

SPEECH ENHANCEMENT WITH WEAKLY LABELLED DATA FROM AUDIOSET

Qiutiang Kong, Haohe Liu, Xingjian Du, Li Chen, Rui Xia, Yuxuan Wang

ByteDance, Shanghai, China

{kongqiutiang, liuhaohe.0379, duxingjian.real, chenli.cloud,
rui.xia, wangyuxuan.11}@bytedance.com

ABSTRACT

Speech enhancement is a task to improve the intelligibility and perceptual quality of degraded speech signal. Recently, neural networks based methods have been applied to speech enhancement. However, many neural network based methods require noisy and clean speech pairs for training. We propose a speech enhancement framework that can be trained with large-scale weakly labelled AudioSet dataset. Weakly labelled data only contain audio tags of audio clips, but not the onset or offset times of speech. We first apply pretrained audio neural networks (PANNs) to detect anchor segments that contain speech or sound events in audio clips. Then, we randomly mix two detected anchor segments containing speech and sound events as a mixture, and build a conditional source separation network using PANNs predictions as soft conditions for speech enhancement. In inference, we input a noisy speech signal with the one-hot encoding of “Speech” as a condition to the trained system to predict enhanced speech. Our system achieves a PESQ of 2.28 and an SSNR of 8.75 dB on the VoiceBank-DEMAND dataset, outperforming the previous SEGAN system of 2.16 and 7.73 dB respectively.

Index Terms— Speech enhancement, weakly labelled data, AudioSet.

1. INTRODUCTION

Speech enhancement (SE) is a task to improve the intelligibility and perceptual quality of degraded speech signal. Speech enhancement has many applications in our life, such as teleconference, mobile phone calls, automatic speech recognition and hearing aids [1]. Early works of speech enhancement applied signal processing methods such as minimum-mean square error short-time spectral amplitude estimator [2] and non-negative matrix factorization (NMF) [3]. Those conventional methods perform well under stationary noise, but have limited performance under non-stationary noise or in low signal-to-noise ratio (SNR) environments. Recently, neural network based methods have been proposed for speech enhancement, such as denoising autoencoder [4], fully connected neural networks [5], recurrent neural networks (RNNs) [6], convolutional neural networks [7, 8] (CNNs) or time do-

main CNNs [9, 10, 11] and generative adversarial networks (GANs) [12, 13]. Those neural network based speech enhancement methods require clean speech and background noise for training. Recently, universal source separation systems [14, 15] have been proposed for source separation without clean training data.

However, previous neural network based speech enhancement methods require clean speech and background noise for training, while collecting clean speech and background noise can be difficult and time consuming. For example, the background noises recorded in the laboratories [16] can be different from real world sounds. It is difficult to collect a large-scale dataset covering a wide range of sounds in our world. In addition, speech datasets such as TIMIT [17] and VoiceBank [18] contain neutral emotion speech, while there can be various emotions of speech in our real life. Recently, a large-scale AudioSet [19] dataset containing hundreds of different sound classes from YouTube was released, which provides a larger variety of sounds than previous speech and noise datasets.

However, the difficulty of using AudioSet for speech enhancement is that audio clips in AudioSet are weakly labelled. That is, each audio clip is only labelled the presence or absence of sound events, without knowing their onset and offset times. Also, AudioSet does not indicate clean speech in audio clips, and speech are usually mixed with other sound events. In this article, we propose a speech enhancement framework trained with weakly labelled data. First, we apply pretrained audio neural networks (PANNs) [20] to select 2-second *anchor segments* that are most likely to contain speech or sound events in an audio clip. One contribution of this work is that we propose an anchor segment mining algorithm to better select anchor segments for creating mixtures. Two randomly selected anchor segments are used to constitute a mixture. Then a convolutional UNet [21] is used to predict the waveform of individual anchor segments. We extend the loss function calculated on spectrogram [22] to a loss function calculated in the waveform domain. For the speech enhancement task, we evaluate various metrics including PESQ, CSIG, etc. that were not discussed in [22].

This paper is organized as follows: Section 2 introduces our speech enhancement system trained with weakly labelled data. Section 3 shows the experiment results. Section 4 con-

cludes this work.

