-frequency masks and recurrent neural networks (RNNs) [16]. Our paper uses a U-Net architecture, which consists of convolutional layers without using RNNs, separates speech and noise, and jointly optimizes VAD to improve noise reduction performance. Additionally, there are several studies [17, 18] on noise reduction using U-Net. However, no study has yet proposed a model that simultaneously optimizes speech/noise separation and VAD models.

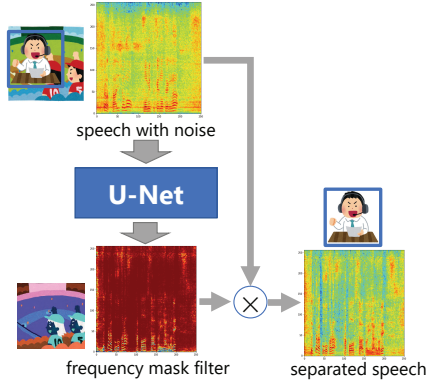


Figure 1: Noise suppression flow using the U-Net model

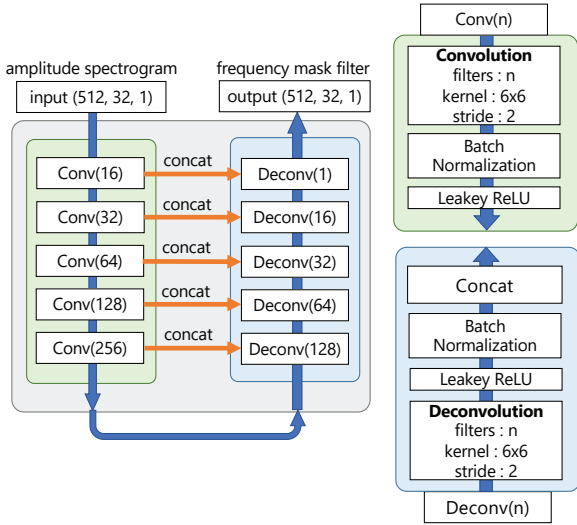


Figure 2: The architecture of the U-Net-based speech/noise separation model

The VAD experiment using live speech from baseball games confirmed that the performance of VAD for noisy speech was improved by applying noise reduction before VAD. Furthermore, the proposed tandem model, composed of U-Net-based speech/noise separation and convolutional neural network (CNN)-based VAD models, further improved the accuracy of VAD by optimizing the two models simultaneously. Based on the above, the contributions of this paper are as follows: first, the accuracy of VAD for noisy speech can be improved by optimizing the two models simultaneously, and second, models trained on the simulated live speech of baseball games can perform well with real-world data.

## 2. Models

### 2.1. Speech/Noise Separation Model

We use a U-Net model as the speech/noise separation model. This model is known to be suitable for source separation [19]. This model can train a frequency mask filter to suppress only the noise for a speech spectrogram containing noise. As shown in Figure 1, when a speech spectrogram containing noise is input into the trained U-Net model, a frequency mask filter can be dynamically generated by the model. By applying the mask filter to the input speech spectrogram, noise-separated speech can be obtained.

The data required for training the speech/noise separation model are clean speech and speech with noise added to it. Since

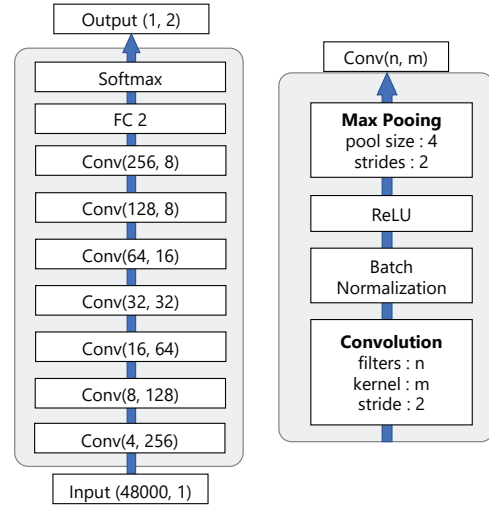


Figure 3: The architecture of the CNN-based VAD model

such a corpus does not exist, the model is trained by dynamically synthesizing clean speech with task-specific noise data. The speech length of the amplitude spectrogram input to the model is one second (16,000 samples). The speech is cut out by a Hamming window of 1,024 points, sliding at 512 points, and converted to an amplitude spectrogram using FFT. Therefore, the amplitude spectrogram of 32 frames from one second of speech, i.e., the (512, 32) shape of a two-dimensional tensor, is input to the U-Net model. This is to match the input conditions of the VAD model. Therefore, the masking process for noise suppression is performed in one-second steps.

