Bone-conducted Speech Enhancement Using Vector-quantized Variational Autoencoder and Gammachirp Filterbank Cepstral Coefficients

Quoc-Huy Nguyen, Masashi Unoki
Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
{hqnguyen, unoki}@jaist.ac.jp

***Abstract*—Bone-conducted (BC) speech potentially avoids the undesired effects on recorded speech due to background noise or reverberation; however, BC speech has lower quality and intelligibility than air-conducted (AC) speech. Since a large-scale BC speech database is hard to obtain (low-resource), current BC speech enhancement methods hardly improve the speech of speakers outside the training dataset. We proposed a method for enhancing BC speech from speakers outside of the training dataset in such a low-resource scenario. The proposed method contained a feature conversion model based on a vector-quantized variational autoencoder incorporating the gammachirp filterbank cepstral coefficients. The proposed method exploited the large-scale clean AC speech database to improve the quality of the BC speech. We conducted three evaluations to determine the effectiveness of the proposed method: perceptual evaluation of speech quality, short-time objective intelligibility, and the syllable error rate of the automatic speech recognition system. The results indicated that the proposed method could improve the sound quality and intelligibility of the BC speech from speakers outside of the training dataset.**

***Index Terms*—bone-conducted speech, speech enhancement, gammachirp filterbank cepstral coefficients, vector-quantized variational autoencoder**

I. Introduction

In our daily environments, speech signals are distorted due to background noise and reverberation. The performance of useful applications such as telephone communication and automatic speech recognition (ASR) systems, therefore, is reduced. A bone-conducted (BC) microphone can alleviate the adverse effect of the surrounding environment on the speech by recording the speech signal transmitted via the skull of the speaker [1]. The recorded speech signal is named BC speech to distinguish it from conventional air-conducted (AC) speech transmitted via air vibration. However, the sound quality and intelligibility of the BC speech are degraded due to the BC characteristics. This degradation varies between the speakers and pronounced utterances [2]. Hence, the quality and the intelligibility of BC speech need to be improved.

Several studies investigated the characteristics of BC speech. The bone conduction is similar to a low-pass filter with a cut-off frequency of about 1 kHz [3], suggesting that the speech fundamental frequency is preserved [4]. Many studies have proposed BC speech enhancement methods, such as long-term spectra-based [2], linear prediction-based (LP-based), and modulation-transfer-functions (MTF) [3], which aimed to model the inverse filter of the bone conduction. These methods require parameters that are heavily reliant on the information of the corresponding clean AC speech, which is not available in the noisy surrounding environment.

Further studies developed BC speech enhancement methods that attempted to estimate the features of AC speech from the BC speech. A common approach is to convert the spectral-envelope-related features, such as LP-based [3] or cepstrum-based features in the Mel [5] or linear frequency scale [6]. The feature conversion models can be either deterministic [7] or statistical such as a Gaussian mixture model [8], deep feed-forward neural network (DNN) [6], denoising autoencoder (DDAE) [5], or long short-term memory network (LSTM) [9]. Although the statistical models give prominent results, these models require a large-scale AC-BC speech database, where the AC and BC speech signals are recorded simultaneously. Such an AC-BC speech database is too costly to obtain, so the available data size is small, i.e. a low-resource scenario. As a result, current statistical models fail on unseen speakers whose speech data are not in the training dataset.

We proposed a BC speech enhancement method to overcome the above problems. We used gammachirp filterbank cepstral coefficients (GCFCCs) as the target features and designed our feature conversion model on the basis of the vector-quantized variational autoencoder (VQVAE) [10], which we named BCE-VQVAE. Our idea is to build a dictionary of AC speech features using VQVAE and 'look up' the BC speech features in the dictionary for enhancement. The BCE-VQVAE could utilize the large-scale clean AC speech database with higher speaker diversity to improve the enhancement results on the small-size AC-BC database. Also, the features based on the gammachirp filterbank were shown to have a high correlation with speech intelligibility [11] as well as the recognition rate of the ASR systems [12]. Thus, the use of the GCFCCs is to improve the quality of enhanced speech.

In the rest of the paper, Section II describes the details, and Section III shows the evaluations of the proposed method.