2. SPEECH ENHANCEMENT WITH WEAKLY LABELLED DATA

2.1. Neural Network Based Speech Enhancement

Recently, neural network based methods have been applied to speech enhancement, and have outperformed conventional speech enhancement methods [5]. The neural network based speech enhancement methods require pairs of noisy speech and clean speech for training. We denote a noisy speech as $x \in \mathbb{R}^L$, and its corresponding clean speech as $s \in \mathbb{R}^L$, where L is the number of samples in an audio clip. Then, a neural network learns a mapping: $f : x \mapsto s$, where f can be modeled by a neural network with learnable parameters, such as fully connected neural networks [5], RNNs [6], CNNs [7, 8] and time domain CNNs [9, 10, 11]. We denote the enhanced speech as $\hat{s} = f(x)$. In training, the parameters of f can be optimized by minimizing a loss function $l(\hat{s}, s)$, such as a mean absolute error (MAE) loss:

$$l_{\text{MAE}} = \|\hat{s} - s\|_1, \quad (1)$$

where $\|\cdot\|_1$ is an l_1 norm. In inference, the enhanced speech \hat{s} can be calculated by $\hat{s} = f(x)$. However, one disadvantage of the above neural network based speech enhancement method is that noisy and clean speech pairs are required for training, which can be difficult and time consuming to obtain. To address this problem, we propose a speech enhancement framework that can be trained with weakly labelled data. That is, training a speech enhancement system from audio clips containing noisy speech.

2.2. Speech Enhancement with Weakly Labelled Data

Our speech enhancement system is trained with a large-scale weakly labelled AudioSet [19] dataset containing 527 kinds of sound classes. Most of audio clips have durations of 10 seconds. AudioSet is weakly labelled, that is, each audio clip is only labelled with tags, but without onset and offset times of sound events. Also, AudioSet does not indicate clean speech, where speech are usually mixed with other sounds under unknown SNR. Previous works have investigated general source separation with weakly labelled data [22]. Our improvement to [22] is that we propose a novel anchor segment mining algorithm in Section 2.5. To begin with, we denote two anchor segments containing different sound events as s_1 and s_2 respectively. The anchor segments s_1 and s_2 are selected from two audio clips that are most likely to contain speech or sound events. The anchor segments s_1 and s_2 are selected to have disjoint audio tags that is described in Section 2.5. In training, we build a neural network to learn a mapping:

$$f(s_1 + s_2, c_1) \mapsto s_1, \quad (2)$$

where $c_1 \in [0, 1]^K$ is a conditional vector that controls what source to separate, and K is the number of sound classes in AudioSet. In training, there is no need for s_1 or s_2 to be clean. The conditional vector c_1 is the audio tagging probability calculated on s_1 . To explain, if s_1 contains both “Speech” and “Water”. When conditioning on the audio tagging probability c_1 , the system (2) will separate both “Speech” and “Water”. In inference, the enhanced speech \hat{s} can be obtained by input a noisy speech x and setting the conditional vector c as the one-hot encoding of “Speech”:

$$\hat{s} = f(x, c). \quad (3)$$

To explain, the training of the speech enhancement system described in (2) does not require clean speech. Still, we can obtain clean speech from noisy speech using the trained speech enhancement system in (3).

2.3. Sound Event Detection for Selecting Anchor Segments

The anchor segments s_1 and s_2 are 2-second segments used to constitute a mixture as input. To begin with, we randomly select two sound classes from AudioSet. For each sound class, we randomly select an audio clip in AudioSet. However, there are no information of when the sound classes occur in audio clips. Therefore, we apply a sound event detection (SED) system [20] to predict the frame-wise presence probability of the sound class. The SED system is a DecisionLevelMax system from PANNs [20], which applies log mel spectrogram as input feature, and uses a 14-layer CNN as a classification model. Each convolutional layer has a kernel size of 3×3 . The convolutional layers are followed by a time distributed fully connected layer with K outputs to predict the frame-wise presence probability of sound classes. The frame-wise predictions are max pooled along the time axis to obtain clip-wise predictions. We denote the weak labels of an audio clip as $y \in \{0, 1\}^K$, and its clip-wise prediction as $\hat{y} \in [0, 1]^K$. The SED system is trained by minimizing a binary cross-entropy loss [20] between predicted and target weak label tags:

$$\text{loss} = - \sum_{k=1}^K y_k \ln \hat{y}_k + (1 - y_k) \ln (1 - \hat{y}_k). \quad (4)$$

The first row of Fig. 1 shows the log mel spectrogram of a 10-second audio clip from AudioSet containing “Speech” and other sound classes. The second row shows the frame-wise SED prediction of “Speech”. We select anchor segment s_1 that is most likely to contain the selected sound event, as shown in the red block in Fig. 1. Similarly, we select anchor segment s_2 from another audio clip. Then, we mix $s_1 + s_2$ as input to (2).

