POSTFILTERING USING AN ADVERSARIAL DENOISING AUTOENCODER WITH NOISE-AWARE TRAINING

Naohiro Tawara[1], *Hikari Tanabe*[1], *Tetsunori Kobayashi*[1], *Masaru Fujieda*[2], *Kazuhiro Katagiri*[2], *Takashi Yazu*[2], *Tetsuji Ogawa*[1]

[1]Waseda University, Japan, [2]OKI Electric Industry Co., Ltd., Japan

ABSTRACT

An adversarial denoising autoencoder (ADAE) with noise-aware training is proposed and successfully applied to post-filtering for linear noise reduction. The ADAE is effective for attenuating interference sounds, however, it is difficult to learn to handle its various unexpected harmful effects (e.g., various types of noise) using a single network. Legacy speech enhancement was introduced as a pre-processor to make it possible to efficiently train the ADAEs by reducing the unexpected variabilities in the inputs to the ADAEs. Time-frequency masking performed well to suppress the variabilities, however, it induced unpleasant distortion, which is difficult for the ADAE to complement. In this paper, a minimum variance distortionless response (MVDR) beamformer, which can avoid troublesome non-linear distortions, is exploited as a preprocessor, and the MVDR outputs are used as the inputs to the ADAE-based post-filter. In addition, noise-dominant signals derived from the MVDR beamformer can improve the accuracy of the ADAE-based post-filter because the residual noise depends on the original noise signals. Experimental comparisons conducted using multichannel speech enhancement demonstrate that ADAE-based post-filtering yields significant improvements over the MVDR- and ADAE-based speech enhancement systems, and noise-aware training of ADAE works well.

Index Terms— Adversarial denoising autoencoder, minimum variance distortionless response, noise-aware training, speech enhancement

1. INTRODUCTION

In accordance with the remarkable progress of deep neural network (DNN) technologies, DNN-based speech enhancement systems have been frequently developed and performed well. Denoising autoencoders (DAEs) are effective for direct mapping from a noise-corrupted signal to a desired clean signal [1]. The DAEs have certain advantages in known environments, whereas their performance deteriorates when there is a mismatch between the training and testing environments [2]. One approach to address this problem is to collect as many types of noises as possible to generalize the model [3, 4]. Covering all types of noise, however, is infeasible, and more complex conditions (e.g., multiple noise sources) may appear in real-world environments [5]. Another approach, such as fine-tuning [6], is able to adapt networks to the testing environment; however, it requires a large amount of noise-corrupted and clean signal pairs.

Incorporating adversarial learning into DAEs is the latest approach aimed at increasing the accuracy of the DAEs [7, 8, 9]. The adversarial learning restricts the DAE from generating realistic signals, which are difficult to distinguish from actual clean signals. Adversarial learning has a stable performance over several noise types but it still needs fine-tuning with a set of supervised data in unknown environments [9].

To improve the robustness of the DAEs against the mismatch in noise conditions, time frequency (TF)-masking was applied and its outputs were taken as inputs to the ADAE-based speech enhancement systems [10]. This method successfully reduced the unexpected variability on the inputs to the ADAE and yielded a significant improvement over an ADAE without prefiltering. In addition, it was demonstrated that the quality of speech slightly improved when noise dominant signals obtained from TF-masking are fed as input to the ADAE. This result indicated the efficiency of noise-aware training [11] for ADAEs. Moreover, further improvements were observed when the oracle noise signals were used. This result indicated that the accuracy of noise estimation is crucial for noise-aware training. In this case, the non-linearity of TF-masking often over-subtracts target speech components from speech signals, making noise-aware training inefficient. The present study, therefore, introduces linear speech enhancement prior to the ADAE, instead of TF-masking. In particular, a minimum variance distortionless response (MVDR) beamformer is exploited because of its advantage that its output ideally has no non-linear distortion in speech components, while residual noise may remain in an enhanced signal.

The rest of this paper is organized as follows. Section 2 explains an ADAE-based speech enhancement with an MVDR beamformer. Section 3 demonstrates the effectiveness of the proposed system on multichannel speech signals

This work was supported by JSPS KAKENHI Grant Number 17K12718

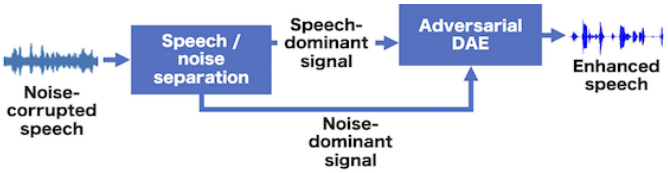

Fig. 1. Schematic diagram of proposed system.

with an interference source. Finally, Section 4 concludes the paper.