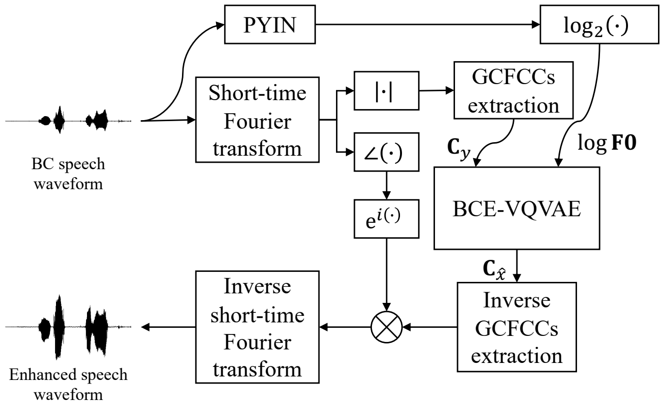


Fig. 1: Block diagram of the proposed method.

II. Proposed Method

The processing flow of the proposed BC speech enhancement method is shown in Fig. 1. We present the detailed information in the below sections.

A. Gammachirp Filterbank Cepstral Coefficients

The gammachirp filterbank (GCF) [13] is a model of the auditory filterbank - a set of band-pass filters reflecting the frequency selectivity on the cochlea [14]. The impulse response of the filter with a center frequency f_c is defined as

$$g(t) = at^{n-1}e^{-2\pi b\mathrm{ERB}(f_c)t}\cos\left(2\pi f_c t + c\ln t\right), \quad (1)$$

where a is the amplitude, n is the order of filter, b is the bandwidth, c is the chirp rate, and the equivalent rectangular bandwidth $\mathrm{ERB}(f_c)$ is defined as

$$\mathrm{ERB}(f_c) = 24.7 + 0.108 f_c\,. \quad (2)$$

For computation, the set $\{f_c(k)\}_{k=1}^{K}$ of K center frequencies are selected in accordance with Slaney's Auditory Toolbox[1].

The GCFCCs can be obtained in the frequency domain from the magnitude spectrogram $\mathbf{M}$ by two steps:

– Step 1: Compute GCF magnitude spectrum

$$g_{k,r} = (\mathbf{VM})_{k,r} = \sum_{l} \frac{\exp\left(c\theta_{k,l}\right)}{B_k\left(2\pi D_{k,l}\right)^n} m_{l,r}\,, \quad (3)$$

where

$$D_{k,l} = \sqrt{B_k^2 + \Delta f_{k,l}^2}\,, \quad (4)$$

$$\theta_{k,l} = \arctan\frac{\Delta f_{k,l}}{B_k}\,, \quad (5)$$

$$B_k = b\mathrm{ERB}(f_c(k))\,, \quad (6)$$

$$\Delta f_{k,l} = \frac{l f_s}{2N} - f_c(k)\,. \quad (7)$$

– Step 2: Compute GCF cepstral coefficients

$$c_{k',r} = \sum_{k=1}^{K} u_{k',k}\log_{10}\left(g_{k,r}\right). \quad (8)$$

In Eq. (3) and Eq. (8), $\mathbf{V} = \{v_{k,l}\}$ is the magnitude frequency response of the k-th filter in the GCF [15] normalized by the filter's bandwidth, $\mathbf{U} = \{u_{k',k}\}$ is the orthonormal basis representing the discrete cosine transform [16], l is the frequency bin index of N frequency bins, r is the frame index, k' is the cepstral coefficient index, and f_s is the sampling frequency. The inverse of the GCFCCs extraction is done by inverting $\mathbf{U}$ and $\mathbf{V}$ straightforwardly.

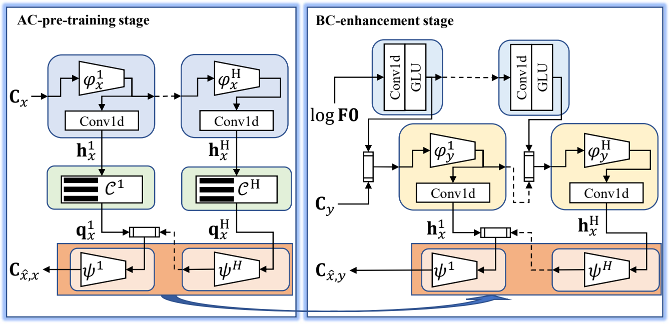


Fig. 2: Block diagram of the proposed BCE-VQVAE model in AC-pre-training and BC-enhancement stages.