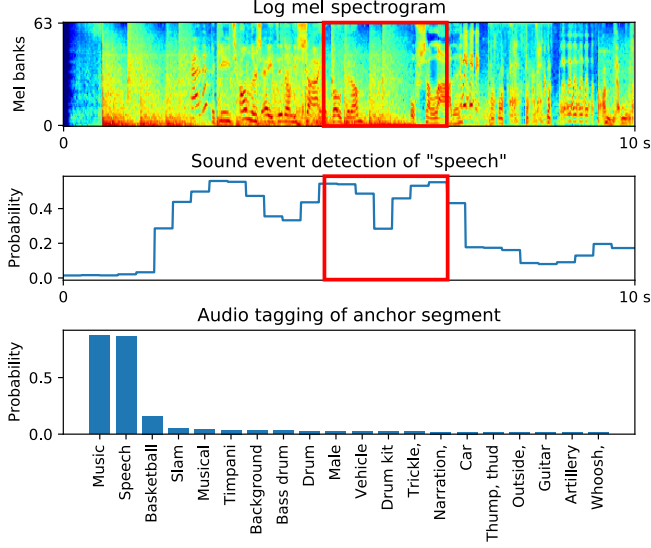


Fig. 1. Top: log mel spectrogram of a 10-second audio clip from AudioSet; Middle: predicted SED probability of “Speech”, where red block shows the selected anchor segment; Bottom: predicted audio tagging probabilities of the anchor segment.

2.4. Audio Tagging for Constructing Conditional Vector

The conditional vector c_1 controls what sources to separate from $s_1 + s_2$. However, there is no ground truth label of s_1 , so c_1 is unknown. In addition, there can be multiple sound events in s_1 . We apply an audio tagging system on s_1 : $c_1 = g_{AT}(s_1)$ to estimate the conditional vector c_1 . The audio tagging system g_{AT} is a 14-layer CNN of PANNs [20]. The 14-layer CNN consists of several convolutional layers. Then, global average and max pooling are applied to summarize the feature maps into a fixed dimension embedding vector. Finally, a fully connected layer is applied on the embedding vector to predict the presence probability of sound events. The training of the audio tagging system applies the binary crossentropy described in (4). The advantage of using audio tagging prediction rather than one-hot encoding of labels to build c_1 is that $g_{AT}(s_1)$ provides a better estimation of sound events probability in s_1 than labels. From the top to bottom in Fig. 1 shows the log mel spectrogram of a 10-second audio clip, the SED result of the audio clip, and the audio tagging probabilities of the selected anchor segment. For example, the predominant sound events in s_1 are “Music” and “Speech”. Other sound classes in s_1 include “Basketball” and “Slam”, etc.

2.5. Anchor Segment Mining

Previous work [22] proposed to select s_1 and s_2 randomly from AudioSet. However, if s_1 and s_2 contain mutual sound classes, the separation result of equation (2) can be incorrect. For example, if both s_1 and s_2 contain clean “Speech”, when

Algorithm 1 Anchor segment mining.

- 1: Mini-batch of anchor segments: $S = \{s_1, \dots, s_B\}$, and their predicted tags: $R = \{r_1, \dots, r_B\}$.
 - 2: **for** $r_1 \in R$ **do**
 - 3: **for** $r_2 \in R$ **do**
 - 4: **if** $r_1 \cap r_2 = \emptyset$ **then**
 - 5: Collect anchor segments of r_1 and r_2 to constitute s_1 and s_2 .
 - 6: Remove r_1 and r_2 from R .
 - 7: **end if**
 - 8: **end for**
 - 9: **end for**
-

the conditional vector c_1 is the one-hot encoding of “Speech”, the system in (2) will learn to only separate “Speech” from s_1 , but not “Speech” from s_2 . For a speech enhancement system, we aim to separate all “Speech” from both s_1 and s_2 . To address this problem, we propose an anchor segment mining method to select s_1 and s_2 to have disjoint conditional vectors. In training, we randomly select B anchor segments to constitute a mini-batch $\{s_1, \dots, s_B\}$, where B is the mini-batch size. Then, we calculate the conditional vectors $\{c_1, \dots, c_B\}$ by $g_{AT}(s_1)$. For each conditional vector c_b , we apply thresholds to predict their present tags r_b , where the thresholds are calculated from PANNs [20] with equal precision and recall for sound classes. Then, we propose a mining algorithm described in Algorithm 1 to select pairs of anchor segments to have disjoint predicted tags from the mini-batch to constitute s_1 and s_2 .

