IEICE TRANS. INF. & SYST., VOL.E102–D, NO.8 AUGUST 2019

LETTER

# Speech Quality Enhancement for In-Ear Microphone Based on Neural Network

**Hochong PARK**[†a)], ***Member***, **Yong-Shik SHIN**[††], ***and*** **Seong-Hyeon SHIN**[†], ***Nonmembers***

**SUMMARY** Speech captured by an in-ear microphone placed inside an occluded ear has a high signal-to-noise ratio; however, it has different sound characteristics compared to normal speech captured through air conduction. In this study, a method for blind speech quality enhancement is proposed that can convert speech captured by an in-ear microphone to one that resembles normal speech. The proposed method estimates an input-dependent enhancement function by using a neural network in the feature domain and enhances the captured speech via time-domain filtering. Subjective and objective evaluations confirm that the speech enhanced using our proposed method sounds more similar to normal speech than that enhanced using conventional equalizer-based methods.

***key words:*** *in-ear microphone, neural network, noise-free speech, speech quality enhancement*

## 1. Introduction

Many studies on capturing noise-free speech in noisy environments have been carried out. A noise cancellation approach captures both speech and noise signals and then cancels noise based on the difference between the two signals [1]. However, this technique has limitations because speech and noise signals inevitably share many common components. A noise blocking approach, which is used to prevent noise from entering the microphone by capturing speech through bone and tissue conduction, is a more effective method for capturing noise-free speech [2]–[6]. The most convenient way of blocking noise is to capture speech from inside an occluded ear by using an in-ear microphone (IEM) [2]–[4].

Because of the different sound-transmitting pathways, speech captured from inside an occluded ear has different sound characteristics from normal speech captured in front of the mouth through air conduction. Accordingly, the speech captured using an IEM is considered degraded and sounds different from normal speech. Hence, to use an IEM successfully, a speech enhancement module as a post processor is required for the captured speech.

The goal of speech enhancement is to recover the signal modification caused by an IEM. A typical form of this modification is the spectral envelope change in all bands, with obvious reduction in high-band level; the harmonic structure is not changed [2]–[4]. Accordingly, a bandwidth extension (BWE) was proposed as an enhancement solution that focuses on recovering the high band [4]. However, the BWE may not be an appropriate enhancement strategy for IEMs, because it recovers the high band without using the correct high-band harmonic structure that remains unchanged in the captured speech. Moreover, the BWE cannot enhance the degraded low-band spectral envelope, which is often the main reason for the low quality of the captured speech.

An equalizer (EQ) can enhance speech quality by adjusting the spectral envelope of the degraded speech close to that of the target speech [2], [3]. It was confirmed from our investigation that the nature of speech modification by the IEM depends on the speech phoneme. However, the conventional EQs proposed in [2], [3] cannot perform the spectral adjustment in a phoneme-dependent manner and therefore cannot enhance the degraded speech to the desired level. Hence, an input-dependent method based on learning or modeling is required, such as a neural network [5] or a Gaussian mixture model [6].

Many methods of speech enhancement based on a neural network have been developed [7]–[12]. They generally have a long processing delay, especially when using a cost function that directly measures speech quality, such as short-time objective intelligibility (STOI) [9]–[12]. In addition, most methods aim to enhance noisy speech by implementing complex functions with a large network. In contrast, speech enhancement for the IEM, which works while capturing speech on a tiny in-ear device, requires a short-delay and low-complexity method that is specialized for the IEM-induced distortion. Therefore, the conventional enhancement methods may not be directly applicable to speech enhancement for the IEM.

In this study, a new speech enhancement method for an IEM is proposed. The IEM has only one microphone inside the ear canal, and a blind method that utilizes only a single-channel input is designed. The proposed method determines an input-dependent enhancement function on a short frame basis by using a neural network in the form of feature mapping. Subsequently, the enhancement is implemented by applying the estimated target features to the input through time-domain filtering. Moreover, a preliminary high-band booster is included to recover roughly the high-band reduction due to the IEM, thus relieving the burden on the neural network. Subjective and objective evaluations confirm the effectiveness of the input-dependent operation of the proposed method in enhancing the speech quality, in

Manuscript received December 5, 2018.
Manuscript revised March 20, 2019.
Manuscript publicized May 15, 2019.
[†]The authors are with the Department of Electronics Engineering, Kwangwoon University, Seoul, Korea.
[††]The author is with RippleBuds Ltd., Korea.
a) E-mail: hcpark@kw.ac.kr
DOI: 10.1587/transinf.2018EDL8249

comparison to conventional EQs [2], [3].