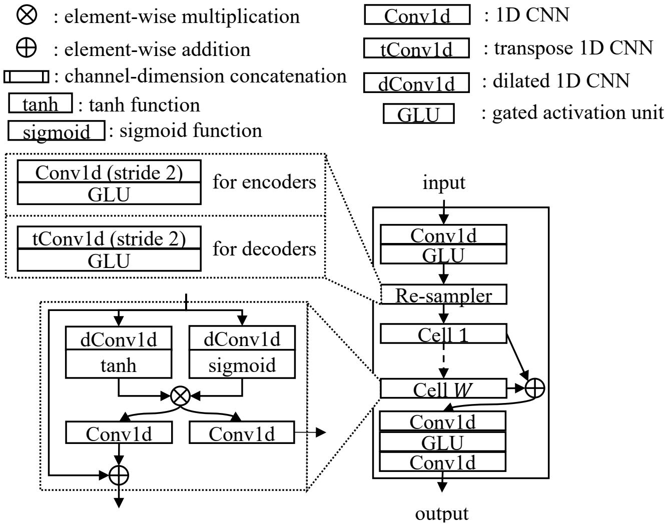


Fig. 3: WaveNet-based architecture for encoders and decoders.

B. Proposed BCE-VQVAE Model

Our idea is to construct a fixed dictionary from AC speech features and make the enhancement model map the BC speech features to the correct items in the dictionary. For this purpose, we designed our BC speech enhancement model by using the VQVAE [10], called BCE-VQVAE.

The VQVAE [10] is a generative model that can capture a data distribution by a finite discrete latent space. This property makes VQVAE suitable for our purpose. The model includes an encoder φ, codebook $\mathcal{C}$, and decoder ψ. With input feature $\mathbf{x}$, VQVAE optimizes the following loss function L

$$L = \left\|\mathbf{x} - \hat{\mathbf{x}}\right\|^2 + \left\|\mathbf{q} - \mathrm{sg}(\mathbf{h})\right\|^2 + \beta\left\|\mathbf{h} - \mathrm{sg}(\mathbf{q})\right\|^2, \quad (9)$$

[1] https://engineering.purdue.edu/~malcolm/interval/1998-010/

where

$$\mathbf{h} = \varphi(\mathbf{x})\,, \tag{10}$$

$$\mathbf{q} = \underset{\tilde{\mathbf{q}}\in\mathcal{C}}{\operatorname{argmin}} \left\|\mathbf{h} - \tilde{\mathbf{q}}\right\|^2, \tag{11}$$

$$\hat{\mathbf{x}} = \psi\left(\mathbf{h} + \operatorname{sg}(\mathbf{q} - \mathbf{h})\right)\,. \tag{12}$$

and the stop-gradient function $\operatorname{sg}(\mathbf{x})$ was defined in [10].

On the basis of VQVAE, we proposed the BCE-VQVAE with two stages of training: AC-pre-training and BC-enhancement stages. The process flow of each stage was illustrated in the block diagram in Fig. 2. Inspired by [17], the BCE-VQVAE contains H stacked VQVAEs (see Fig. 2), where φ_x^h, φ_y^h, ψ^h, and $\mathcal{C}^h$ are the AC and BC encoders, decoder, and codebook, respectively, at level h ($h = 1, \ldots, H$). The encoders and decoders have WaveNet-based architecture (WNA) [18] illustrated in Fig. 3. Also, we defined

$$\mathbf{h}_\nu^h = \operatorname{Conv1d}\left(\varphi_\nu^h\left(\ldots\varphi_\nu^2\left(\varphi_\nu^1\left(\mathbf{C}_\nu\right)\right)\right)\right)\,, \tag{13}$$

$$\mathbf{q}_x^h = \underset{\tilde{\mathbf{q}}\in\mathcal{C}^h}{\operatorname{argmin}} \left\|\mathbf{h}_x^h - \tilde{\mathbf{q}}\right\|^2, \tag{14}$$

$$\mathbf{e}_x^h = \mathbf{h}_x^h + \operatorname{sg}\left(\mathbf{q}_x^h - \mathbf{h}_x^h\right)\,, \tag{15}$$

