Speech Enhancement Network with Unsupervised Attention using Invariant Information Clustering

Yosuke SUGIURA*, Shunta Nagamori*, and Tetsuya SHIMAMURA*
* Faculty of Engineering, Saitama University, Saitama, Japan
E-mail: ysugiura, shima@mail.saitama-u.ac.jp Tel: +81-048-858-3776

***Abstract*—In this paper, we propose a new framework for speech enhancement using supervised attention trained by Invariant Information Clustering (IIC). For suppressing an overfitting in the speech enhancement network, the multitask learning with the speaker-invariant information is adopted at the latent representation layer. Several simulations reveal the effectiveness of this method through the speech enhancement experiments.**

I. Introduction

Recently, the demands of speech communication and speech recognition have been increasing as the devices controlled with a voice user interface have been spread. Since such devices are mostly used in noisy environments, speech enhancement technique is growing in importance. In this paper, we focus on a single microphone speech enhancement technique.

Most of the state-of-the-art speech enhancement are operated in frequency domain or time-frequency domain [1], [2]. Although some of them produce excellent speech enhancement results, they usually require a little higher computational complexity because of the use of the Short Time Fourier Transform (STFT) or the wavelet analysis. The end-to-end speech enhancement, which is the time-domain speech enhancement, has the advantage of requiring low computational complexity. However, it is a challenging task since the waveform is more easily corrupted by noise than the spectral features.

Among the existing end-to-end models, Wave-U-Net [3] architecture significantly provides the outstanding performance. Wave-U-Net is composed of the stack of the downsampling fully-convolutional layers and the upsampling fully-convolutional layers. Although the several modifications of Wave-U-Net have been developed [4], there remains a problem that the detailed structures are still degraded in a high noise-level situation.

For accelerating a feature learning, several architectures are introduced or additional metrics are imposed to the loss functions. The former approach includes Wave-U-Net with self-attention mechanism [5], speech enhancement transformer (SETransformer) [6], and so on. The most of the former methods utilize the self-attention [7]. The self-attention can classifies the common structures within the input data, and so that helps the network to understand the important features composing the desired output. Although the self-attention module has a strong effect as intra-attention, this approach, including the multi-head attention, is insufficient to exploit the time-invariant information, such as the speaker information.

The latter approach is so called multi-task training. Speech enhancement GAN [8] and its modifications [11], [9] introduce an adversarial training as the multi-task training. For more effective training, some methods focused on the latent representation of the autoencoder. High-level SEGAN (HLGAN) [10] adds a regularization for the latent variable so as to reduce the distance of the latent variables for the clean speech and the noisy speech. Adversarial Latent Representation Learning (ALRL) [12] adopts an adversarial loss for calculating the distance of the latent variable in HLGAN. They explicitly assume that the optimal latent variable can be obtained from the clean speech input, whereas there is no theoretical validation that this assumption is correct.

In this paper we address to improve learning of the latent representation using the context invariant features. To obtain such the features, we introduce new multi-task training based on information invariant clustering (IIC) [13]. The paired speech which has the different utterance but uttered by the same speaker helps IIC to understand the speaker independent features. Some experiments were conducted to evaluate the performance of the proposed method. Through the comparison with the results of the several conventional models, we reveal that the proposed method can solve the incompatibility between the distortion loss and the adversarial loss, and improve the speech enhancement performance.

II. Speech Enhancement using Wave-U-Net

In this paper, we adopt the model of Wave-U-Net [3] for speech enhancement. Figure 1 shows Wave-U-Net architecture. Wave-U-Net has a bottleneck architecture composing the encoder and decoder to extract the important features which is called a latent variable. The encoder halves the dimension of the feature map in each downsampling block, while the decoder doubles that in each upsampling block. The feature skip-connections are introduced from the encoder layers and the decoder layers with the same level to restore the fine structures. The output layer having a channel-wised convolution produces the enhanced speech of 16,384 samples.

