CycleGAN with Dual Adversarial Loss for Bone-Conducted Speech Enhancement

Qing Pan*, Yan Pan*, Jian Zhou*, Huabin Wang*, Liang Tao*, and Hon Keung Kwan†

*Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei, Anhui, China 230601
†Department of Electrical and Computer Engineering, University of Windsor, Windsor, Ontario, Canada N9B 3P4
{e19301131@stu.ahu.edu.cn, e21301331@stu.ahu.edu.cn, jzhou@ahu.edu.cn, wanghuabin@ahu.edu.cn, taoliang@ahu.edu.cn kwan1@uwindsor.ca}

Abstract—Compared with air-conducted speech, bone-conducted speech has the unique advantage of shielding background noise. Enhancement of bone-conducted speech helps to improve its quality and intelligibility. In this paper, a novel CycleGAN with dual adversarial loss (CycleGAN-DAL) is proposed for bone-conducted speech enhancement. The proposed method uses the adversarial loss and the cycle-consistent loss simultaneously to learn forward and cyclic mapping, in which the adversarial loss is replaced with the classification adversarial loss and the defect adversarial loss to consolidate the forward mapping. Compared with conventional baseline methods, it can learn feature mapping between bone-conducted speech and target speech without additional air-conducted speech assistance. Moreover, the proposed method also avoids the over-smooth problem which is occurred commonly in conventional statistical-based models. Experimental results show that the proposed method outperforms baseline methods such as CycleGAN, GMM, and BLSTM.

Index Terms—Bone-conducted speech enhancement, dual adversarial loss, Parallel CycleGAN

I. INTRODUCTION

Speech is the most convenient way of communication. When a person is vocalizing, the vocal cord vibration signals are radiated through the acoustic cavity and the lips, producing an air-conducted sound. Meanwhile, this voicing behavior is also accompanied by skull and throat vibrations. The bone conduction microphone can collect such weak signals from the skull and the larynx to obtain bone-conducted (BC) speech.

Compared with the air-conducted (AC) speech, the BC speech is insusceptible to adverse environmental noise due to its distinctive generation mechanism. The BC speech has a wide range of applications in military and smart civilian fields [1]. However, because of the low-pass characteristics of the skull, the voice signal collected by the bone conduction microphone loses many high-frequency components, resulting in the BC speech sounds dull, and its intelligibility being low.

To improve the speech quality and intelligibility of the BC speech, early methods usually adopt the source-filter model to decompose the BC speech into a spectral envelope and excitation features and mainly focus on the conversion of the low-dimensional spectral envelope of the BC speech to that of the AC speech [2]–[4].

Recently, neural networks have been used for nonlinear transformation between spectral envelopes [5], [6]. Zheng et al. proposed a BC speech enhancement method based on the Bi-directional Long Short-Term Memory (BLSTM) type of recurrent neural networks and attention mechanism (AB- BLSTM). Moreover, fusion features have also been studied in BC speech enhancement [7]. However, BC speech enhancement fused with AC speech features is less effective in a strong noise environment. Hence, enhancing BC speech independently has important theoretical significance and practical application. The generative adversarial network (GAN) has attracted much attention because of its ability in simulating data distributions. GANs have been applied in speech processing tasks such as speech bandwidth expansion (BWE) and voice conversion (VC). Kaneko et al. applied GAN to sequence to sequence (Seq2Seq) VC [8]. Experimental results show that GAN-based training is better than traditional mean-square error-based training.

In this article, we propose a novel CycleGAN with dual adversarial loss (CycleGAN-DAL) model to conduct BC speech enhancement. The proposed model does not need to explicitly model the relationship between the missing high-frequency components of BC speech and the low-frequency components of AC speech. Moreover, the discriminator is trained by using the dual countermeasure loss, and the countermeasure loss is divided into classification countermeasure loss and defect countermeasure loss. This is a blind enhancement method that does not need prior information such as the high-frequency components of BC speech and the low-frequency components of AC speech.

II. THE PROPOSED CYCLEGAN-DAL ARCHITECTURE FOR BC SPEECH ENHANCEMENT

A. The Proposed CycleGAN-DAL Model

The proposed CycleGAN-DAL model is shown in Fig. 1. Two different adversarial losses for the forward-inverse mapping, i.e., a classification adversarial loss and a defect adversarial loss are proposed.