Figure 2 shows the structure of the U-Net model used in this paper. The model consists of five convolutional layers and five inverse convolutional layers. The convolutional layer performs two-dimensional convolutional operations with a filter size of  $6 \times 6$  and a stride size of two while also applying batch normalization [20]. The activation function is the leaky rectified linear unit (Leaky ReLU) function [21]. When the input is less than or equal to zero, the slope is set to 0.2. In the inverse convolutional layer, the inverse convolution operation is also performed with a filter size of  $6 \times 6$  and stride size of two, and batch normalization is also applied in the same way. For the activation function in the inverse convolutional layers, the ReLU function is used from the first to the fourth layer, and the sigmoid function is used in the fifth layer to obtain filter coefficients in the range of 0 to 1. Loss is calculated based on the mean absolute error function, considering the difference between the noise-masked speech and the input clean speech.

### 2.2. VAD model

A CNN is also used to train the VAD model. The model structure consists of a one-dimensional convolutional layer and a fully connected layer, as shown in Figure 3, based on previous research on a real-time VAD model in noisy environments [22, 23]. The input for this model is a raw waveform with a 48 kHz sampling rate—a fixed one-second speech sample. The model has two sorts of output labels: speech segment and non-speech segment. The model decides whether the one second of input data is a speech segment or not frame-by-frame. In this VAD model, the sampling frequency of the input audio is assumed to be 48 kHz, and the model structure is optimized for this. Since we deal with 16 kHz audio in this study, the audio input to the VAD model is up-sampled from 16 kHz to 48 kHz.

The network structure of the VAD model used consists of seven one-dimensional convolutional layers and one fully connected layer. In the convolutional layer, the convolution oper-

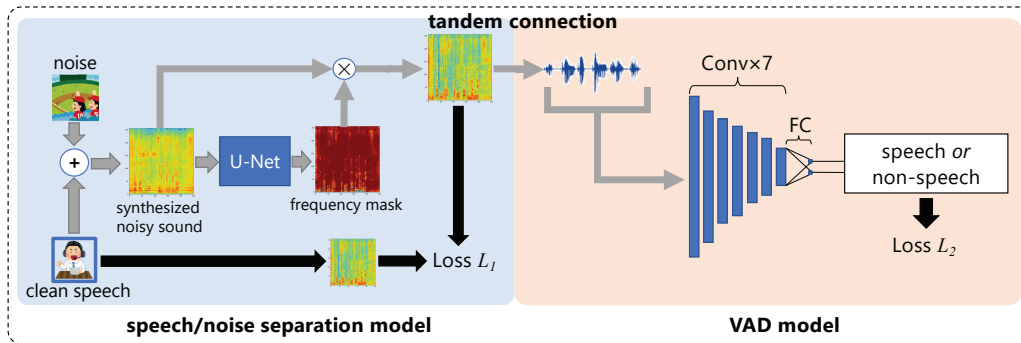


Figure 4: Tandem connection of the speech/noise separation model and the VAD model

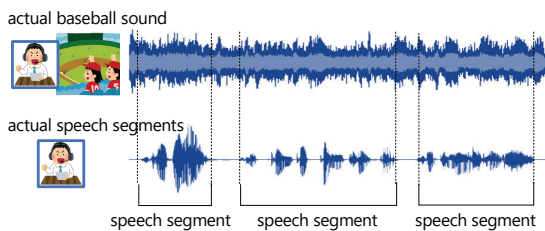


Figure 5: An example of a sound waveform from baseball game audio

ation is performed with a filter size of  $6 \times 6$  and a stride size of two, and batch normalization is performed. The activation function uses the ReLU function, and max pooling is applied with a window size of four and a stride size of two. The number of output nodes in the fully connected layers is two, the activation function uses the softmax function, and the loss value is calculated using cross-entropy.