The model parameters are updated by minimizing the reconstruction loss calculated by

$$L_R = \|\boldsymbol{y} - \boldsymbol{d}\|_1, \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{d}$ indicate the reconstruction and the desired clean signal, respectively.

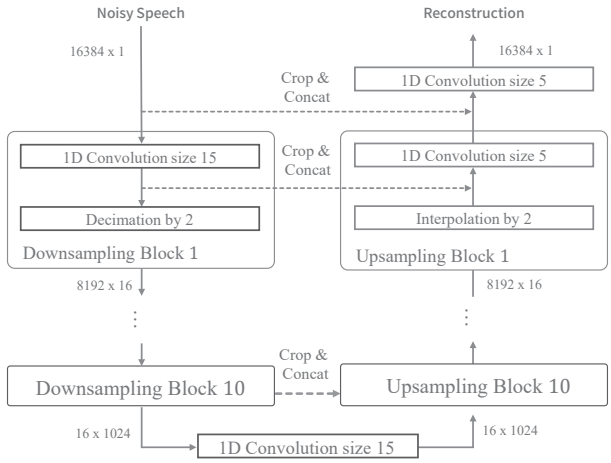

Fig. 1. Structure of Wave-U-Net [3]

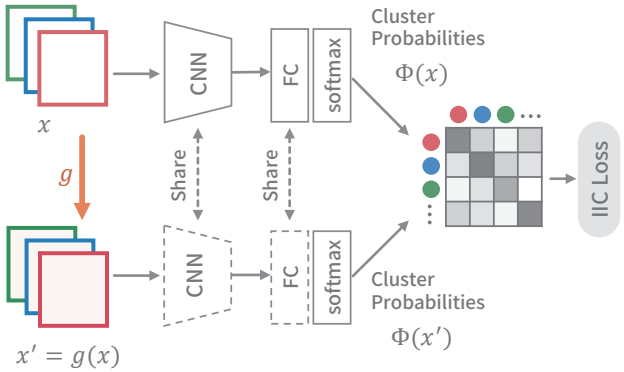

Fig. 2. Architecture of the original IIC [13].

The proposed method additionally imposes the spectral normalization (SN) [14] to all weight parameters for stable learning. The SN regularization guarantees the weights in each layer to satisfy l-1 Lipschitz continuity as

$$\max_{z} \frac{\|\boldsymbol{W}\boldsymbol{z}\|_2}{\|\boldsymbol{z}\|_2} \leq 1 \;\Rightarrow\; \|\boldsymbol{z}\|_2 \geq \|\boldsymbol{W}\boldsymbol{z}\|_2 , \tag{2}$$

where $\boldsymbol{z}$ is the input features for each layers and $\boldsymbol{W}$ is the weight.

III. Reguralization by Context-Invariant Clustering

In this section, we introduce a speech enhancement framework using IIC [13]. IIC is a strong algorithm for unsupervised classification, which discovers clusters underlying unlabelled data samples. The network of IIC simply composes of a single fully-connected layer and a softmax function, and is normally connected to the output of the feature extraction network. Figure 2 shows the network architecture of IIC. IIC requires a source of paired samples $(\boldsymbol{x}, \boldsymbol{x'})$, where $\boldsymbol{x}, \boldsymbol{x'}$ could be different data but belong to the same object class. Namely $\boldsymbol{x}$ and $\boldsymbol{x'}$ have a certain projective relation of $\boldsymbol{x} = g(\boldsymbol{x'})$. The goal of IIC is to learn the network model Φ by maximizing the mutual information between encoded data:

$$\max_{\Phi} I\left(\Phi(\boldsymbol{x}), \Phi(\boldsymbol{x'})\right). \tag{3}$$

Ref. [13] reported that IIC can provide superior performance for the clustering task to the supervised clustering algorithm.