## 2. Proposed Speech Quality Enhancement Method

### 2.1 Methodology of Speech Quality Enhancement

To analyze the characteristics of the signal modified in the IEM, speech signals were captured simultaneously using the IEM [13] and a normal microphone held in front of the mouth; the signals are denoted by $x(t)$ and $y(t)$, respectively. $x(t)$ corresponds to the raw captured speech to be used as an input for enhancement, and $y(t)$ corresponds to normal speech to be used as a target for enhancement.

Figure 1 shows the spectra of $x(t)$ and $y(t)$ and the difference in the spectral envelope between $x(t)$ and $y(t)$ for different phonemes corresponding to two different speakers. The correct harmonic structure of speech is maintained in $x(t)$. In contrast, a change in the spectral envelope occurs in all bands of $x(t)$, with obvious high-band reduction, because of the occluded ear effect. Moreover, the shape of the spectral envelope mismatch varies with respect to $x(t)$ with different phonemes, because different phonemes resulting from the unique shapes of the mouth and jaw lead to different shapes of the transmitting pathways. Hence, for enhancing the speech quality, a spectral envelope correction function that depends on the input $x(t)$ should be obtained, which is denoted by $S(f;x)$ in the form of frequency response.

In the proposed method, $S(f;x)$ is estimated using a neural network. Because the neural network aims to correct the spectral envelope mismatch, the loss in the neural network should be the spectral envelope error between the neural network output and its target. In this case, the contribution of each band to the loss depends on the band energy, if loss weighting is not used. Therefore, the straightforward learning of the neural network tends to be biased in the direction in which it focuses on reducing the errors in the high-energy bands, making it difficult to recover the significant level reduction in the high bands with low energy.

To solve this problem in training the neural network, in the proposed method, a strategy of pre-boosting the high band is employed, instead of using loss weighting. In other words, by using prior information that the energy in the high bands of $x(t)$ has been reduced significantly, $x(t)$ is first inputted to a high-band booster and its output is then applied to the neural network. A fixed high-band booster is designed empirically after analyzing the average high-band envelope difference between $x(t)$ and $y(t)$ by using the given training dataset. It is implemented via a second-order infinite impulse response (IIR) shelving filter; the IIR filter is selected because low complexity and short processing delay are required. Accordingly, in this two-stage procedure, the high-band boosting plays the role of preliminary rough level matching of the high bands, and the neural network focuses on fine spectral shaping in all the bands.

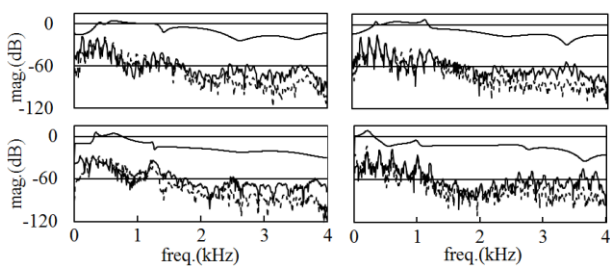


**Fig. 1** Spectra of speech captured by a normal microphone (lower solid) and the IEM (lower dotted), and the spectral envelope difference between the two (upper). Each plot shows the case of /er/ (upper left), /u/ (upper right), /g/ (lower left), and /o/ (lower right).

The neural network for estimating $S(f;x)$ operates in the feature domain. The spectral coefficients or signal samples are not qualified as features, because they show too many details of the signal, while overlooking the overall relationship between $x(t)$ and $y(t)$, leading to a poor learning performance of the neural network. Instead, a low-dimensional feature that effectively expresses the spectral envelope is preferred, such as the linear predictive coefficient, Mel-frequency cepstral coefficient (MFCC), or band energy. No significant differences were found between these candidates in terms of the performance, after conducting many experiments on them, and a method with lower complexity, i.e., the band energy, was finally selected for the proposed method.