### 2.3. Tandem Connection

As shown in Figure 4, the speech/noise separation model and the VAD model described in Section 2.1 and 2.2 are connected in tandem and trained simultaneously. A raw waveform is input to the VAD model, while the output of the separation model is the spectrogram form. Therefore, the separated speech spectrogram by the separation model is transformed into a raw waveform by inverse fast Fourier transform (IFFT), and it then proceeds into the VAD neural network. The phase information used in the IFFT is calculated based on the input speech for the separation model.

In the tandem connection model, two sorts of loss values are computed: mean absolute error loss ( $L_1$ ) from the separation model and cross-entropy loss ( $L_2$ ) from the VAD model. For updating the parameters of the VAD network, only  $L_1$  loss is used. On the other hand, the combined loss of  $L_1$  and  $L_2$  is used to update the parameters of the separation model. That is:

$$L_1 + 0.02 \times L_2.$$

Combining the loss values from the separation model with the VAD model allows the separation network to be well-trained while taking into account the presence of speech. Therefore, the performance of noise suppression will be improved, and the performance of VAD will also be improved.

## 3. Dataset of Live Sound of Baseball Games

We evaluate the proposed VAD model on live speech from baseball games. The VAD of live sports audio, such as that from baseball games, is challenging because it contains not only the

voice of the announcer and commentator but also many loud noises, such as cheers of the audience and celebratory music. Figure 5 shows a sound waveform from baseball game audio. As shown in Figure 5, it is difficult to estimate the speech segment just by looking at the waveform.

In the speech separation task with live baseball game audio, the training data required for the separation model are so many kinds of sounds from baseball games and the actual live speech contained in the audio. However, it is difficult to collect a large amount of these data in practice. In this study, we collected an existing clean speech corpus and the noises in baseball stadiums and synthesized them to create pseudo-baseball game audio. For the existing clean speech corpus, we obtained 202 persons' utterances (50 hours) randomly selected from the corpus of spontaneous Japanese [24] (CSJ, totaling 600 hours of speech), which is commonly used in Japanese speech recognition research. Since each speech file in the CSJ is labeled with information about the speech segment, it is possible to automatically generate labels for speech and non-speech segments for the VAD model. In addition to this, we collected 29 videos, such as live TV game videos and talk videos of excited speakers, and we extracted only the speech segments from these videos as clean speech. The purpose of this is to reproduce the announcer's voice when the baseball game gets exciting. Finally, we collected 50 hours of clean speech and 15 hours of baseball noise audio. On the other hand, 67 videos of Japanese professional baseball games and high school baseball games were collected from the video distribution site YouTube to be used as noise data. In addition, the test set audio for the proposed VAD consisted of the audio of three baseball games collected from "radiko" [25], which is a service that distributes Japanese radio broadcasts over the Internet. This is a completely different set of live baseball game audio from the noise data used when the model is trained.

## 4. Experiment

We compare three sorts of VAD approaches to investigate the effectiveness of the proposed approach of simultaneously training the speech/noise separation and the VAD models as follows:

**VAD only** : Train the VAD model from speech with noise (pseudo-synthesized audio). The separation model is not used.

**Separate** : Train the separation model and the VAD model separately, and then apply the VAD model after speech separation.

**Tandem (proposed)** : Train the tandem connection model consisting of the separation and the VAD.

For each model, the input speech (or its amplitude spectrogram) duration is 1 second and shifted by 0.5 seconds. The VAD per-

formance is measured frame-by-frame (one frame, one second) using the F1-score, which is the harmonic mean of the recall rate and the precision rate of each speech segment class and non-speech segment class.

#### 4.1. Experimental Setup

When training the noise reduction and VAD models separately, the mini-batch size was 256 for both the separation model and the VAD model. A one-second amplitude spectrum of (32, 512) shapes was input into the separation model. On the other hand, one-second audio with a sampling frequency of 48 kHz was also input into the VAD model. Since the sampling frequency of clean speech and baseball noise was 16 kHz, sounds were up-sampled to 48 kHz to match the VAD model. Adam [26] was used as the optimization function for both models, and the learning rate was set to 1e-03 and 1e-04 for the separation and VAD models, respectively.

