Speech Enhancement in the Presence of Background Music Considering Speech and Music Characteristics

Jeongwoo Woo[1,a)] Masato Mimura[1,b)] Kazuyoshi Yoshii[1,c)]
Tatsuya Kawahara[1,d)]

Abstract: Speech enhancement in the presence of background music is not so different from noise reduction if music is treated as just noise. However, music has definite characteristics which are made by human beings, unlike noise which can be any. In order to consider characteristics of background music instead of noise reduction, we introduce a generative adversarial network (GAN). We combine two multi-scale discriminators for speech and music with Conv-TasNet modified for speech enhancement. We train it jointly with SI-SDR and the GAN objective. Experimental evaluations through speech recognition demonstrate that the proposed model is improved from the baseline model. It is notable that the more music interference is large, the more the proposed method is effective. Comparing the spectrogram of enhanced speech by the proposed and baseline model demonstrate that the baseline model tends to cut off noise excessively, in contrast the proposed model reconstructs more faithfully.

Keywords: Speech Enhancement, GAN, Background Music Interference

1. Introduction

Speech enhancement in the presence of background music is not so different from noise reduction if music is treated as just noise. However, music has definite characteristics which are made by human beings, unlike noise which can be any. We do not recognize white noise or the nerve-racking noise as music. In order to consider characteristics of background music instead of noise reduction, we introduce a generative adversarial network (GAN).

The GAN [1] is a framework proposed by Goodfellow et al. for estimating generative models via an adversarial process, in which two models, a generator and a discriminator, are trained simultaneously. With the adversarial training, the discriminator is trained to discriminate real data from fake data generated by the generator, and the generator is trained to generate the fake data which would not be discriminated from real data. In this process, the generative model learn to make its distribution approximate to the real data distribution, not to make its output approximate to each certain point in the real data distribution. This learning the real data distribution makes the model possible to consider speech and music characteristics.

[1] Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
[a)] woo@sap.ist.i.kyoto-u.ac.jp
[b)] mimura@sap.ist.i.kyoto-u.ac.jp
[c)] yoshii@sap.ist.i.kyoto-u.ac.jp
[d)] kawahara@sap.ist.i.kyoto-u.ac.jp

In this paper, we propose combining two discriminative networks for speech and music with a speech enhancement model on time domain. We exploit a fully-convolutional time-domain audio separation network (Conv-TasNet) [6] as a speech enhancement model and modify a discriminator of MelGAN [3] as a discriminative network.

We jointly train the proposed model with scale-invariant source-to-distortion ratio (SI-SDR) [4] and the training objective of GAN. We evaluate our method through speech enhancement on the metric of source-to-distortion ratio (SDR) [9] and speech recognition on the metric of word error rate (WER).

The rest of this paper are organized as follows. We introduce the discriminative network for considering speech and music characteristics in Section 2, describe combining discriminative networks with speech enhancement model and training it jointly in Section 3, present the experiment procedures in Section 4, analyze the experiment results in Section 5, and conclude this paper in Section 6.

2. Considering Speech and Music Characteristics

We modify a discriminator of MelGAN to consider speech and music characteristics. MelGAN is a non-autoregressive feed-forward convolutional architecture to perform audio waveform generation in the GAN framework. Since it achieved the first method that successfully trains GAN for raw audio generation without additional process, we expect

that its discriminator is appropriate to consider speech and music characteristics during speech enhancement. The MelGAN discriminator is multi-scale architecture [10] with 3 discriminators (D_1, D_2, D_3) that have an identical network structure but operate on different audio scales. D_1 operates on the scale of raw waveform, whereas D_2, D_3 operate on raw waveform downsampled by a factor of 2 and 4, respectively. The downsampling is performed using average pooling with a kernel size of 4 and a stride. Multiple discriminator structure has an inductive bias such that each discriminator learns features for different frequency range of the waveform. Specifically, the discriminator operating on downsampled audio does not have access to high frequency component, therefore, it is biased to learn discriminative features on the low frequency components only.