We utilize IIC to accelerate the understanding of the speaker dependent features at the latent representation in Wave-U-Net. Figure 3 shows the learning architecture of Wave-U-Net using IIC. In the proposed method, IIC is adopted to the latent variable of Wave-U-Net. The input paired data is composed of two frame data segmented at the random position within one noisy speech utterance. The batch data is set to the stack of the various speaker's paired data. According to [13], the loss function of IIC to be minimized can be written by

$$L_I = P\left(\log(P) - \alpha\log(P_j) - \alpha\log(P_i)\right), \tag{4}$$

$$P = \frac{1}{n}\sum_{i=1}^{n} \Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x'_i})^T, \tag{5}$$

$$P_j = \frac{1}{n}\sum_{i=1}^{n} \Phi(\boldsymbol{x}_i), \tag{6}$$

$$P_i = \frac{1}{n}\sum_{i=1}^{n} \Phi(\boldsymbol{x'_i}), \tag{7}$$

where $\alpha(>0)$ is a constant to adjust the influence of the marginal entropies. In our method, the projection function Φ operates the time shift and so can be represented by

$$\Phi\colon = \{x(n+k)|x(n) \in \boldsymbol{x}\}, \tag{8}$$

where k is an integer constant. The overall loss function is given by

$$L = L_R + \beta L_I, \tag{9}$$

where β is a strength of regularization.

Due to the nature of the unsupervised clustering algorithm, the number of classes can be set arbitrarily. In this paper, the number of classes is set to 5, which is less than the number of the speakers included in the dataset described below. The aim of this setting is that IIC could roughly integrate the speaker independent features while avoiding the overfitting.

Using this learning architecture, the encoder can learn the context invariant features, and thus we expect Wave-U-Net to flexibly adjust the reconstruction process due to the speaker dependent features underlying the input data.

IV. Experimental Setup

A. Dataset

To evaluate the performance of the proposed architecture on the speech enhancement task, we employed VCTK speech dataset [15] and DEMAND dataset [16], which are the same datasets used in [8]. For the training set, 10 types of noise and 10 different sentences with 4 signal-to-noise ratio (SNR)

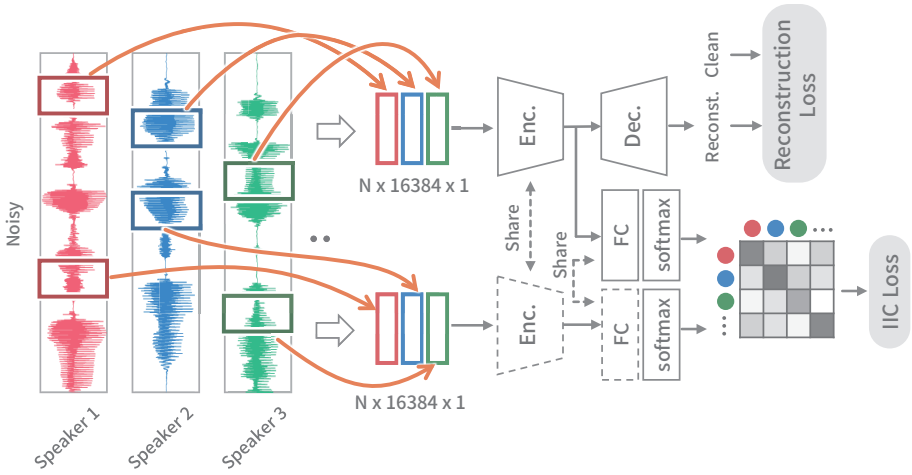

Fig. 3. Learning Architecture of Wave-U-Net with the context invariant clustering.

(15, 10, 5, and 0 dB) were used. For the test set, 5 types of noise and 4 different sentences with 4 SNR (17.5, 12.5, 7.5, and 2.5 dB) were used. Each speech data has the sampling frequency of 16kHz. During training, the speech waveform segments with length of 16,384 samples were extracted from the training data with 50% overlap. During testing, the length of the speech waveform segment was the same as that of training, but the ratio of the overlap was changed to 75%. As in [8], a high-frequency pre-emphasis filter of coefficient 0.95 was applied to each segment of the waveform. The segments of the waveform produced by the trained model were de-empasized and eventually concatenated to reconstruct the enhanced speech waveform.

