Domain Adversarial Training for Speech Enhancement

Nana Hou*, Chenglin Xu*†, Eng Siong Chng*†, Haizhou Li‡
* [1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
† Temasek Laboratories, Nanyang Technological University, Singapore
‡ Department of Electrical and Computer Engineering, National University of Singapore, Singapore
E-mail: nana001@e.ntu.edu.sg, {xuchenglin, aseschng}@ntu.edu.sg, haizhou.li@nus.edu.sg

***Abstract*—The performance of deep learning approaches to speech enhancement degrades significantly in face of mismatch between training and testing. In this paper, we propose a domain adversarial training technique for unsupervised domain transfer, that 1) overcomes domain mismatch, and 2) provides a solution to the scenario where we only have noisy speech data, and we don't have clean-noisy parallel data in the new domain. Specifically, our method includes two parts that are jointly trained, 1) an enhancement net to map noisy speech to clean speech by indirectly estimating a mask with a spectrum approximation loss, and 2) a domain predictor to distinguish between domains. As the proposed approach is able to adapt to a new domain only with noisy speech data in target domain, we call it an unsupervised learning technique. Experiments suggest that our approach delivers voice quality comparable with other supervised learning techniques that require clean-noisy parallel data.**

I. Introduction

Speech enhancement technique aims to reduce additive noise to speech signals in order to improve speech intelligibility and quality. It can be used as a pre-processing module in automatic speech recognition (ASR), speaker identification systems, and hearing aids design [1], [2], [3]. Among the conventional solutions are spectral subtraction [4], Wiener filtering [5], minimum-mean-square-error (MMSE)-based spectral amplitude estimator [6], and subspace algorithms [7]. In recent years, various deep learning frameworks were studied that benefit from their ability to learn from data distribution, such as deep neural network (DNN) method [8], recurrent neural network (RNN) method [9], and denoising autoencoder (DAE) method [10].

For many machine learning tasks, such as speech enhancement, the training and testing data are usually assumed to have the same probability distribution. However, practical scenarios often fail to meet this assumption. To address such training-testing mismatch, a popular technique is to adapt a model trained under one training condition, the source domain, towards another testing condition, the target domain. For example, the study in [11] suggests adapting the last layers of pre-trained speech enhancement generative adversarial network (SEGAN) with the dataset of new language and noise to reduce the mismatch between different languages and noise, but this technique asks for clean-noisy parallel speech data that are not always available in practice.

Recently, another method widely-used in image processing is domain adaptation [12]. This technique attempts to adapt features with a domain discriminator structure via domain adversarial training (DAT) in face of test data in the new domain. In speech and speaker recognition, it was used to adapt acoustic models or produce speaker-invariant features to overcome the mismatch between training and testing [13], [14], [15].

Inspired by previous study in DAT, we propose an unsupervised domain transfer approach by adapting the enhancement net without the need of clean-noisy parallel speech data in the new domain. In our scenarios, the training data in source domain consist of clean-noisy speech pairs, but those in the target domain only consist of noisy speech. We propose a complete pipeline to overcome the mismatch across domain, that we call domain adversarial training approach to speech enhancement, or SE-DAT. It has the following main advantages:

- SE-DAT can be adapted to target domain with only noisy speech data, without the need of clean-noisy speech pairs.
- SE-DAT architecture is concise and requires no deep structure like feature extractor in the work [16] to learn adapted feature representation.
- We also introduce the dynamic features [17], [18] into SE-DAT, which takes the temporal context of features into consideration to ensure the continuity of the enhanced speech [19], [20].

The rest of this paper is organized as follows. In section 2, we describe the proposed SE-DAT technique. In section 3, experimental settings and results are presented. Section 4 concludes the study.