In the tandem model, the mini-batch size was set to 256. Separate optimization functions (Adam) were prepared for the two models in the tandem model: the learning rates of the VAD network and the U-Net were set to 1e-05 and 1e-07, respectively. The combination of the loss values from the two models has been described in Section 2.3.

The training and evaluation data have also been described in Section 3. Ten (about 4 hours) of the 50 hours of clean speech were used to validate the models.

#### 4.2. Results

Table 1 shows the VAD results for the test audio from baseball games. Table 1 (a) indicates the VAD results when only the VAD model was trained, Table 1 (b) shows the results when the two models were trained separately, and Table 1 (c) shows the results when the two models were tandemly connected and trained simultaneously. These tables show the number of segments (frames) estimated in the speech and non-speech segments, the F1-score for each class, and their macro-averages.

First, comparing Table 1 (a) and (b), the F1-score (macro-average) improved from 84.3% to 85.9% by applying VAD after the noise separation rather than the VAD model alone. In particular, the F1-score for non-speech segment detection was greatly improved, indicating that the separation model could suppress the false detection of speech. However, there was no significant improvement in speech segment detection. Figure 6 (a) shows a spectrogram of input noisy speech, which was noise-suppressed by the separation model. Looking at the 800 Hz to 1,000 Hz frequency band from 1.8 to 4.2 seconds, not only the noise but also the speech was suppressed. Thus, it is possible that the separation model also suppresses speech.

On the other hand, the proposed method (the tandem connection model) improved the F1-score (macro-average) by 2.6 points compared to training the two models separately. Moreover, the F1-score of the non-speech segment increased by 3.7 points compared to the results of training the models separately and by 6.4 points compared to the results of training the VAD model alone. Furthermore, the F1-score of the speech segment also increased by 3.5 points compared to when the models were trained separately. Figure 6 (b) also shows the noise suppression results for the tandem connection model (same audio as in Figure 6 (a)). When the noise suppression models were trained separately, the speech was also suppressed; however, in the tandem model, the suppressed speech was retained.

In summary, the experimental results confirmed that the performance of the VAD for the noisy live speech of baseball games was improved by simultaneously optimizing the speech/noise separation model and the VAD model. Furthermore, although the proposed model was trained on simulated data, the model worked well for actual baseball game sounds.

Table 1: The number of speech and non-speech segments estimated by three VAD approaches and F1-scores of each class

(a) "VAD only"			
actual label	estimated class		F1-score [%]
	non-speech	speech	
non-speech	1,072	264	74.8
speech	458	5,403	93.7
macro ave.	—	—	84.3
(b) "separate"			
actual label	estimated class		F1-score [%]
	non-speech	speech	
non-speech	1,172	160	77.7
speech	514	5,339	94.1
macro ave.	—	—	85.9
(c) "tandem (proposed)"			
actual label	estimated class		F1-score [%]
	non-speech	speech	
non-speech	1,083	253	<b>81.2</b>
speech	249	5,612	<b>95.7</b>
macro ave.	—	—	<b>88.5</b>

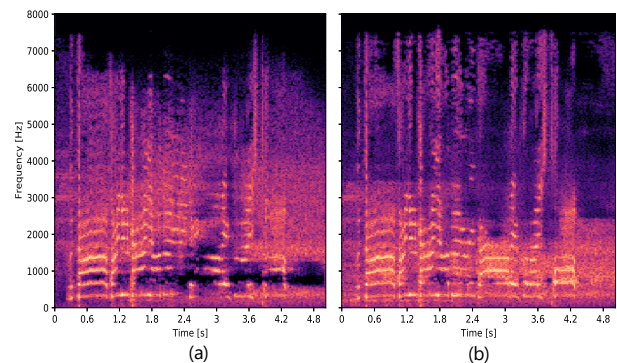


Figure 6: Sound spectrograms of the input audio which are noise-suppressed by (a) the separation model and (b) the tandem model