$$\mathbf{C}_{\hat{x},x} = \psi^1\left(\left[\mathbf{e}_x^1 \quad \psi^2\left(\ldots\psi^{H-1}\left(\left[\mathbf{e}_x^{H-1} \quad \psi^H\left(\mathbf{e}_x^H\right)\right]\right)\right)\right]\right)\,, \tag{16}$$

$$\mathbf{C}_{\hat{x},y} = \psi^1\left(\left[\mathbf{h}_y^1 \quad \psi^2\left(\ldots\psi^{H-1}\left(\left[\mathbf{h}_y^{H-1} \quad \psi^H\left(\mathbf{h}_y^H\right)\right]\right)\right)\right]\right)\,. \tag{17}$$

where $\mathbf{C}_x$ and $\mathbf{C}_y$ are the GCFCCs of AC and BC speech, respectively, and the symbol ν is in $\{x, y\}$. In the AC-pre-training stage, $\left\{\varphi_x^h, \psi^h, \mathcal{C}^h\right\}_{h=1}^H$ are trained to capture the distribution of $\mathbf{C}_x$ using the loss function as

$$L_{\mathrm{AC}} = \gamma L_{\mathrm{SISDR}}(\mathbf{s}_x, \mathbf{s}_{\hat{x}}) + \left\|\mathbf{C}_x - \mathbf{C}_{\hat{x},x}\right\|^2 + \sum_h \left(\left\|\mathbf{q}_x^h - \operatorname{sg}(\mathbf{h}_x^h)\right\|^2 + \beta\left\|\mathbf{h}_x^h - \operatorname{sg}(\mathbf{q}_x^h)\right\|^2\right)\,, \tag{18}$$

where

$$L_{\mathrm{SISDR}}(\mathbf{s}_x, \mathbf{s}_{\hat{x}}) = -10\log_{10}\frac{\left(\mathbf{s}_x^\top\mathbf{s}_{\hat{x}}\right)^2}{\left\|\mathbf{s}_x\right\|^2\left\|\mathbf{s}_{\hat{x}}\right\|^2 - \left(\mathbf{s}_x^\top\mathbf{s}_{\hat{x}}\right)^2}\,. \tag{19}$$

$\mathbf{s}_x$ and $\mathbf{s}_{\hat{x}}$ are waveform of AC speech and the reconstructed waveform from $\mathbf{C}_{\hat{x}}$ (as in Fig. 1). The L_{SISDR} is the scale-invariant signal-to-distortion ratio loss [19], which makes the output magnitude features able to match well with the phase information. In the BC-enhancement stage, the $\left\{\varphi_y^h\right\}_{h=1}^H$ are trained to enhance $\mathbf{C}_y$ using the loss function as

$$L_{\mathrm{BC}} = \gamma L_{\mathrm{SISDR}}(\mathbf{s}_x, \mathbf{s}_{\hat{x}}) + \left\|\mathbf{C}_x - \mathbf{C}_{\hat{x},y}\right\|^2 + \sum_h \beta\left\|\mathbf{h}_y^h - \operatorname{sg}(\mathbf{q}_x^h)\right\|^2\,. \tag{20}$$

1) F0 Injection: The cut-off frequency of bone conduction is about 1 kHz, while the F0 of human speech varies from 60 to 300 Hz. Therefore, the F0 information should be preserved in BC speech. Thus, we propose to concatenate the log F0 information to the inputs of the BC encoders (see Fig. 2) to improve BC encoders. Probabilistic YIN (PYIN) [20], which is one of the most powerful frame-wise F0 estimation algorithms, was used for F0 estimation.

2) Data Augmentation: We trained the model with augmented data, then fine-tuned it on the AC-BC dataset. From the clean AC speech, we synthesize artificial BC speech by modeling bone conduction and recording noise. We model bone conduction by applying the measured power transfer function from the oral cavity to the temporal region on the magnitude spectrum of AC speech [21] while randomly distorting the phase spectrum of which frequencies are above 3 kHz. The recording noise is modeled by the white noise with the signal-to-noise ratio randomly from 15 to 25 dB. According to our preliminary experiments, the fine-tuning process hardly converges on the AC-BC dataset.

III. Evaluations

A. Data, Configurations, and Experimental Setups

The AC-BC speech dataset was recorded from 14 speakers (ten males and four females) in a soundproof room. The BC microphone was the TEMCO HG70. The speech data from 12 speakers were used for training (seen speakers). The remaining data from one male speaker and one female speaker were used for evaluation (unseen speakers). Each speaker pronounced 46 Japanese utterances. For the AC-pre-training stage, the clean Japanese utterances from the 'parallel100' and 'nonpara30' subcorpora of the JVS dataset [22] were used. All the signals were re-sampled to 16 kHz.