Each individual discriminator is a Markovian window-based discriminator [2] consisting of a sequence of convolutional layers with a stride and a large kernel size. While the standard GAN discriminator learns to classify between distributions of entire waveform sequence, window-based discriminator learns to classify between distributions of small chunks. Since the discriminator loss is computed over the overlapping windows where each window is very large as equal to the receptive field of the discriminator, it learns to maintain coherence across patches.

The training objective is the hinge loss version of the GAN objective [5]. The training objective of the discriminator is shown as follows:

$$\begin{aligned}\min_{D_k} \mathbb{E}_{\mathbf{x}}[\min(0, 1 - D_k(\mathbf{x}))] \\ + \mathbb{E}_{\mathbf{s},\mathbf{z}}[\min(0, 1 + D_k(G(\mathbf{s}, \mathbf{z})))], \forall k = 1, 2, 3\end{aligned} \tag{1}$$

where $\mathbf{x}$ represents the raw waveform, $\mathbf{s}$ represents the conditioning information (eg. mel-spectrogram), and $\mathbf{z}$ represents the Gaussian noise vector.

3. Speech Enhancement with Discriminative Network

We exploit the Conv-TasNet for speech enhancement, which is proposed for multi-speaker source separation, operates on time domain. The training objective of Conv-TasNet is maximizing the SI-SDR which is computed as follows:

$$\begin{cases} \mathbf{s}_{target} \coloneqq \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \\ \mathbf{e}_{noise} \coloneqq \hat{\mathbf{s}} - \mathbf{s} \\ \text{SI-SDR} \coloneqq 10 \log_{10} \frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{noise}\|^2} \end{cases} \tag{2}$$

where $\hat{\mathbf{s}}$ and $\mathbf{s}$ are the estimated and original clean sources, respectively, $\|\mathbf{s}\|^2 = \langle \mathbf{s}, \mathbf{s} \rangle$ denotes the signal power. Scale invariance is ensured by normalizing $\hat{\mathbf{s}}$ and $\mathbf{s}$ to zero-mean prior to the calculation.

Since there are two target sources, the final training objective of speech enhancement is minimizing the loss fucntion $L^{(ENH)}$ defined as follows:

$$L^{(ENH)} = -\frac{\text{SI-SDR}_{speech} + \text{SI-SDR}_{music}}{2} \tag{3}$$

where, SI-SDR_{speech} and SI-SDR_{music} are the SI-SDR of speech and music targets, respectively. In the original Conv-TasNet paper, Utterance-level permutation invariant training (uPIT) is applied to address the source permutation problem, but there is no need to address it with speech and music targets.

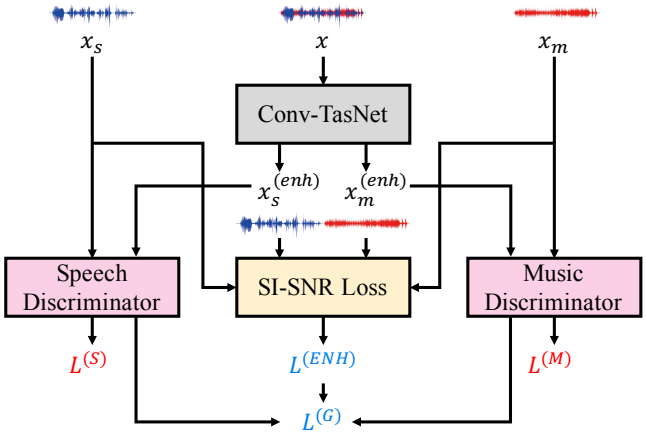

Fig. 1 The architecture of the proposed model.

We combine two MelGAN discriminators for each target with Conv-TasNet. For adversarial training, we regard the Conv-TasNet as a generator G with one discriminator for speech $D^{(S)}$ and another one for music $D^{(M)}$. Generally, the Gaussian noise $\mathbf{z}$ is provided to the generator of GAN as an input, but we modify the model without noise. Since G is conditioned on the input mixture waveform $\mathbf{x}$, there is a variability in the input of the G even in the absence of noise. Hence, noise is not necessary anymore, which is consistent with Mathieu et al. [7]. The training objective is composed of the objective of GAN and minimizing the enhancement loss function $L^{(ENH)}$ shown in Equation (3).