## 5. Conclusions

We proposed a VAD approach that used a tandem connection of the speech/noise separation model and the VAD model and trained them simultaneously on the VAD task for live baseball game speech with loud noise. Simultaneous optimization training of the tandem model was expected to enhance noise suppression effectiveness and prevent the fatal suppression of speech segments. The proposed model for VAD was evaluated on actual live audio of baseball games. Although the model was trained from pseudo-simulated sounds by synthesizing clean speech and live audio of baseball games, the model achieved good VAD performance on the actual live sound. Finally, the proposed model improved the F1-score by 4.2 points (from 84.3% to 88.5% in the F1-score) compared to the VAD model alone.

In future work, we will use a multi-task model in which part of the layers of the separation and VAD models is shared. Further improvement in VAD accuracy can be expected by constructing a model that shares the feature extraction part of the separation model and the VAD model.

## 6. Acknowledgements

This work was supported by the Hoso Bunka Foundation. Besides, a part of this work was also supported by SPS KAKENHI Grant Number 21H00901.

## 7. References

- [1] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7378–7382.
- [2] T. Sainath, R. J. Weiss, A. Senior, K. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proceedings of INTERSPEECH2015*, 2015, pp. 1–5.
- [3] L. Bai, Z. Zhang, and J. Hu, "Voice activity detection based on deep neural networks and Viterbi," *IOP Conference Series: Materials Science and Engineering*, vol. 231, p. 012042, 2017.
- [4] R. Gemello, F. Mana, and R. De Mori, "Non-linear estimation of voice activity to improve automatic recognition of noisy speech," in *Proceedings of INTERSPEECH2005*, 2005, pp. 2617–2620.
- [5] Y. Jung, Y. Choi, and H. Kim, "Self-adaptive soft voice activity detection using deep neural networks for robust speaker verification," in *Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.
- [6] X. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016.
- [7] G. W. Lee and H. K. Kim, "Multi-Task Learning U-Net for Single-Channel Speech Enhancement and Mask-Based Voice Activity Detection," *Applied Sciences*, vol. 10, no. 9, 2020.
- [8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lecture Notes in Computer Science (International Conference on Medical Image Computing and Computer-Assisted Intervention)*, vol. 9351, pp. 234–241, 2015.
- [9] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7092–7096.
- [10] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.
- [11] S. Mirsamadi and I. Tashev, "Causal Speech Enhancement Combining Data-Driven Learning and Suppression Rule Estimation," in *Proceedings of INTERSPEECH2016*, 2016, pp. 2870–2874.
- [12] I. Tashev and S. Mirsamadi, "DNN-based Causal Voice Activity Detector," in *Proceedings of Information Theory and Applications Workshop*, 2017, pp. 1–5.
- [13] L. Wang, J. Zhu, and I. Kodrasi, "Multi-task single channel speech enhancement using speech presence probability as a secondary task training target," *arXiv preprint: arXiv:2011.07547*, 2020, pp.1–5.
- [14] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 12, p. 2136–2147, 2015.
- [15] M. Hermans and B. Schrauwen, "Training and Analysing Deep Recurrent Neural Networks," in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26, 2013, pp. 1–9.
- [16] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," in *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, 2014, pp. 1–13.
- [17] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for Speech Enhancement," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 249–253.
- [18] M. N. Ali, A. Brutti, and D. Falavigna, "Speech Enhancement Using Dilated Wave-U-Net: an Experimental Analysis," in *2020 27th Conference of Open Innovations Association (FRUCT)*, 2020, pp. 3–9.
- [19] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing Voice Separation with Deep U-Net Convolutional Networks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 745–751.
- [20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.
- [21] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.
- [22] "hcmmlab/vadnet: Real-time Voice Activity Detection in Noisy Environments using Deep Neural Networks," <https://github.com/hcmmlab/vadnet>.
- [23] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" in *Proceedings of INTERSPEECH2018*, 2018, pp. 147–151.
- [24] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, 2003, pp. 7–12.
- [25] "radiko," <https://radiko.jp/>.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015, pp. 1–15.