2. SPEECH ENHANCEMENT WITH ADAE AND MVDR BEAMFORMER

Figure 1 illustrates the schematic of the proposed system. First, an observed noise-corrupted signal is divided into noise- and speech-dominant signals by a speech / noise separation module. Then, the obtained speech-dominant signals are taken as the input to an ADAE to attenuate remaining noises. Here, the noise-dominant signal is feed into the ADAE as auxiliary information to make training the ADAE easier. The rest of this section gives a brief explanation of each module.

2.1. Speech and noise separation with MVDR

The role of this module is to divide a noise-corrupted signal into speech- and noise-dominant signals. In the current study, we introduce an MVDR beamformer to obtain speech- and noise-dominant signals.

The MVDR beamformer is a linear filter that enhances the target direction by directing null toward interfering directions. The weight matrix is selected to minimize the output power while maintaining the unity gain in the target direction. Thus, the optimization problem is described as

$$\min_{\mathbf{w}} \mathbf{w}^{\mathrm{H}}\mathbf{R}\mathbf{w} \quad \text{s.t.} \mathbf{w}^{\mathrm{H}}\mathbf{a} = 1, \tag{1}$$

where $\mathbf{w}$,$\mathbf{a}$, and $\mathbf{R}$ denotes a weight matrix, steering vector, and a spacial covariance matrix, respectively. The optimal weight matrix $\mathbf{w}^{\mathrm{MVDR}}$ that minimizes Eq. 1 is obtained as

$$\mathbf{w}^{\mathrm{MVDR}} = \frac{\mathbf{R}^{-1}\mathbf{a}}{\mathbf{a}^{\mathrm{H}}\mathbf{R}^{-1}\mathbf{a}}. \tag{2}$$

Using the obtained weight matrix, the target-dominant and noise-dominant signals are derived as

$$\mathbf{x}_{tar} = \mathbf{w}^{\mathrm{H}}_{\mathrm{MVDR}}\mathbf{x}, \tag{3}$$
$$\mathbf{x}_{int} = \mathbf{x} - \mathbf{x}_{tar}, \tag{4}$$

where $\mathbf{x}$, $\mathbf{x}_{tar}$, and $\mathbf{x}_{int}$ denote a observed, target-enhanced, and noise dominant signals, respectively.

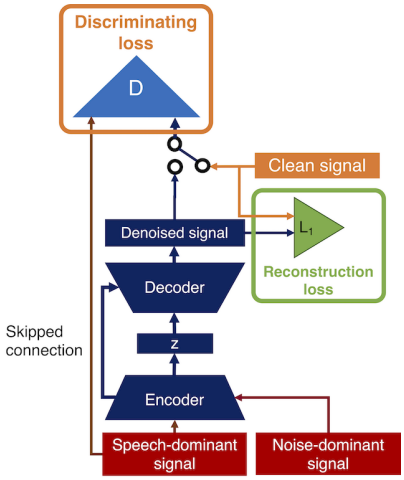

Fig. 2. Architecture of an adversarial denoising autoencoder (ADAE) with an auxiliary reference input.

2.2. Speech enhancement with ADAE

The speech- and noise-dominant signals derived from the MVDR beamformer are fed into an ADAE module to attenuate the remaining noises.

ADAE is a variant of DAEs that imposes a constraint on the DAE to generate realistic denoised signals. The ADAE is composed of a generator and a discriminator as described in Fig. 2. The role of the generator is to provide a mapping from a noise-corrupted signal to a denoised signal. The generator receives a one-second voice dominant signal (i.e., 16384 samples at 16 kHz) and generates the waveform of the same length as the input signal. The generator is trained to minimize $L1$ loss between clean and output signal from the generator. Here, the lower $L1$ can be achieved by concatenating noise-dominant signal with input of the ADAE. We call this noise-aware training for the ADAE. The output from the generator is taken into a discriminator. The role of the discriminator is to distinguish whether a given signal is a denoised-signal or an actual clean signal. The current paper introduces a conditional discriminator that takes enhanced and denoised signals. The configuration of the discriminator is the same as that of the encoder in the generator. The detailed structure of network is same as [10].