II. Domain adversarial training for speech enhancement (SE-DAT)

With SE-DAT, we assume the model learns the mapping between noisy speech sample $x \in X$ and its corresponding clean sample $y \in Y$. y and x form a clean-noisy speech pair. We also assume that the noisy sample x and its clean sample y belong to a distribution $\mathcal{S}(x, y)$, also called source domain. Suppose that we now have some noisy speech samples in the new domain without the corresponding clean speech samples, we hope to adapt the model so that it works both in the source domain and the new domain. The unpaired dataset is assumed to belong to the other distribution $\mathcal{T}(x, y)$, also called target domain. Finally, we assign the binary domain label, $d \in [0, 1]$, to each noisy sample at the training stage

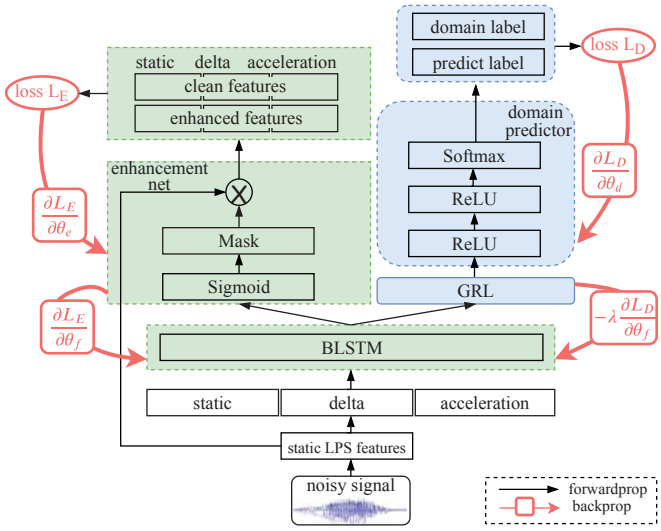

Fig. 1. SE-DAT includes two parts, an enhancement net E (green) that generates the enhanced speech and a domain predictor D (blue) that distinguishes between domains the input comes from. The two parts are jointly trained to minimize the loss of the enhancement net L_E and to maximize the loss of the domain predictor L_D at the same time through a GRL.

to indicate which domain the noisy samples come from. We illustrate the proposed technique in Figure 1.

A. Dynamic features

In the training process, the training speech data from both the source domain and the target domain are extracted with a shifting window into static log-power-spectrum (LPS) features, that is also called the static feature. A frame of speech is represented by a vector of static features. A limitation of using only LPS features is that each frame is represented independently and we cannot guarantee that the produced frame sequence is smooth and sounds natural. Hence, we introduce the dynamic features [17], [18] that take the temporal context of features into consideration. In this work, the dynamic features are the derivatives of the LPS features, including delta features (first-order time derivatives) and acceleration features (second-order time derivatives). We can approximate the delta features and acceleration features as follows:

$$f_D(t) = \frac{\sum_{l=1}^{L} l * (f_S(t+l) - f_S(t-l))}{\sum_{l=1}^{L} 2l^2} \tag{1}$$

where $f_S(t)$ and $f_D(t)$ are the static feature and the delta feature respectively at frame t. L is the order of computing the derivatives and is set to 2 in this study. The acceleration features denoted as f_A are obtained by applying equation (1) on the delta features f_D. The original LPS feature, delta feature, and acceleration feature jointly form a new feature $F = [f_S, f_D, f_A]$ for a speech frame.

B. The enhancement net

The enhancement net E aims to map input noisy speech to clean speech by estimating a mask, where one bidirectional long short-term memory (BLSTM) layer produces the adapted representations v for input feature frame $F = [f_S, f_D, f_A]$. Such representations v is used by two nets: the enhancement net E and the domain predictor D. In the enhancement net E, suppose that the representations v_i of input sample x_i arrives from the source domain, we take the dot product between the static LPS feature f_S and its estimated mask. We then take the enhanced static feature $\hat{y}_S$ to obtain its dynamic features $\hat{y}_D$ and $\hat{y}_A$ according to equation (1). Finally, we compute the spectrum approximation loss between the enhanced feature frame $\hat{y} = [\hat{y}_S, \hat{y}_D, \hat{y}_A]$ and the corresponding clean feature frame.