B. Settings

The model of Wave-U-Net that we useed is completely the same as in [3]. We omitted the over-clustering loss shown in [13]. The settings of the other hyper parameters are shown as below. The model parameters of Wave-U-Net and IIC were updated by Adam [17] with $\beta_1 = 0.9$ and $\beta_2 = 0.9$. The learning rate is 0.0001 for Wave-U-Net but 0.01 for IIC. The number of epochs is 200 with random minibatches of size 28, which is the same value as the number of speakers. As mentioned above, the proposed architecture used the spectral normalization for stable training. The number of classes in IIC is less than the number of speakers as 5, since we aimed to induce the integration of the speaker independent features and the avoidance of the overfitting.

C. Objective Evaluation

To assess the quality of the reconstruction signals, we used six objective metrics including PESQ [18], CSIG, CBAK, COVL [19], Segmental SNR (SSNR), and STOI [20]. PESQ measures speech quality, which returns a score from 4.5 to -0.5, with higher scores indicating better quality. CSIG is a MOS predictor of speech distortion (from 1 to 5), CBAK is a MOS predictor of intrusiveness of background noise (from 1 to 5), and COVL is a MOS predictor of overall processed speech quality (from 1 to 5). STOI whose score ranges from 0 to 1 is a measure used to predict the intelligibility of speech.

For evaluation, we compared other five end-to-end models in addition to the above three models: SEGAN [8], HLGAN [10], WGAN-GP, SERGAN [11], SETransformer [6], Attention Wave-U-Net [5], ALRL [12], and Wave-U-Net with IIAD [21]. Table I shows the experimental results of the different models. This table summarizes that the proposed method provides best performance at PESQ and SSNR. As seen from this table, the Wave-U-Net with IIAD also demonstrates better results, where this method utilizes the self-attention based adversarial training instead of the normal adversarial training. Therefore, it is expected that the combination of the adversarial training with the proposed method would more boost the speech enhancement performance.

Figures 4 illustrates the resulting spectrograms of p257_070.wav which is a noisy female speech signal with low SNR included in the test set. Focusing on the silenced region enclosed in the white dashed box (1), the proposed method provides second-best performance. Meanwhile, focusing on the white dashed box (2), the proposed method can successes in keeping the weak speech component in comparison with the other methods.

V. Conclusion

In this paper, we proposed a new training architecture for the end-to-end speech enhancement network. The proposed learning architecture helps the encoder to understand the context invariant features. From the experimental results, we reveal the effectiveness of the proposed method.

TABLE I
OBJECTIVE EVALUATION RESULTS OF 10 DIFFERENT MODELS ON THE TEST SET OF VCTK DATASET

Model	PESQ	CSIG	CBAK	COVL	SSNR	STOI
Noisy	1.97	3.35	2.44	2.63	1.68	0.921
SEGAN [8]	2.16	3.43	2.94	2.80	7.73	-
HLGAN [10]	2.48	3.65	3.19	3.05	9.21	-
WGAN-GP [11]	2.54	-	-	-	-	0.937
SERGAN [11]	2.62	-	-	-	-	0.940
SETransformer [6]	2.62	-	-	-	-	0.93
Wave-U-Net [3]	2.40	3.52	3.24	2.96	9.97	-
AttWave-U-Net [5]	2.63	3.95	3.30	3.29	9.35	-
ALRL [12]	2.57	**4.79**	3.23	3.16	9.73	0.937
Wave-U-Net with IIAD [21]	2.80	4.11	**3.37**	**3.45**	10.0	**0.944**
Wave-U-Net with IIC	**2.80**	4.05	3.35	3.39	**10.3**	0.942