In the training phase, the generator and the discriminator are alternately optimized with the following adversarial procedure. First, fixing the parameter of generator G, the parameter of discriminator D is optimized by minimizing the following loss function:

$$\mathcal{L}_{\mathrm{cGAN}}(D) = \mathbb{E}_{\mathbf{x}_{tar},\mathbf{x}_c\sim p_{\mathrm{data}}(\mathbf{x}_{tar},\mathbf{x}_c)}[(1 - D(\mathbf{x}_{tar},\mathbf{x}_c))^2] + \mathbb{E}_{\mathbf{x}_{tar},\mathbf{x}_{int}\sim p_{\mathrm{data}}(\mathbf{x}_{tar},\mathbf{x}_{int})}[(D(\mathbf{x}_{tar},G(\mathbf{x}_{tar},\mathbf{x}_{int})))^2], \tag{5}$$

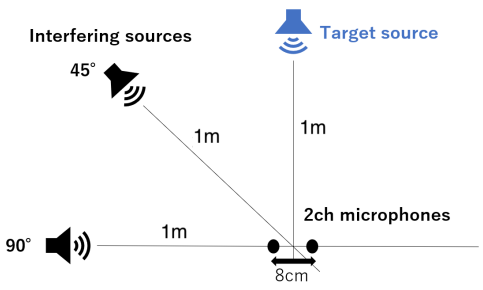

Fig. 3. Experimental environment with two microphones, a target source and interference sources.

DB id	noise type	use
09	exhibition hall (booth)	training
11	exhibition hall (aisle)	training
13	station (concourse)	training
14	station (aisle)	training
18	factory (machine)	training
20	factory (metal)	training
26	street	training
28	intersection	training
30	crowd	testing
47	elevator hall	testing

Table 1. Noise types which is selected from JEIDA and used for training and testing.

where $p_{\text{data}}(\mathbf{x}_{tar}, \mathbf{x}_c)$ denotes an empirical distribution over a pair of speech-dominant and clean signals $(\mathbf{x}_{tar}, \mathbf{x}_c)$; further, $p_{\text{data}}(\mathbf{x}_{tar}, \mathbf{x}_{int})$ denotes an empirical distribution over a pair of speech-dominant and noise-dominant signals $(\mathbf{x}_{tar}, \mathbf{x}_{int})$. By minimizing eq. (5), discriminator D attempts to discriminate whether the input is a clean or a denoised signal. Then, fixing the parameters of the discriminator, the generator G is optimized by minimizing the following loss function:

$$\mathcal{L}_{\text{cGAN}}(G) = \mathbb{E}_{\mathbf{x}_{tar},\mathbf{x}_{int},\mathbf{x}_c \sim p_{\text{data}}(\mathbf{x}_{tar},\mathbf{x}_{int},\mathbf{x}_c)}[1 - D(\mathbf{x}_{tar}, G(\mathbf{x}_{tar}, \mathbf{x}_{int})))^2 + \lambda ||\mathbf{x}_c - G(\mathbf{x}_{int}, \mathbf{x}_{tar})||_1], \quad (6)$$

where λ is a weight between the adversarial and the reconstruction losses, and set to 100 for the training. By minimizing eq. (5), the generator G attempts to generate denoised signals, which is difficult to distinguish from clean signals. After alternating the optimization of eqs. 5 and 6, the generator G generates denoised signals of high quality.

3. SPEECH ENHANCEMENT EXPERIMENT

Speech enhancement experiments were conducted to demonstrate the proposed ADAE with noise-aware training.

3.1. Experimental setup

Figure 3 shows the experimental environment. The target source was placed in front of two channel microphones. The distance between each source and the microphones was 1 m, and the distance among the microphones was 8 cm.

To simulate this condition, the dataset is made by the following procedure. The dry sources of target signals were 8000 utterances spoken by 124 speakers (62 male, 62 female) for training, and 500 utterances spoken by 14 speakers (7 male, 7 female) for testing. All utterances were selected from the Japanese newspaper article sentences read speech corpus (JNAS) [12], yielding approximately 50 different sentences for each speaker and noise condition. The dry sources of interference noises were eight types of noises for training and two types of noises for testing. All interference signals were selected from the JEIDA Noise Database [13]. Table 3 lists noise types used for training and testing. The target and interference signals were synthesized by convoluting their dry sources with the impulse responses measured in fixed positions. The impulse responses of each position were selected from the multi-channel impulse response database (MIRD) [14]. Reverberation time equals 160 ms. The convoluted speech and interference signals were mixed at five SNRs of -10, -5, 0, 5, and 10 dB. Note that the combinations of experimental conditions regarding speakers, utterances, and noise-types differed between training and testing.