The spectrum approximation loss for enhancement net E [18] is given as follows:

$$\begin{aligned} L_E(\theta_f, \theta_e) = \|\hat{y}_S - y_S\|_F^2 + w_D\|\hat{y}_D - y_D\|_F^2 \\ + w_A\|\hat{y}_A - y_A\|_F^2 \end{aligned} \tag{2}$$

where θ_f and θ_e are parameters of the BLSTM layer and the rest enhancement net respectively. $\|\cdot\|_F$ is the Frobenius norm. w_D and w_A are the weights of cost contributed by the delta and acceleration features. Besides, if the representations v_i for input sample x_i arrives from the target domain (without the paired clean-noisy speech), we don't calculate the loss L_E for this input noisy sample due to no clean reference.

C. The domain predictor

SE-DAT aims to overcomes the mismatch between the source and target domain without the need of clean-noisy parallel data, which is achieved by the domain predictor. In domain predictor D, we set the i-th domain label as d_i for the representation v_i to indicate where v_i comes from. If v_i comes from the source domain, d_i is set to 0 (if $v_i \sim \mathcal{S}(v)$, set $d_i = 0$), otherwise d_i is set to 1 (if $v_i \sim \mathcal{T}(v)$, set $d_i = 1$). The cross-entropy loss for domain predictor D is defined as:

$$\begin{aligned} L_D(\theta_f, \theta_d) = -\frac{1}{N}\sum_{i=1}^{N}[d_i \log P(v_i \in \mathcal{S}(v)) \\ + (1 - d_i) \log P(v_i \in \mathcal{T}(v))] \end{aligned} \tag{3}$$

where θ_f and θ_d are parameters of the BLSTM layer and the domain predictor D respectively. N is the number of input training samples.

We now jointly train the two parts: the enhancement net E and the domain predictor D for 1) seeking the parameters θ_f to maximize the loss of the domain predictor D, 2) simultaneously seeking the parameters θ_d to minimize the loss of domain predictor D, and 3) seeking θ_e to minimize the loss of the enhancement net E. Such optimization can be achieved by the gradient reversal layer (GRL). The role of GRL is an identity transform during the forward propagation. During the backpropagation, the GRL multiplies the gradient from the domain predictor D by $-\lambda$ and then passes it to the BLSTM layer. The whole cost function of the SE-DAT is formulated below:

$$L(\theta_f, \theta_e, \theta_d) = L_E(\theta_f, \theta_e) - \lambda L_D(\theta_f, \theta_d) \tag{4}$$

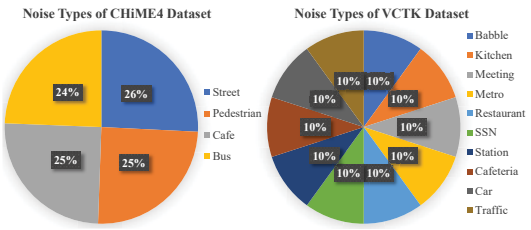

Fig. 2. The statistics of noise types of CHiME4 dataset (left) and VCTK dataset (right).

where λ is the gradient reversal coefficient that controls the trade-off between two objectives during training. λ is defined as:

$$\lambda = \frac{2}{1 + \exp\left(-10 * \frac{j+k*J}{K*J}\right)} - 1 \tag{5}$$

where j denotes the index of current batch and J is the total number of batches. k presents the index of current epoch and K is the total number of epochs. In this way, standard stochastic gradient solvers (SGD) can be applied for the search of the best parameters $(\theta_f, \theta_e, \theta_d)$ as follows:

$$\begin{aligned} \theta_f &\leftarrow \theta_f - \mu\left(\frac{\partial L_E}{\partial \theta_f} - \lambda \frac{\partial L_D}{\partial \theta_f}\right) \\ \theta_e &\leftarrow \theta_e - \mu \frac{\partial L_E}{\partial \theta_e} \\ \theta_d &\leftarrow \theta_d - \mu \frac{\partial L_D}{\partial \theta_d} \end{aligned} \tag{6}$$

where μ is the learning rate.

At the testing stage, only the noisy speech is enhanced by the enhancement net E, while the domain predictor D is discarded.

III. Experiments

We would like to validate the proposed SE-DAT by adapting the enhancement net from source domain to target domain.