The classification adversarial loss function $\mathcal{L}_{adv}^C(G_{X\rightarrow Y}, D^C_{Y})$ is responsible for classifying the generated speech features $G_{X\rightarrow Y}(x)$ as fake and the ground truth AC speech features $y$ as real, which is defined by

$$\mathcal{L}_{adv}^C(G_{X\rightarrow Y}, D^C_{Y}) = E[\log D^C_{Y}(y)] + E[\log(1 - D^C_{Y}(G_{X\rightarrow Y}(x)))] \tag{1}$$
where $D^C_Y$ denotes the classification discriminator.

To constrain the generator to produce a more accurate AC speech feature, the defect adversarial loss function $L_{adv}^D(G_X \rightarrow Y, D^D_Y)$ is adopted to judge whether the generated spectral feature $G_X \rightarrow Y(\hat{x})$ of the BC speech is damaged compared to that of the real AC speech spectrum $y$, which is defined by

$$L_{adv}^D(G_X \rightarrow Y, D^D_Y) = E[\log(1 - D^D_Y(G_X \rightarrow Y(\hat{x})))] \quad (2)$$

where $D^D_Y$ denotes the defect discriminator.

We employ a cycle-consistent loss to keep the input and output features consistent in the aspect of semantics, the cycle-consistent loss is defined by

$$L_{cyc}(G_Y \rightarrow X, G_X \rightarrow Y) = E[||G_Y \rightarrow X(G_X \rightarrow Y(\hat{x})) - x||_1] \quad (3)$$

where $|| \cdot ||_1$ denotes the $l_1$ norm. The proposed cycle-consistent loss encourages $G_X \rightarrow Y$ and $G_Y \rightarrow X$ to find $(x, y)$ pairs with the same contextual information.

To embed semantic information of the BC speech in the mapping, an identity mapping loss $L_{id}$ is introduced in the CycleGAN-DAL model, which is defined by

$$L_{id}(G_X \rightarrow Y) = E[||G_X \rightarrow Y(y) - y||_1] \quad (4)$$

In summary, the objective loss function $L_{dual}$ of the proposed CycleGAN-DAL is as follows

$$L_{dual} = L_{adv}^C(G_X \rightarrow Y, D^C_Y) + L_{adv}^D(G_X \rightarrow Y, D^D_Y) + \lambda_{cyc}L_{cyc}(G_X \rightarrow Y, G_Y \rightarrow X) + \lambda_{id}L_{id}(G_X \rightarrow Y) \quad (5)$$

B. The Structure Of The CycleGAN-DAL Model

1) Generator: The generator of the proposed CycleGAN-DAL is shown in Fig. 2 and the parameter settings are shown in Table I. In Fig. 2, $h$, $w$, and $c$ represent the height(i.e., batch size), width(i.e., number of frames), and number of channels(i.e., dimension of MCEPs), respectively. $k$, $c$, and $s$ represent kernel size, the number of channels, and stride size, respectively. As can be seen in Fig. 2, the generator consists of an encoder, a converter, and a decoder.

The encoder, which is used to capture the latent representation of the input BC speech features, comprises two stacked convolutional network blocks, each of which is composed of
a one-dimensional convolutional layer, an instance normalization (IN) layer, and an activation layer.

The converter consists of 6 residual network blocks, each of which comprises two stacked convolutional networks. Note that the first convolutional network consists of a convolutional layer, an IN layer, and a GLU activation layer, while the second convolutional network consists of a convolutional layer, an IN layer, and a skip connection layer.

The decoder is a deconvolution network. A spectrum shuffling layer is added before each convolution layer to improve the speech clarity of enhanced BC speech.

\[
\text{h:24 \quad k:3\times3 \quad w:128 \quad c:128 \quad s:1\times2}
\]

\[
\text{k:3\times3 \quad c:256 \quad s:2\times2}
\]

\[
\text{k:3\times3 \quad c:512 \quad s:2\times2}
\]

\[
\text{k:6\times3 \quad c:1024 \quad s:1\times2}
\]

![Discriminator Structure of CycleGAN-DAL.](image)

2) Discriminator: We adopt two discriminators, i.e., \(D^C\) and \(D^D\) in the proposed CycleGAN-DAL model, and the network structure of both is the same, as shown in Fig. 3.