As for the direction of interference sources, two different conditions were evaluated. In the single interference condition, an interference source was placed at 90 degrees to the target source in training time while it was placed at 45 or 90 degrees to the target source in testing time. In the multiple interferences condition, two interference sources were placed at 45 and 90 degrees to the target source in training and testing time. For all conditions, the target source was fixed in front of the microphones.

A signal distortion rate (SDR) between the estimated and the clean speeches is calculated using the BSS Eval toolbox [15] to evaluate the quality of speech. In order to measure the perceptual performance, a perceptual evaluation of speech quality (PESQ), based on the ITU standard P.862 [16], is also measured.

3.2. Evaluation items

The following four denoising systems were compared:

- **MVDR:** MVDR beamforming;
- **ADAE:** ADAE denoising;
- **MVDR+ADAE:** MVDR beamforming followed by post-filtering using ADAE; and
- **MVDR+ADAE-NAT:** MVDR beamforming followed by post-filtering using ADAE with noise-aware training.

Table 2. Speech enhancement performance of developed systems in the single interference condition. An interference source was placed at 90 degrees in training set, and placed at 45 or 90 degrees in test set. PESQ and SDR were averaged over 500 test utterances for each condition.

		training:90 [deg]										training:45 and 90 [deg]				
		test:90 [deg] (closed condition)					test:45 [deg] (open condition)					test: 45 and 90 [deg] (closed condition)				
Eval.	SNR [dB]	Obs.	MVDR	ADAE	MVDR +ADAE	MVDR +ADAE +NAT	Obs.	MVDR	ADAE	MVDR +ADAE	MVDR +ADAE +NAT	Obs.	MVDR	ADAE	MVDR +ADAE	MVDR +ADAE +NAT
SDR	-10	-10.00	-4.25	0.59	4.98	**9.54**	-10.00	-3.75	-0.34	2.70	**3.41**	-10.00	-4.00	1.04	6.14	**8.99**
	-5	-5.00	0.55	4.27	8.45	**12.23**	-5.00	0.98	3.61	6.62	**6.92**	-5.00	0.76	3.96	8.85	**10.98**
	0	0.00	4.98	7.25	11.36	**14.43**	0.00	5.33	6.72	9.89	**10.19**	0.00	5.15	6.50	10.91	**12.49**
	5	5.01	8.69	9.33	13.77	**16.09**	5.00	8.68	8.88	12.06	**12.72**	5.01	8.69	8.33	12.01	**13.21**
	10	10.1	11.40	10.40	15.54	**17.09**	10.1	10.8	10.10	13.09	**14.56**	10.01	11.11	9.73	12.58	**13.84**
PESQ	-10	1.07	1.26	0.74	2.00	**2.43**	1.23	1.46	0.73	**1.95**	1.67	0.99	1.36	0.71	1.99	**2.32**
	-5	1.28	1.64	1.27	2.36	**2.78**	1.43	1.80	1.23	**2.26**	2.02	1.10	1.72	1.19	2.40	**2.60**
	0	1.57	2.01	1.80	2.67	**3.07**	1.75	2.14	1.74	**2.57**	2.36	1.37	2.08	1.63	2.73	**2.82**
	5	1.85	2.36	2.21	2.90	**3.32**	2.06	2.42	2.16	**2.84**	2.68	1.70	2.39	2.06	2.96	**3.00**
	10	2.19	2.71	2.53	3.12	**3.51**	2.40	2.70	2.50	**3.09**	3.00	2.07	2.70	2.41	3.13	**3.14**

3.3. Experimental results

Tables 2 shows the speech enhancement performance in the single interference condition. From this result, we can see that **MVDR + ADAE** yielded a significant improvement over individual **MVDR** and **ADAE**. This result indicates that the residual noise in the voice-dominant signal obtained by the MVDR beamformer was successfully attenuated by the subsequent ADAE. Moreover, noise-aware training (i.e., **MVDR+ADAE+NAT**) yielded further improvements especially at low SNRs. This result demonstrates the effectiveness of the noise-aware training for ADAE. However, comparing the results of 45 and 90 degrees cases, the improvement of SDR obtained by noise-aware training in the former case was relatively small than that of the latter case. This was because the noise-dominant signal of test data significantly differed from that of training data due to the difference of position of the intereference source. This problem was solved by addiding both conditions to training data.