A. Database

To evaluate the effectiveness of SE-DAT, We resort to two corpora: one is CHiME-4 dataset [21]; the other is the dataset released by Cassia Valentini-Botinhao [22], which is same in SEGAN [23] and Wave-U-Net [24], referred to as VCTK dataset hereafter. We use CHiME-4 dataset as source domain data and VCTK dataset as target domain data in order to validate SE-DAT in reducing the mismatch across domains.

1) Source domain: CHiME-4 dataset: In the CHiME-4 dataset, the simulated data are generated by artificially mixing clean speech data with noisy backgrounds of four types, i.e. cafe, bus, street, and pedestrian area. We use the simulated training set (7,128 utterances) and simulated development set (1,600 utterances) as source domain data.

TABLE I
Comparisons with SE-DAT-0 and SE-DAT in terms of the PESQ, CSIG, CBAK, COVL and SSNR scores on VCTK test set. "Zero-effort" means that we use the untreated noisy speech of VCTK test set. Higher scores are better for all metrics.

Method	PESQ	CSIG	CBAK	COVL	SSNR
Zero-effort	1.97	3.35	2.44	2.63	1.68
SE-DAT-0	2.12	3.38	2.46	2.66	1.76
SE-DAT	**2.26**	**3.72**	**2.77**	**2.98**	**4.11**

2) Target domain: VCTK dataset: In the VCTK dataset, a total of 40 different conditions are considered [22]: 10 types of noise (2 artificial and 8 from the Demand database [25]) with 4 signal-to-noise ratio (SNR) each (15, 10, 5, and 0 dB). There are 14 male and 14 female training speakers. We use the VCTK dataset at a ratio of 9:1 as the training set (1,0415 utterances) and development set (1,157 utterances). With respect to test set, a total of 20 different conditions are considered [22]: 5 types of noise (all from the Demand database) with 4 SNR each (17.5, 12.5, 7.5, and 2.5 dB). There are 1 male and 1 female test speakers.

As shown in Figure 2, the conditions of the CHiME-4 dataset and the VCTK dataset are different in noise types. Besides, the training speakers and SNR are totally different, which fits our purpose: evaluating the effectiveness of SE-DAT in reducing the mismatch between two different domains. To show the effectiveness of DAT, we use VCTK dataset in the noisy target domain without the need of its corresponding clean speech.

B. Experiment setup

The two datasets are sampled at 16 kHz sampling rate and 16 bits/sample. We applied 512-point STFT to extract LPS, the delta features and acceleration features. One BLSTM layer is used with 512 units, which is followed by one feed-forward layer of 257 logistic units with sigmoid activation in enhancement net E. The domain predictor D consists of three feed-forward layers with two ReLU activations and one softmax activation. The learning rate μ is set to 0.001, and the batch size is 32. The weights for delta features w_D and for acceleration features w_A are empirically set to 4.5 and 10.0 respectively [19], [27]. Early stop and learning rate adjustment strategy are also adopted in the experiments. The start halving improvement, halving factor and the end halving improvement are 0.003, 0.5 and 0.001 respectively.

To evaluate SE-DAT, two models were trained using aforementioned datasets:

- SE-DAT-0: The model, with λ set to 0, is trained only on source domain CHiME-4 data (with clean-noisy parallel data) and tested on target domain VCTK test set. This model serves as the reference baseline for testing, where model adaptation is no attempted.
- SE-DAT: SE-DAT is trained by both source domain CHiME-4 data (with clean-noisy parallel data) and target domain VCTK data (noisy speech without clean speech counterpart) to verify the effectiveness, which attempts

TABLE II
TRAINING DETAILS OF DIFFERENT METHODS ON VCTK DATASET.

Method	Training set	Clean for supervision	Feature domain	Test set
SEGAN [11]	VCTK set	Yes	time domain	VCTK test set
CNN-GAN [26]	VCTK set	Yes	frequency domain	VCTK test set
Wave-U-NET [24]	VCTK set	Yes	time domain	VCTK test set
SE-DAT	CHiME4 simu set for source domain VCTK set for target domain	Yes for source domain No for target domain	frequency domain	VCTK test set

TABLE III
COMPARISONS WITH DIFFERENT METHODS IN TERMS OF THE PESQ, CSIG, CBAK, COVL AND SSNR SCORES ON VCTK TEST SET. "ZERO-EFFORT" MEANS THAT WE USE THE UNTREATED NOISY SPEECH OF VCTK TEST SET.