The speech enhancement is conducted on a short frame basis to ensure short processing delay, in which the sampling rate of the signal is 8 kHz and the frame size is set to 20 ms. To obtain the band energy, a 20 ms-frame-based spectrum is computed using a 256-point discrete Fourier transform (DFT) with overlap. The spectrum is then divided into 18 Bark-scale bands, and the band energy in the log scale is computed, resulting in an 18-dimensional feature vector.

Based on the above investigation, the overall structure of the proposed enhancement method is designed, as shown in Fig. 2. For each frame, the raw captured speech $x(t)$ is converted into the enhanced speech $y_o(t)$ as follows. The high-band booster output $x_H(t)$ is computed, and the feature vector $\boldsymbol{X}$ of $x_H(t)$, consisting of its band energy, is computed and inputted to the trained neural network. The neural network then outputs the estimate of the target feature vector $\boldsymbol{Y}_o$, corresponding to the band energy of $y_o(t)$. The band energy conversion from $\boldsymbol{X}$ to $\boldsymbol{Y}_o$ is conducted by time-domain filtering, rather than by spectral-domain filtering that requires inverse DFT and an overlap-and-add operation to compute $y_o(t)$, in order to achieve low complexity and short processing delay, which is a strict design constraint in this study. In particular, a set of IIR biquad peaking filters for each band $b$, denoted by a transfer function $H_b(z)$, is de-

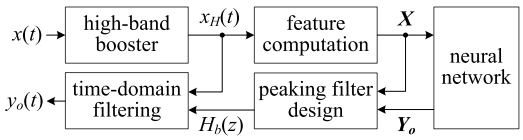


**Fig. 2** Overall structure of the proposed speech enhancement method.

signed [14], where the gain of the peaking filters is set as the difference between $\boldsymbol{X}$ and $\boldsymbol{Y}_o$ and the gain interpolation from the previous frame is applied to ensure a smooth spectral change. The $Q$ value that controls the bandwidth is set to 4.0 in all bands [14]. Then, the resulting peaking filters $H_b(z)$ are applied to $x_H(t)$ with the band energy $\boldsymbol{X}$ to obtain the final output $y_o(t)$ with the band energy $\boldsymbol{Y}_o$.

In summary, the proposed method enables the design of the time-varying IIR filters using neural network such that they can correct the spectral envelope modification in a phoneme-dependent way. Then, it applies the filters to the input to implement the speech enhancement.

### 2.2 Neural Network

The neural network was designed to be as simple as possible while providing acceptable performance, in order to reduce the computational complexity. Accordingly, a basic multilayer neural network with two hidden layers, each with 180 and 60 neurons, was selected. A sigmoid activation function is used in all the layers. Then, the developed enhancement module can run successfully on our target in-ear device.

For training the neural network, the feature vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are computed from the training dataset, after applying the high-band booster to $x(t)$. The neural network is trained such that it searches for the best mapping function from $\boldsymbol{X}$ to $\boldsymbol{Y}$ by using the stochastic gradient descent (SGD) method with a cross-entropy cost function. The training is run by 500 epochs with a mini-batch size of one and learning rate of 0.005.

## 3. Performance Evaluation

As different IEMs have different characteristics depending on their physical structure and shape, the speech database (DB) for training the neural network as well as the performance evaluation should be generated using the target IEM. Therefore, the speech DB was generated locally in our laboratory by recording a set of speech signals by using both the target IEM [13] and a normal microphone. The DB contains speech signals obtained from 10 speakers, and its size is approximately 12 min. The training dataset contains speech signals obtained from eight speakers, and the testing dataset contains speech signals obtained from two speakers, one male and one female, who are not included in the training dataset. Thus, the training is done in a speaker-independent manner.

Figure 3 shows the spectra of the enhanced speech $y_o(t)$ obtained using the proposed method and its target $y(t)$, along with the spectral envelope difference between the two for the same signals shown in Fig. 1. The mismatch in the spectral envelope is significantly reduced, though not completely eliminated. The spectrograms shown in Fig. 4 also confirm that the enhanced speech $y_o(t)$ gets closer to its target $y(t)$ than the raw captured speech $x(t)$.