Both the short-time Fourier transform (STFT) and the PYIN algorithm use a 32-ms window with a hop length of 8 ms. The GCF consists of 32 gammachirp filters with $c = -2$, $b = 1$, $n = 4$, $f_{\min} = 60$ Hz, and $f_{\max} = 8$ kHz.

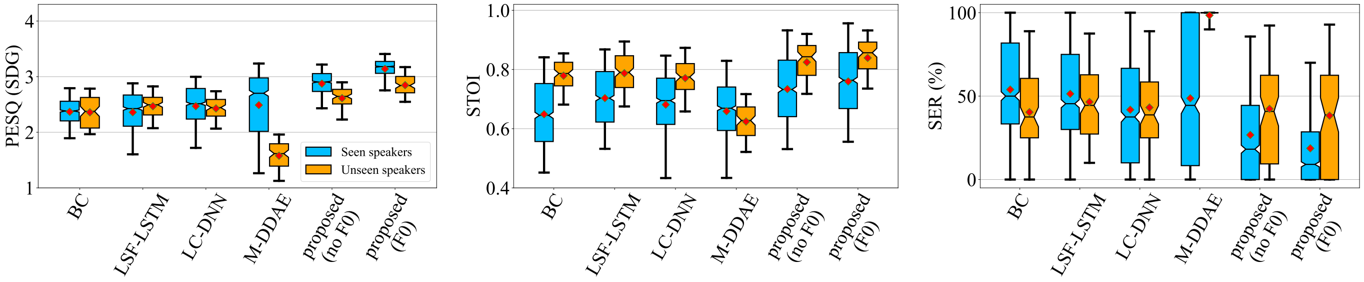
The BCE-VQVAE had $H = 3$ stacked VQVAEs. Each Conv1d layer had 128, 64, and 64 output channels for $h = (1, 2, 3)$, respectively. The kernel size of all Conv1d layers was three. In the WNA, we set $W = 6$ and the dilation list is $(1, 2, 4, 1, 2, 4)$ (see Fig. 3). The Adam optimizer [23] was used with a learning rate of 10^{-4} and a 16-sample batch per iteration. Each sample was a 4.088-s segment of speech signals. When fine-tuning the model on the AC-BC dataset, the learning rate was 10^{-5}, and the batch size was two.

We used three objective evaluation metrics, i.e., perceptual evaluation of speech quality (PESQ) [24], short-time objective intelligibility (STOI) [25], and syllable error rate (SER) of the ASR systems. PESQ is a metric to measure speech quality while STOI is well-known to evaluate intelligibility. The range of PESQ is from -0.5 to 4.5 while that of STOI is from 0 to 1. A higher score means better speech quality and intelligibility.

The SER is used to measure how well the speech signals can be recognized in a practical ASR system regarding the acoustic units of utterances, i.e., syllables. Since we used the Japanese corpus, the SER is equivalent to the character error rate of the katakana transcriptions. In Japanese, katakana is a writing system representing the syllables. The SER was computed from the two audio signals via three steps:

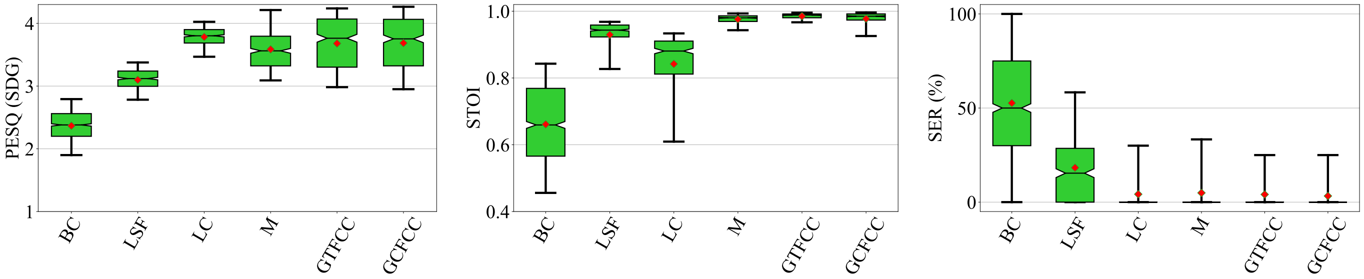
1) Transcribe both speeches to texts using Google ASR[2].
2) Convert the texts to katakana.