The objective of GAN can be expressed as follows:

$$\begin{aligned} L^{(S)} &= \min_{D_k^{(S)}} \mathbb{E}_{\mathbf{x}_s}[\min(0, 1 - D_k^{(S)}(\mathbf{x}_s)] \\ &\quad + \mathbb{E}_{\mathbf{x}}[\min(0, 1 + D_k^{(S)}(G(\mathbf{x})^{(S)}))] \\ L^{(M)} &= \min_{D_k^{(M)}} \mathbb{E}_{\mathbf{x}_m}[\min(0, 1 - D_k^{(M)}(\mathbf{x}_m)] \\ &\quad + \mathbb{E}_{\mathbf{x}}[\min(0, 1 + D_k^{(M)}(G(\mathbf{x})^{(M)}))] \\ \min_G \mathbb{E}_{\mathbf{x}}[&\sum_{i \in \{S,M\}} \sum_{k=1,2,3} -D_k^{(i)}(G(\mathbf{x})^{(i)})] \end{aligned} \tag{4}$$

where $\forall k = 1, 2, 3$, $\mathbf{x}$ represents the mixture waveform, $\mathbf{x}_s$ represents the speech waveform, $\mathbf{x}_m$ represents the music waveform, $G(\mathbf{x})^{(s)}$ represents the speech waveform generated by G given $\mathbf{x}$, and $G(\mathbf{x})^{(m)}$ represents the music waveform generated by G given $\mathbf{x}$. With the enhancement loss function, we use the following final objective to train G:

$$\begin{aligned} L^{(G)} = \min_G \mathbb{E}_{\mathbf{x}}[&\sum_{i \in \{S,M\}} \sum_{k=1,2,3} -D_k^{(i)}(G(\mathbf{x})^{(i)})] \\ &+ \lambda L^{(ENH)} \end{aligned} \tag{5}$$

and we fix $\lambda = 10$.

In the original MelGAN work, there is a feature mapping objective which is an additional training objective of the

Table 1 WER and SDR on CSJ-anime at all SNR level for Baseline model and Proposed model with cascading Clean ASR and Mixture ASR

WER: Word error rate, SDR: Signal-to-distortion ratio

Speech Enhacement	ASR	WER (%) 5 dB	0 dB	-5 dB	average	SDR (dB) 5 dB	0 dB	-5 dB
Baseline	–					20.94	18.38	15.64
	Clean ASR	13.89	17.97	27.59	19.78			
	Mixture ASR	13.82	16.10	22.58	17.50			
Proposed	–					20.93	18.31	15.47
	Clean ASR	13.80	17.26	26.50	19.18			
	Mixture ASR	13.75	15.78	21.72	**17.08**			

generator to minimize the L1 distance between the discriminator feature maps of real and synthetic audio. Since this can be seen as a similarity metric and we already have a similarity metric SI-SNR, we do not apply the feature mapping objective. The architecture of proposed model is shown in Figure 1.

4. Experimental Details

In this section, we describe the experimental details. We evaluate the proposed model on speech enhancement and speech recognition performance.

4.1 Dataset

Experimental mixture data were generated by mixing utterances from speech database with background music. Both speech and music are sampled at 16 kHz. As the speech database, we used the Corpus of Spontaneous Japanese Academic Presentation Speech (CSJ-APS). The CSJ-APS has a duration of around 260 hours and consists of live recordings of academic presentations in nine different academic societies. The societies range from engineering, humanities, and social and behavioral sciences. For background music, we used around 30 hours of background music used in Japanese animations.

For the training dataset, we added background music to the speech with randomly sampled source-to-noise ratio (SNR) levels from a normal distribution with a mean of 0 dB and a standard deviation of 5 dB. For the test dataset, we added background music from animations not used for the training dataset to the speech of official CSJ-APS testset 1 with various SNR levels such as 5 dB, 0 dB and -5 dB. This test dataset is referred as the CSJ-anime.