2.6. Separation Model

We apply convolutional UNets [21, 22] on the spectrogram of mixture to build separation systems. To begin with, the waveform of a mixture is transformed into a spectrogram. A UNet consists of an encoder and a decoder. The encoder consists of 12 convolutional layers with kernel sizes of 3×3 to extract high-level representations. Downsampling layers with sizes of 2×2 are applied to every two convolutional layers. The decoder is symmetric to the encoder with 12 convolutional layers. Transposed convolutional layers are used to upsample feature maps after every two convolutional layers. Shortcut connections are added between encoder and decoder layers with same hierarchies. In each convolutional layer, the conditional vector c_1 is multiplied with a learnable matrix, and is added to the feature maps as biases. This bias information controls what sound events to separate from a mixture. The decoder outputs a spectrogram mask with values between 0 and 1, and is multiplied to the mixture spectrogram to obtain the separated spectrogram of s_1 . Then, an inverse short time Fourier transform (ISTFT) is applied on the separated spectrogram using the phase of mixture to obtain \hat{s}_1 . The separation system is trained by minimizing the loss function (1).

3. EXPERIMENTS

Our speech enhancement system is trained on the balanced subset of the weakly labelled AudioSet [19] containing 20,550 audio clips with 527 sound classes. The audio clips have durations of 10-second. Audio clips are weakly labelled, and there can be multiple sound events in an audio clip. There are 5,251 audio clips containing “Speech”. To begin with, we resample all audio clips to 32 kHz to be consistent with the configuration of PANNs [20]. The sound event detection and audio tagging systems from PANNs are used to select anchor segments as described in Section 2.5. To build the separation system, we extract spectrograms of mixtures using short time Fourier transform (STFT) with a window size 1024 and a hop size 320. All anchor segments have durations of 2 seconds. We set mini-batch size to 24. Adam optimizer [23] is used for training. We trained the system for 1 million iterations using a single Tesla-V100-SXM2-32GB GPU card in one week.

We evaluate our proposed speech enhancement system directly on the test set of the VoiceBank [18] and DEMAND [16] datasets without training on them. There are 824 paired noisy and clean speech for testing in VoiceBank-DEMAND. Each audio clip has a sample rate of 48 kHz. The noisy speech have four SDR settings of 15, 10, 5 and 0 dB. There are 10 types of noise, including 2 types of synthetic noise and 8 types of noise from DEMAND. There are 28 speakers from VoiceBank. The major difference between our speech enhancement method with previous works is that, we do not use the training data from VoiceBANK-DEMAND, and directly evaluate our speech enhancement system on the test clips.

Following previous works of speech enhancement [24, 12, 25], we apply Perceptual evaluation of speech quality (PESQ) [26], Mean opinion score (MOS) predictor of signal distortion (CSIG), MOS predictor of background-noise intrusiveness (CBAK), MOS predictor of overall signal quality (COVL) [27] and segmental signal-to-ratio noise (SSNR) [28] to evaluate the speech enhancement performance. Table 1 shows that noisy speech without enhancement achieves PESQ, CSIG, CBAK, COVL, SSNR of 1.97, 3.35, 2.44, 2.63 and 1.68 dB respectively. Our proposed speech enhancement system achieves a PESQ of 2.28, outperforming the Wiener [24] and SEGAN [12] systems. Our system achieves a CBAK of 2.96 and an SSNR of 8.75 dB, outperforming the Wiener and SEGAN systems of 2.68 and 5.07 dB, indicating the effectiveness of training speech enhancement with weakly labelled data. On the other hand, our system achieves a CSIG of 2.43 and COVL of 2.30, lower than other systems, indicating that our speech enhancement may lost details of speech, especially the high frequency component shown in Fig. 2. The left and right columns of Fig. 2 visualizes two speech enhancement examples of our proposed system. From top to bottom rows show the log mel spectrogram of noisy speeches, target clean speeches and enhanced speeches respectively. Considering that our system is trained with