As can be seen in Fig. 3, the input of the discriminators \(D^C\) and \(D^D\) are the speech features of 128 successive frames and the output is a sigmoid operation of the full connection layer. To capture the inter-frame spectral texture, the discriminators utilize 2D convolution. The encoder block of each discriminator includes three stacked 2D convolution networks, each of which consists of a 2D convolutional layer, an IN layer, and a GLU activation layer. The detailed parameter settings of each discriminator are shown in Table I.

### III. EXPERIMENT

#### A. Experimental Conditions

To confirm the effectiveness of the ideas proposed in Section II, we conducted objective and subjective evaluation experiments. We perform BC speech enhancement experiments on two public BC speech datasets, i.e., the TMHINT dataset proposed in [9] and the AEUCHSAC&BC-2017 dataset proposed in [10].

In this work, all BC speech is downsampled to 16 kHz, and 24-dimensional MCEPs, \(F_0\) and \(A_p\) were extracted using the WORLD vocoder respectively from every 25ms frame of speech shifted by 5 ms. For the proposed CycleGAN-DAL model training, the training set and the test set are divided according to a ratio of 4:1.

To improve the model stability, least squares loss is adopted instead of negative log-likelihood target adversarial loss. The Adam optimizer is used and the trade-off coefficient \(\lambda_{\text{cycle}}\) is set to 10. We set \(\lambda_{\text{id}} = 5\) only for the first \(10^4\) iterations to guide the learning process, and the batch to 1. The learning rate of the generator is 0.0002, while the learning rate of the discriminator is 0.0001.

#### B. Objective Evaluation

To evaluate the quality and intelligibility of the enhanced speech, short-time objective intelligibility (STOI) [11], log-spectral distance (LSD) [12], and P.563 [13] are used for objective evaluation. To show the effectiveness of the proposed CycleGAN-DAL model for BC speech enhancement, the Gaussian mixture model (GMM) based BC speech enhancement [14], the BLSTM-based BC speech enhancement method [5], the starGAN-based speech enhancement method [15], and the conventional CycleGAN method [16] were implemented respectively as baseline methods for comparison.

As shown in Table II, it can be observed that the enhanced AC-like speech STOI value obtained by our method is the highest, in advance of the GMM, the BLSTM, and the two methods based on traditional CycleGAN (NMC and PMC). Table III lists the LSD of the comparison methods.

![MOS with 95% confidence w.r.t different BC speech enhancement method.](image)

#### C. Subjective Evaluation

For subjective test analysis, Mean Opinion (MOS) has been taken to measure the naturalness of the enhanced AC-like speech. A Total of 10 subjects (5 females and 5 males between 18 to 30 years of age and with no known hearing impairments) took part in the subjective test. Here, we randomly played utterances from the systems. In the MOS test, subjects were asked to rate the played utterances on a scale of 1-5, the grades of the enhanced AC-like speech are 5 = excellent, and 1 = bad. In this experiment, we conduct the MOS test among seven systems, the results of the MOS test with 95% confidence intervals are shown in Fig. 4. The score of our method is 4.06, which outperforms other methods.

### IV. CONCLUSION

In this work, we have proposed a parallel cyclegan BC speech enhancement method with dual adversarial loss. This
method replaces adversarial loss with classified adversarial loss and defect adversarial loss. The classification antagonism loss distinguishes the generated speech from the real speech, and the defect antagonism loss measures the difference between the spread spectrum of BC speech and the spectrum of AC speech. The experimental results have shown that this method can recover the lost high-frequency components, obtain the spectrum with better harmonic structure, greatly improve the quality and intelligibility of BC speech, and improve the quality of speech communication in a harsh noise environment.

**ACKNOWLEDGMENT**

This work was supported in part by the National Natural Science Foundation of China under Grant 61372137, and in part by the Natural Science Foundation of Anhui Province under Grant 1908085MF209.

**REFERENCES**

[1] H. S. Shin, H.-G. Kang, and T. Fingscheidt, “Survey of speech enhancement supported by a bone conduction microphone,” in *Speech Communication; 10. ITG Symposium*, 2012, pp. 1–4.

[2] A. Shahina and B. Yeongnarayana, “Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach,” *EURASIP Journal on Advances in Signal Processing*, 2007.

[3] T.-N. Phung, M. Unoki, and M. Akagi, “A study on restoration of bone-conducted speech in noisy environments with LP-based model and gaussian mixture model,” *Journal of Signal Processing*, vol. 16, pp. 409–417, Jan. 2012.