4.2 Baseline Model

In this experiment, we trained speech enhancement Conv-TasNet as a baseline model. It is trained with a configuration which is N=256, L=20, B=256, H=512, P=3, X=8, R=4, C=2, following the the hyper-parameter notations in the original paper [6]. We used Adam optimizer to train the network with a learning rate of 1e-3.

4.3 Proposed Model Configuration

The hyper-parameter setting of the proposed model is the same as that of the baseline model. We used Adam optimizer to train the network with a learning rate of 1e-4.

4.4 ASR Model for Speech Recognition

We integrate the proposed model or the baseline model with the ASR models for evaluating on the speech recognition performance. We implemented an attention-based encoder-decoder model for ASR of which detail is following. The encoder consists of two CNN layers with a stride of 2 followed by batch normalization of each, 5-layer bidirectional LSTM of which the number of hidden unit is 320. The decoder is 2-layer LSTM of which the number of hidden unit is 320. The input acoustic feature for the encoder is a 40-channel log-mel filterbank feature. The output of the decoder is a sequence of subwords defined by the byte-pair-encoding (BPE) [8]. The number of the BPE units is 9,515.

There are two types of ASR model used: one referred to as clean ASR is trained on clean speech data, the other referred to as mixture ASR is trained on speech and music mixture data.

5. Results and Discussion

We evaluate the performance in terms of SDR for speech enhancement, and WER for speech recognition. The results are shown in Table 1. The speech enhancement performance of the proposed model is degraded very slightly from that of the baseline model, but speech recognition performance of the proposed model is improved from the baseline model. Relative WER improvements on average over SNR levels with the clean ASR and mixture ASR are 3.03% and 2.4%, respectively. It is notable that the results show the more music interference is large, the more the proposed method is effective. Relative WER improvements of 5 dB, 0 dB, -5 DB SNR levels are 0.65%, 3.95%, 3.95%, respectively with the clean ASR, and 0.51%, 1.99%, 3.81%, respectively with the mixture ASR. A large WER improvements is with observed the clean ASR, as the mixture ASR is more robust to music interference.

The spectrograms of original clean speech, speech enhanced by the proposed model, and speech enhanced by the baseline are shown in Figure 2. Attending the red boxes, it appears that the proposed model reconstructed some parts which are not reconstructed with the baseline model. The baseline model tends to cut off the noise, but it tends to cut off excessively. We can observe the sound of breathing is omitted by the baseline model through listening. Although the performance of speech recognition is not improved so