Table 3 shows the speech enhancement performance in multiple interference conditions (i.e., two interference sources were placed at 45 and 90 degrees to a target). In this condition, the performance of **MVDR** significantly deteriorated especially measured in SNR. This could be because the number of interfering sources was larger than the that of microphones, and relatively a large noise remained in the voice-dominant signal obtained by MVDR. In this case, the proposed **MVDR+ADAE** and **MVDR+ADAE+NAT** still yielded significant improvements over **Obs** and **MVDR**. This result indicates that ADAE performed well even if the noise could not be attenuated completely by MVDR. Moreover, the improvement of performance obtained by noise-aware training in the multiple interference condition was relatively large compared with that in the single interference condition. This was because noise information especially effective to attenuate the residual noise in voice-dominant signal.

Table 3. Speech enhancement performance of developed systems in the multiple interference condition. Two interference sources were placed at 45 and 90 degrees to the target source.

Eval. metric	SNR	Obs.	MVDR	ADAE	MVDR +ADAE	MVDR +ADAE +NAT
SDR	-10dB	-10.00	-4.47	-4.00	1.81	**11.20**
	-5dB	-5.00	0.37	0.71	5.62	**13.65**
	0dB	0.00	4.74	4.40	8.39	**15.52**
	5dB	5.01	8.45	7.99	9.94	**16.66**
	10dB	10.02	11.11	10.06	10.56	**17.34**
PESQ	-10dB	1.21	1.51	1.16	1.93	**2.59**
	-5dB	1.37	1.84	1.67	2.27	**2.96**
	0dB	1.66	2.18	2.15	2.59	**3.24**
	5dB	1.96	2.48	2.58	2.84	**3.44**
	10dB	2.31	2.79	2.92	3.06	**3.58**

4. CONCLUSION

An ADAE with noise-aware training was proposed and applied to post-filtering for linear noise reduction. Specifically, speech-dominant and noise-dominant signals derived by MVDR beamformer were taken as the inputs to the ADAE-based post-filter. Experiments using multichannel speech enhancement were conducted, demonstrating that the proposed approach yielded significant improvements over the MVDR- and ADAE-based speech enhancement systems.

From the experimental result, it was shown that a key to successful noise aware training for ADAE was the accuracy of noise estimation in pre-filtering. We, therefore, plan to introduce more sophisticated source separation algorithms as pre-filtering. The direction of the target source was given in the current experiment. We also plan to apply the proposed framework to a blind source condition.

5. REFERENCES

[1] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *INTERSPEECH*, 2013, pp. 436–440.

[2] Liao Chien-Feng, Tsao Yu, Lee Hung-Yi, and Wang Hsin-Min, "Noise adaptive speech enhancement using domain adversarial training," *arXiv preprint arXiv:1807.07501*, 2018.

[3] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui 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, 2015.

[4] Jitong Chen, Yuxuan Wang, Sarah E Yoho, DeLiang Wang, and Eric W Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[5] Anurag Kumar and Dinei Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," in *INTERSPEECH*, 2016, pp. 3738–3742.

[6] Xue Feng, Yaodong Zhang, and James Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *ICASSP*. IEEE, 2014, pp. 1759–1763.

[7] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[8] Chris Donahue, Bo Li, and Rohit Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," *arXiv preprint arXiv:1711.05747*, 2017.

[9] Santiago Pascual, Maruchan Park, Joan Serrà, Antonio Bonafonte, and Kang-Hun Ahn, "Language and noise transfer in speech enhancement generative adversarial network," in *ICASSP*, 2018, pp. 5019–5023.

[10] Naohiro Tawara, Tetsunori Kobayashi, Masaru Fujieda, Kazuhiro Katagiri, Takashi Yazu, and Tetsuji Ogawa, "Adversarial autoencoder for reducing nonlinear distortion," in *APSIPA*, 2018, pp. 1669–1673.

[11] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *INTSPEECH*, 2014, pp. 2670–2674.

[12] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano, and Shuichi Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[13] Shuichi Itahashi, "A noise database and japanese common speech data corpus," *Journal of the Acoustical Society of Japan*, vol. 47, no. 12, pp. 951–953, 1991, in Japanese.

[14] "Multi-channel impulse response database," https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/.

[15] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[16] 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.