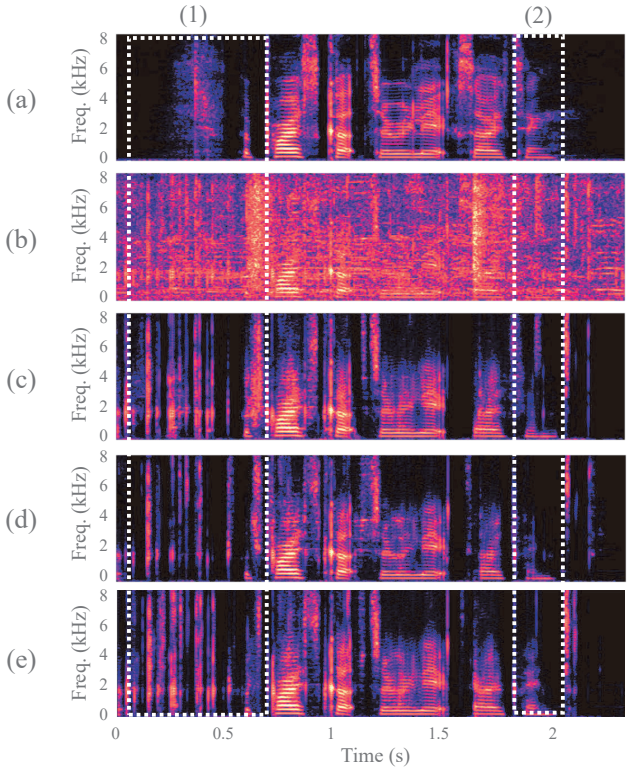

Fig. 4. Spectrograms of speech speech p257_070.wav in the test: (a) clean speech, (b) noisy speech, (c) enhanced by Wave-U-Net, (d) enhanced by Wave-U-Net with IIAD, (e) enhanced by Wave-U-Net with IIC (proposed).

REFERENCES

[1] H. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. of ICLR 2019*, New Orleans, USA, April 2019.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le, R. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. of LVA/ICA 2015*, Liberec, Czech Republic, Aug. 2015.

[3] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," in *Proc. of NIPS 2018*, Montreal, Canada, Nov. 2018.

[4] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 27, no.7, pp.1179–1188, July 2019.

[5] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for speech enhancement," in *Proc. of WASPAA 2019*, New York, USA, Dec. 2019.

[6] W. Yu, J. Zhou, H. B. Wang, and L. Tao, "SETransformer: speech enhancement transformer," *Cognitive Computation*, pp. 1–7, Oct. 2020.

[7] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. of ICML 2019*, California, USA, June 2019.

[8] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. of Interspeech 2017*, Stockholm, Sweden, Aug. 2017.

[9] S. Pascual, J. Serrà, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," in *Proc. of INTERSPEECH 2019*, Graz, Austria, Sep. 2019.

[10] F. Yang, Z. Wang, J. Li, R. Xia, and Y. Yan, "Improving generative adversarial networks for speech enhancement through regularization of latent representations," *Speech Communication*, vol. 118, pp. 1–9, April 2020.

[11] D. Baby and S. Verhulst, "SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. of ICCASP 2019*, Brighton, UK, 2019.

[12] Y. Qiu, and R. Wang, "Adversarial latent representation learning for speech enhancement," *Proc. of INTERSPEECH 2020*, Shanghai, China, Oct. 2020.

[13] X. Ji, J. F. Henriques, and S. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. of ICCV 2019*, Seoul, Korea, April 2019.

[14] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. of ICLR 2018*, Vancouver, Canada, April 2018.

[15] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. of Oriental COCOSDA*, Sep. 2013.

[16] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol.133, no.5, pp. 3591—3591, Sep. 2013.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc, of ICLR 2015*, Vancouver, Canada, May 2015.

[18] ITU-T Std., *P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, ITU-T, 2007.

[19] Y. Hu, and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol.16, no.1, pp.229–238, Dec. 2007.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. of ICASSP 2010*, Texas, USA, Mar. 2010.

[21] Y. Sugiura and T. Shimamura, "Adversarial training using inter/intra-attention architecture for dpeech enhancement network," in *Proc. of APSIPA 2020*, Auckland, New Zealand, Dec. 2020.