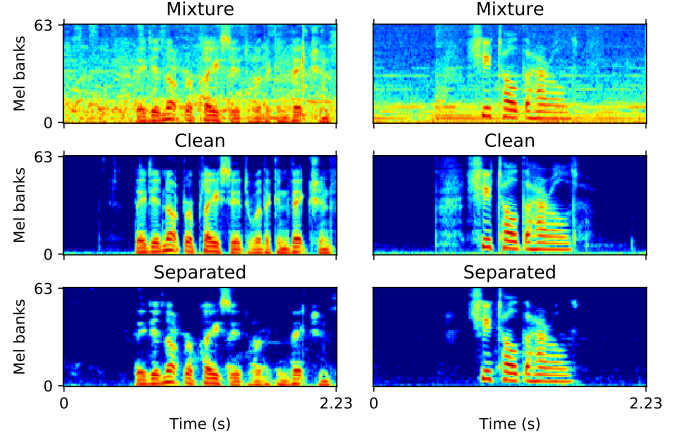


Fig. 2. The left and right columns show two examples of speech enhancement. Top: log mel spectrogram of noisy speech; Middle: ground truth clean speech; Bottom: enhanced speech.

Table 1. Speech enhancement results

	PESQ	CSIG	CBAK	COVL	SSNR
Noisy	1.97	3.35	2.44	2.63	1.68
Wiener [24]	2.22	3.23	2.68	2.67	5.07
SEGAN [12]	2.16	3.48	2.94	2.80	7.73
Wave-U-Net [25]	2.40	3.52	3.24	2.96	9.97
Proposed	2.28	2.43	2.96	2.30	8.75

weakly labelled data only, and does not use any training data from VoiceBank-DEMAND. We show that training a speech enhancement system from weakly labelled data is possible. We provide our speech enhancement demos in the following links¹².

4. CONCLUSION

In this work, we propose a speech enhancement system trained with weakly labelled data from AudioSet. Our system does not require clean speech and background noise to train the speech enhancement system. We propose to use the sound event detection and audio tagging system from pretrained audio neural networks (PANNs), and an anchor segment mining algorithm for selecting anchor segments. We build conditional UNet sound separation systems for speech enhancement. Our proposed systems outperform the Wiener and SEGAN systems evaluated with the VoiceBank-DEMAND dataset in the PESQ, CBAK and SSNR metrics without using any training data from VoiceBank-DEMAND. In future, we will continue to investigate general source separation with weakly labelled data.

¹<https://www.youtube.com/watch?v=q3hVnpNcpBI>

²<https://www.youtube.com/watch?v=DzQvn820u8E>

5. REFERENCES

- [1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC press, 2013.
- [2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing (TASLP)*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [3] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 10, pp. 2140–2151, 2013.
- [4] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *INTERSPEECH*, 2013, pp. 436–440.
- [5] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2014.
- [6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation (LVA-ICA)*. Springer, 2015, pp. 91–99.
- [7] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *INTERSPEECH*, 2017.
- [8] S. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement," in *INTERSPEECH*, 2016, pp. 3768–3772.
- [9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [10] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5069–5073.
- [11] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6875–6879.
- [12] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017.
- [13] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5024–5028.
- [14] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J Weiss, Kevin Wilson, and John R Hershey, "Unsupervised sound separation using mixture invariant training," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- [15] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J Weiss, Kevin Wilson, and John R Hershey, "Unsupervised speech separation using mixtures of mixtures," in *Workshop on International Conference on Machine Learning (ICML)*, 2020.
- [16] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proceedings of Meetings on Acoustics*, 2013.
- [17] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.
- [18] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *International Conference Oriental COCOSDA with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013.
- [19] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.
- [20] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pre-trained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [21] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *International Society for Music Information Retrieval (ISMIR)*, 2017, pp. 745–751.
- [22] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source separation with weakly labelled data: An approach to computational auditory scene analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 101–105.
- [23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
- [24] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996, vol. 2, pp. 629–632.
- [25] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.
- [26] ITU-T Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.
- [27] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE/ACM Transactions on audio, speech, and language processing (TASLP)*, vol. 16, no. 1, pp. 229–238, 2007.
- [28] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*, Prentice Hall, 1988.