Method	Training	PESQ	CSIG	CBAK	COVL	SSNR
Zero-effort	–	1.97	3.35	2.44	2.63	1.68
SEGAN [11]	supervised	2.16	3.48	2.94	2.80	7.73
CNN-SEGAN [26]	supervised	2.34	3.55	2.95	2.92	–
Wave-U-Net [24]	supervised	**2.40**	3.52	**3.24**	2.96	**9.97**
SE-DAT	unsupervised	2.26	**3.72**	2.77	**2.98**	4.11

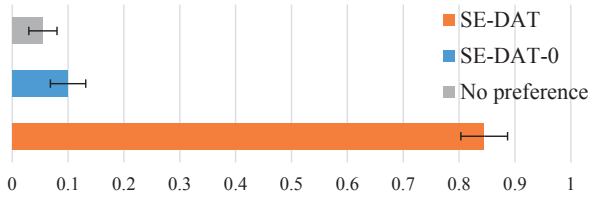

Fig. 3. Results of the quality preference test with 95% confidence intervals for different methods.

to use the noisy target domain data to overcome the mismatch across domains.

C. Experimental results

1) Objective evaluation: To evaluate the quality of the enhanced speech, we compute the following objective measures.

- PESQ: Perceptual evaluation of speech quality, using the wide-band version recommended in ITU-T P.862.2 [28] (from -0.5 to 4.5).
- CSIG: Mean opinion score (MOS) prediction of the signal distortion attending only to the speech signal [29] (from 1 to 5).
- CBAK: MOS prediction of the intrusiveness of background noise [29] (from 1 to 5).
- COVL: MOS prediction of the overall effect [29] (from 1 to 5).
- SSNR: Segmental SNR [30] (from 0 to ∞).

All metrics compare the enhanced signal with the clean reference on the VCTK test set (824 utterances), using the tookit in [31]. As shown in Table I, we note that SE-DAT-0 trained on CHiME4 simulated training set alone does not perform well on the VCTK test set in the new domain. With the domain mismatch, we observe that the performance of SE-DAT-0 is almost same as the noisy speech without enhancement. By applying DAT, we observe that the proposed SE-DAT approach drastically improves all the performance when we train only on noisy target domain data.

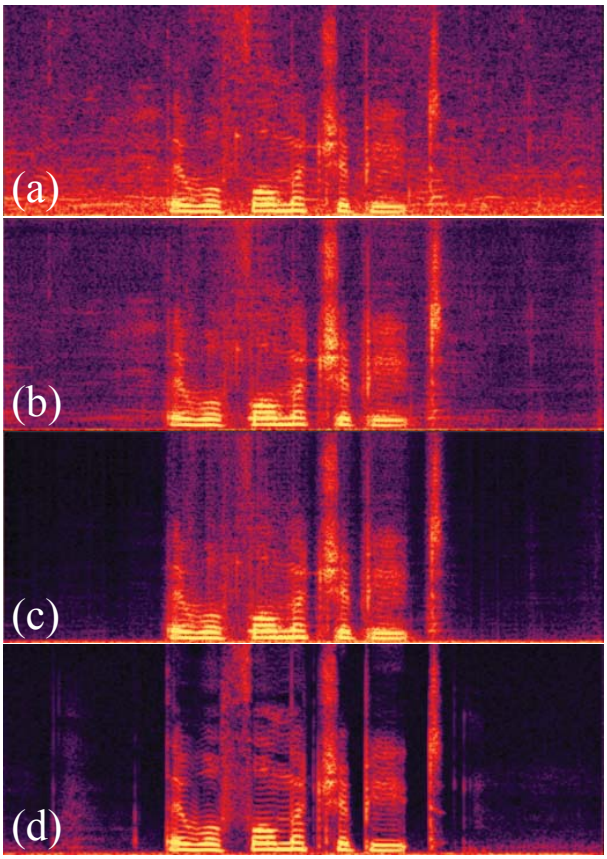

Fig. 4. Comparisons of spectrums. (a) denotes the spectrum of the noisy speech and (b) is the spectrum of the corresponding enhanced speech by SE-DAT-0. (c) represents the spectrum of the corresponding enhanced speech by SE-DAT and (d) is the spectrum of the corresponding clean speech.