[2] https://cloud.google.com/speech-to-text



(a) PESQ (b) STOI (c) SER

Fig. 4: PESQ, STOI, and SER of BC speech, the enhanced speech by current methods and by our methods on seen and unseen speakers.



(a) PESQ (b) STOI (c) SER

Fig. 5: PESQ, STOI, and SER of BC speech and the enhanced speech using the AC speech features in the current methods and in our methods.

3) Compute the SER by the following formula

$$\text{SER}\,(\%) = \frac{S + D + I}{N} \times 100\%\,, \tag{21}$$

where S, D, and I are the numbers of substitutions, deletions, and insertions, respectively, required to transform one text into another.

The value of the SER is positive and can be larger than 100%. The lower the SER, the more accurately the degraded speech is recognized by the ASR systems.

B. Evaluations

The box plots in Fig. 4 shows the evaluation results of the proposed methods compared with some state-of-the-art methods including line spectral frequencies (LSF) with LSTM [9], linear low-order cepstral coefficients (LC) with DNN [6], and Mel-scale log-spectrogram (M) with DDAE [5]. Also, we compared two versions of the proposed methods: with and without F0 injection. The results are shown for both the seen and unseen speakers (see Section III-A). The line at the notch, the lower, and the upper edges of each box denotes the median, the first, and third quartiles of the data. The upper and lower whiskers denote the 5-th and 95-th percentiles. The mean is identified with the red diamond marker.

From Fig. 4a and 4b, the proposed methods significantly enhanced the BC speech and outperformed the current methods in terms of PESQ and STOI for both seen and unseen speakers. From Fig. 4c, the proposed methods had considerable improvement on seen speakers in terms of SER, yet the improvement was slight for unseen speakers. The proposed method with F0 information also gave higher results than the one without F0 information.

Furthermore, we evaluated the enhanced speech when the features of AC speech were given. The purpose of this evaluation was to reveal which features were effective for BC speech enhancement. Fig. 5 shows the evaluation results of the enhanced speech by GCFCCs and other features including LSF [9], LC [6], and M [5] on the whole AC-BC dataset. We also considered a special case of GCFCCs when $c = 0$, which is the gammatone filterbank cepstral coefficients (GTFCCs). From Fig. 5, the results indicated that the M, GTFCCs, and GCFCCs outperformed in STOI. The mean and median PESQ scores given by the GTFCCs and GCFCCs were slightly higher than the M but lower than the LC. Although all the LC, M, GTFCCs, and GCFCCs provided zero-median in SER, the deviations of the SER given by GTFCCs and GCFCCs were smaller.

C. Discussion

From the evaluation results, our proposed method gave the best results when dealing with both seen and unseen speakers. The incorporation of F0 information could improve the results significantly for seen speakers. Thus, the F0 injection seemed to overfit the seen data. Our method modified the magnitude features while keeping the phase information unmodified. When there is a mismatch between the unmodified phase and the modified magnitude, distortions can occur, which prevents the SER from improving.

We observed that, overall, the auditory-filterbank-based features such as GCFCCs and GTFCCs gave slightly better results than other features. The Mel filters have a symmetrical triangle shape, while the shapes of the gammatone and gammachirp filters resemble the frequency selectivity of human auditory system. The bandwidth and frequency scale of Mel filters and gammatone/gammachirp filters are also different. These results affirm the importance of auditory-filterbank-based features in enhancing BC speech signals.

IV. Conclusions

We proposed a method for bone-conducted (BC) speech enhancement using gammachirp filterbank cepstral coefficients (GCFCCs) features incorporated into BC speech enhancement vector-quantized variational autoencoder (BCE-VQVAE) to exploit the clean air-conducted (AC) speech database. We conducted three evaluations including perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and syllable error rate (SER) in Google ASR systems to evaluate the effectiveness of the proposed method. The results indicated that the proposed method outperformed current methods for both seen and unseen speakers. Especially, when dealing with unseen speakers, the proposed method gave the best improvement in terms of PESQ and STOI. The results also revealed that auditory-filterbank-based features are important for speech quality and intelligibility. For future work, the enhancement of phase information and further research in effectively using F0 information should be considered to enhance BC speech further.