[4] K. Vijayan and K. S. R. Murty, “Comparative study of spectral mapping techniques for enhancement of throat microphone speech,” in *2014 Twentieth National Conference on Communications (NCC)*, 2014, pp. 1–5.

[5] C. Zheng, X. Zhang, M. Sun, J. Yang, and Y. Xing, “A novel throat microphone speech enhancement framework based on deep BLSTM recurrent neural networks,” in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 1258–1262.

[6] J. Zhou, Y. Hu, H. Lian, H. Wang, L. Tao, and H. K. Kwan, “Multimodal voice conversion under adverse environment using a deep convolutional neural network,” *IEEE Access*, vol. 7, pp. 170878–170887, 2019.

[7] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, “Time-domain multi-modal bone/air conducted speech enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.

[8] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, “Sequence-to-sequence voice conversion with similarity metric learning using generative adversarial networks,” in *Interspeech 2017*, 2017, pp. 1283–1287.

[9] H.-P. Liu, Y. Tsao, and C.-S. Fu, “Bone-conducted speech enhancement using deep denoising autoencoder,” *Speech Communication*, vol. 104, pp. 106–112, 2018.

[10] G. Zhang, X. Zhang, C. Zhang, and L. Li, “A novel encoder-decoder model via ns-lstm used for bone-conducted speech enhancement,” *IEEE Access*, vol. 6, pp. 62638–62644, 2018.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[12] A. Gray and J. Markel, “Distance measures for speech processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[13] L. Malfait, J. Berger, and M. Kastner, “P.563—the ITU-T standard for single-ended speech quality assessment,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[14] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[15] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 266–273.

[16] T. Kaneko and H. Kameoka, “CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks,” in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

---

**TABLE II**

| Corpus | Speaker | BC | GMM_approxn | GMM_em | BLSTM | StarGAN | CycleGAN | ours |
|--------|---------|----|-------------|--------|-------|---------|---------|-----|
| Dataset AEUCHSAC&BC-2017 | Female | 0.63 | 0.68 | 0.65 | 0.80 | 0.75 | 0.78 | 0.79 |
| | Male | 0.61 | 0.66 | 0.64 | 0.78 | 0.74 | 0.77 | 0.81 |
| | Average | 0.620 | 0.670 | 0.645 | 0.790 | 0.745 | 0.775 | 0.800 |
| Dataset TMHINT | 01 | 0.73 | 0.76 | 0.75 | 0.86 | 0.83 | 0.84 | 0.85 |
| | 02 | 0.70 | 0.71 | 0.69 | 0.80 | 0.78 | 0.83 | 0.84 |
| | 03 | 0.73 | 0.78 | 0.76 | 0.90 | 0.85 | 0.88 | 0.88 |
| | 04 | 0.68 | 0.73 | 0.71 | 0.84 | 0.80 | 0.80 | 0.85 |
| | 05 | 0.60 | 0.64 | 0.61 | 0.70 | 0.74 | 0.77 | 0.86 |
| | Average | 0.687 | 0.727 | 0.707 | 0.817 | 0.795 | 0.823 | 0.842 |

**TABLE III**

| Corpus | Speaker | BC | GMM_approxn | GMM_em | BLSTM | StarGAN | CycleGAN | ours |
|--------|---------|----|-------------|--------|-------|---------|---------|-----|
| Dataset AEUCHSAC&BC-2017 | Female | 1.73 | 1.45 | 1.41 | 1.08 | 1.12 | 1.00 | 0.99 |
| | Male | 1.82 | 1.48 | 1.44 | 1.16 | 1.13 | 1.03 | 1.01 |
| | Average | 1.775 | 1.465 | 1.425 | 1.120 | 1.125 | 1.013 | 1.000 |
| Dataset TMHINT | 01 | 1.58 | 1.32 | 1.31 | 1.07 | 1.08 | 0.98 | 0.96 |
| | 02 | 1.21 | 1.00 | 1.00 | 0.98 | 1.03 | 0.94 | 0.94 |
| | 03 | 1.19 | 1.06 | 1.05 | 0.89 | 1.01 | 0.90 | 0.91 |
| | 04 | 1.20 | 1.06 | 1.10 | 0.98 | 1.07 | 0.97 | 0.95 |
| | 05 | 1.17 | 0.99 | 1.01 | 1.11 | 1.09 | 1.00 | 0.99 |
| | Average | 1.235 | 1.067 | 1.063 | 0.992 | 1.043 | 0.955 | 0.945 |