To further showcase the ability of SE-DAT, a speech utterance (spectrum) from VCTK test set is shown in Figure 4. The original noisy speech is shown in (a) and the corresponding clean speech is in (d). We observe in (b) that SE-DAT-0 cannot reduce the noise effectively in unseen speech of the new domain and most of the noise components still remain. On the contrast, despite training without the clean speech utterances in the new domain, the proposed SE-DAT still can significantly remove the noise components as shown in (c).

2) Subjective evaluation: The AB preference test was conducted to assess the subjective perceptual quality of the en-

hanced speech. In the AB preference test, each paired samples A and B were randomly selected from the proposed SE-DAT model and the SE-DAT-0 model. 10 subjects participated in the preference test. Each listener was asked to choose the sample with better quality from each pair. The subjective results of quality preference test are presented in Figure 3. The results suggest that the speech quality of SE-DAT significantly outperforms that of SE-DAT-0.

3) Comparisons with other methods: We further compare the proposed SE-DAT with some recent methods conducted on VCTK dataset although this comparison is not fair. As shown in Table II, the methods like SEGAN, CNN-GAN, and Wave-U-Net are all trained using clean-noisy paired VCTK speech to supervise the learning of the network without mismatch problem. We are glad to see that the proposed SE-DAT transferred the knowledge from the source domain to the target domain without supervision information from clean speech in the target domain as reference during training. In addition, SEGAN and CNN-GAN directly extract the features from time domain, which means there is no phase problem. CNN-GAN and the proposed SE-DAT use the spectrum features through STFT and re-use the phase of noisy speech. As shown in Table III, despite the unfair conditions, the proposed SE-DAT still performs better in CSIG and COVL, which means it produces less speech distortion and achieves a better overall quality.

IV. Conclusions

In this paper, we propose a domain adversarial training technique to speech enhancement (SE-DAT) to overcome the mismatch across domains and provide a solution for speech denoising to the scenario where we don't have clean-noisy parallel data in the new domain. SE-DAT achieves significant improvement on VCTK dataset compared with the model where no effort is made to overcome the mismatch. SE-DAT also delivers voice quality comparable with other supervised learning techniques that require clean-noisy parallel data. In the future, we will explore its ability for speech recognition as a pre-processing module.

Acknowledgement

This research is supported by Temasek Laboratories@NTU, Nanyang Technological University, Singapore.

References

[1] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4041–4044.

[2] J. Ortega-García and J. González-Rodríguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 929–932.

[3] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *The Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1001–1004, 2005.

[4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, vol. 4. IEEE, 1979, pp. 208–211.

[5] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[7] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on speech and audio processing*, vol. 3, no. 4, pp. 251–266, 1995.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. 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.

[9] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[10] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[11] S. Pascual, M. Park, J. Serrà, A. Bonafonte, and K.-H. Ahn, "Language and noise transfer in speech enhancement generative adversarial network," *arXiv preprint arXiv:1712.06340*, 2017.

[12] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[13] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[14] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," *arXiv preprint arXiv:1806.02786*, 2018.

[15] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," 2018.

[16] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," *arXiv preprint arXiv:1807.07501*, 2018.

[17] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[18] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 4, 2016.

[19] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6–10.

[20] C. Xu, W. Rao, E. S. Chng, and H. Li, "A shifted delta coefficient objective for monaural speech separation using multi-task learning," in *Proceedings of Interspeech*, 2018, pp. 3479–3483.

[21] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[22] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks." in *Interspeech*, 2016, pp. 352–356.

[23] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[24] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[26] N. Shah, H. A. Patil, and M. H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial net-

work," in *Proceedings, APSIPA Annual Summit and Conference*, vol. 2018, 2018, pp. 12–15.

[27] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[28] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, 2005.

[29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[30] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.

[31] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.