Acknowledgment

This work was supported by a Grant-in-Aid for Scientific Research (B) (No. 17H01761), JSPS KAKENHI Grant (No.20J20580), and Fund for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233).

References

[1] Shiming Zhang, Yosuke Sugiura, Nozomiko Yasui, and Tetsuya Shimamura. Quantifying Noise Robustness of Bone-Conducted Speech. In *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 582–585, 2020.

[2] Tetsuya Shimamura and Takeshi Tomikura. Quality Improvement of Bone-Conducted Speech. In *Proceedings of the 2005 European Conference on Circuit Theory and Design, 2005.*, pages III/73–III/76 vol. 3, 2005.

[3] Thang Tat Vu, Kenji Kimura, Masashi Unoki, and Masato Akagi. A Study on Restoration of Bone-Conducted Speech with MTF-Based and LP-Based Models. *Journal of Signal Processing*, 10(6):407–417, 2006.

[4] Mizanur Shahidur Rahman and Tetsuya Shimamura. Pitch characteristics of bone conducted speech. In *European Signal Processing Conference*, pages 795–799. IEEE, 2010.

[5] Hungping Liu, Yu Tsao, and Chiou-Shann Fuh. Bone-conducted speech enhancement using deep denoising autoencoder. *Speech Communication*, 104(June):106–112, 2018.

[6] Daiki Watanabe, Yosuke Sugiura, Tetsuya Shimamura, and Hisanori Makinae. Speech enhancement for bone-conducted speech based on low-order cepstrum restoration. In *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 212–216, 2017.

[7] Mizanur Shahidur Rahman and Tetsuya Shimamura. Intelligibility enhancement of bone conducted speech by an analysis-synthesis method. In *Midwest Symposium on Circuits and Systems*, pages 1–4. IEEE, 2011.

[8] Trung Nghia Phung, Masashi Unoki, and Masato Akagi. A Study on Restoration of Bone-Conducted Speech in Noisy Environments with LP-based Model and Gaussian Mixture Model. *Journal of Signal Processing*, 16(5):409–417, 2012.

[9] Huy Quoc Nguyen and Masashi Unoki. Improvement in bone-conducted speech restoration using linear prediction and long short-term memory model. *Journal of Signal Processing*, 24:175–178, 2020.

[10] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6306–6315. Curran Associates, Inc., 2017.

[11] Toshio Irino and Roy D. Patterson. The gammachirp auditory filter and its application to speech perception. *Acoustical Science and Technology*, 41(1):99–107, 2020.

[12] Md Jahangir Alam, Patrick Kenny, Pierre Dumouchel, and Douglas O'Shaughnessy. Robust feature extractors for continuous speech recognition. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 944–948, 2014.

[13] Toshio Irino and Roy D. Patterson. A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America*, 101(1):412–419, 1 1997.

[14] William Yost. *Fundamentals of Hearing: An Introduction*. Brill, 2013.

[15] Toshio Irino and Masashi Unoki. An analysis/synthesis auditory filterbank based on an iir implementation of the gammachirp. *Journal of the Acoustical Society of Japan (E)*, 20:397–406, 1999.

[16] Nasir Ahmed, T. Natarajan, and Kamisetty Ramamohan Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

[17] Tuan Vu Ho and Masato Akagi. Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pages 140–144. International Speech Communication Association, 10 2020.

[18] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop*, page 125, 2016.

[19] Shuai Li, Hongqing Liu, Yi Zhou, and Zhen Luo. A si-sdr loss function based monaural source separation. In *2020 15th IEEE International Conference on Signal Processing (ICSP)*, volume 1, pages 356–360, 2020.

[20] Matthias Mauch and Simon Dixon. PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 659–663. Institute of Electrical and Electronics Engineers Inc., 2014.

[21] Teruki Toya, Peter Birkholz, and Masashi Unoki. Measurements of transmission characteristics related to bone-conducted speech using excitation signals in the oral cavity. *Journal of Speech, Language, and Hearing Research*, 63(12):4252–4264, 12 2020.

[22] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. JVS corpus: free Japanese multi-speaker voice corpus. 8 2019.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[24] Antony W. Rix, John G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2, pages 749–752, 2001.

[25] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 4214–4217, 2010.