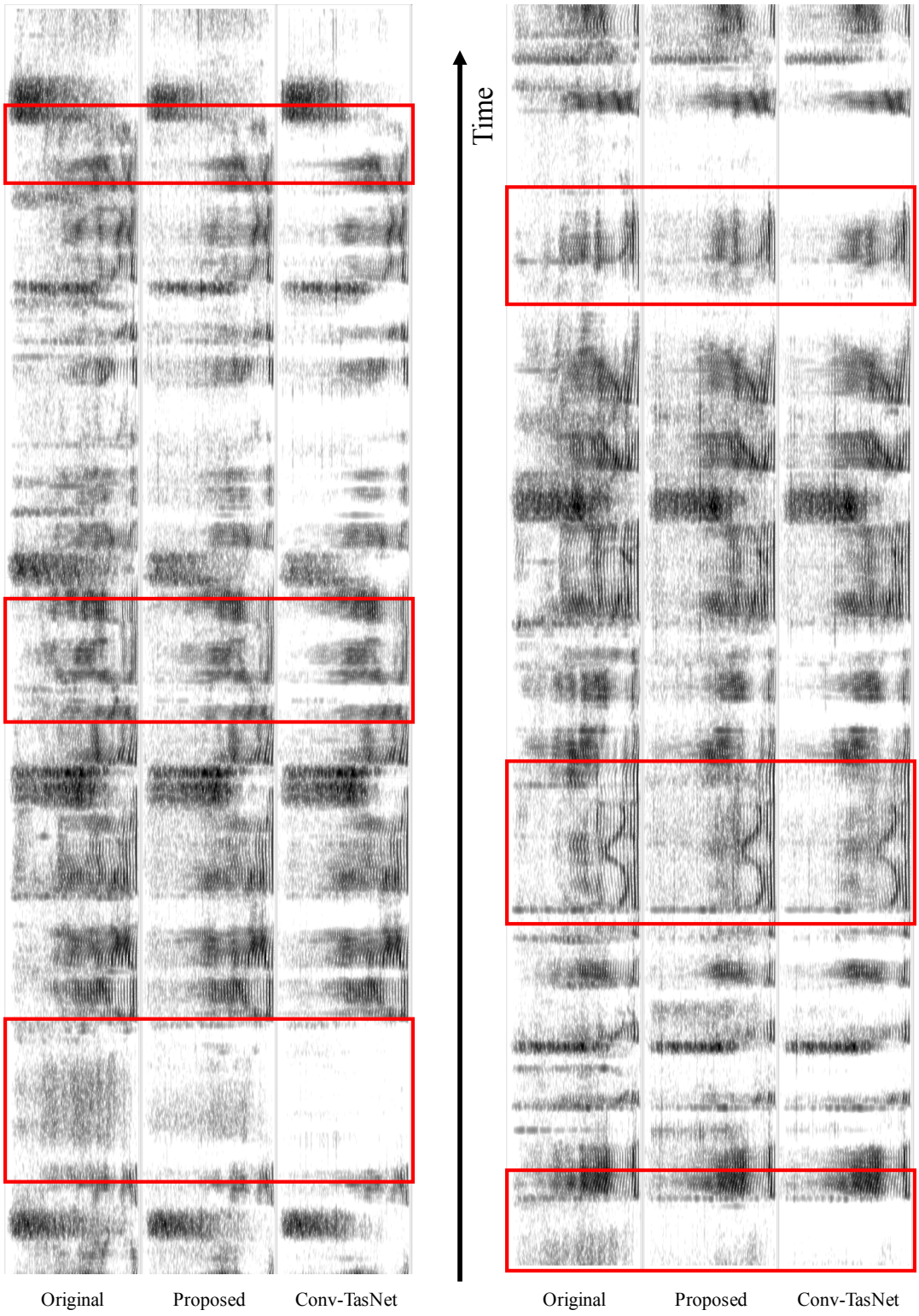

Fig. 2 The spectrograms of original clean speech, speech enhanced by proposed model, and speech enhanced by Conv-TasNet.

much, the proposed model takes into account characteristic with these spectrograms.

6. Conclusion

In this paper, we proposed the method combining speech enhancement with multi-scale discriminator and jointly training with the objective of GAN, for considering speech and music characteristics during speech enhancement. It is more effective for the ASR model trained on clean speech, which is more vulnerable to noise, than that trained on mixture speech. The spectrogram of the enhanced speech is reconstructed by the proposed method more faithfully. These results show that the proposed method considers the speech characteristic during the speech enhancement.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Nets, *Advances in Neural Information Processing Systems* (Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. and Weinberger, K. Q., eds.), Vol. 27, Curran Associates, Inc., pp. 2672–2680 (2014).

[2] Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A.: Image-to-Image Translation with Conditional Adversarial Networks, pp. 5967–5976 (2017).

[3] Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y. and Courville, A. C.: MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis, *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc., pp. 14910–14921 (2019).

[4] Le Roux, J., Wisdom, S., Erdogan, H. and Hershey, J.: SDR - Half-baked or Well Done?, pp. 626–630 (2019).

[5] Lim, J. H. and Ye, J. C.: Geometric GAN (2017).

[6] Luo, Y. and Mesgarani, N.: Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. PP, pp. 1–1 (2019).

[7] Mathieu, M., Couprie, C. and Lecun, Y.: Deep multi-scale video prediction beyond mean square error (2016).

[8] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, pp. 1715–1725 (2016).

[9] Vincent, E., Gribonval, R. and Févotte, C.: Performance measurement in blind audio source separation, *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 14, pp. 1462 – 1469 (2006).

[10] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J. and Catanzaro, B.: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, pp. 8798–8807 (2018).