To confirm the superiority of the proposed method, its performance was compared to that of EQ, because the proposed method is similar to EQ in terms of operation. As in [2], [3], the reference EQ for performance comparison was designed after computing the average spectral envelope difference between $x(t)$ and $y(t)$ by using the training dataset.

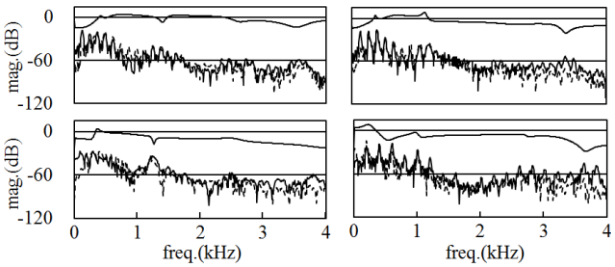


**Fig. 3** Spectra of enhanced speech $y_o(t)$ (lower dotted) and its target $y(t)$ (lower solid), and the spectral envelope difference between the two (upper) for the signals in Fig. 1.

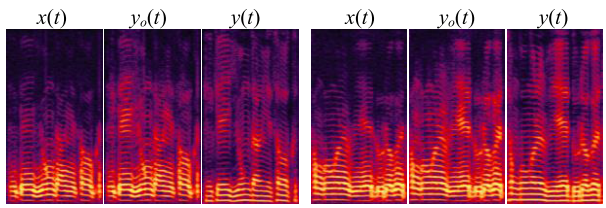


**Fig. 4** Spectrograms of two utterances; each spectrogram has a time length of 1.3 s and a bandwidth of 4 kHz.

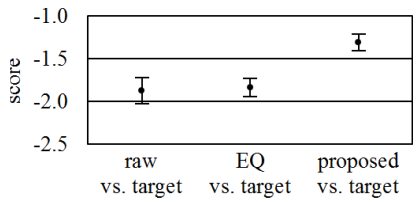


**Fig. 5** Results of subjective evaluation in terms of the comparison category rating.

Subjective evaluation is conducted based on the comparison category rating (CCR), where the quality difference between the two speech signals is measured [15]. Seven subjects participated in the evaluation. Figure 5 shows the CCR results with a 95% confidence interval. For "A vs. target," where A is one to be evaluated, the scores $-1$, $-2$, and $-3$ indicate that the target is slightly better, better, and much better than A, respectively [15]. The scores of "raw vs. target" and "EQ vs. target" show that the perceptual speech quality cannot be enhanced by simply adjusting the spectral envelope via a fixed EQ, despite the less muffled sound owing to the high-band boosting by the EQ. In contrast, from the "proposed vs. target" score, the speech processed using the proposed method is found to be significantly closer to normal speech in terms of perceptual speech quality than the raw captured speech, even though a slight sound difference is still perceived.

The objective performance of the enhancement methods is measured via the average log-spectral distortion (ALSD) that analyzes the degree of spectral matching be-

**Table 1** Results of objective evaluation in terms of the average log-spectral distortion (ALSD) and the PESQ.

| | ALSD (dB) | | | PESQ score |
|---|---|---|---|---|
| | 0~2kHz | 2~4kHz | 0~4kHz | |
| raw vs. target | 8.09 | 6.43 | 7.37 | 2.67 |
| EQ vs. target | 8.20 | 6.02 | 7.37 | 2.52 |
| proposed vs. target | 7.18 | 5.68 | 6.55 | 2.71 |

tween the two signals [16]. In the low bands below 2 kHz, the spectral envelope difference between $x(t)$ and $y(t)$ has positive and negative values arbitrarily depending on the phoneme, which makes the average spectral envelope difference converge to zero. Therefore, the reference EQ, determined from the average spectral envelope difference, has an approximately zero gain in the low bands and cannot reduce the ALSD in these bands, as shown in Table 1. The poor performance in the low bands is the main reason why the reference EQ cannot improve the perceptual speech quality. The proposed method reduces the ALSD in both the low and high bands by virtue of its time-varying operation depending on the input $x(t)$. As another objective evaluation, the perceptual evaluation of speech quality (PESQ) score is measured [17]. Table 1 shows that the proposed method provides a higher PESQ score than the reference EQ.

## 4. Conclusion

An IEM placed inside an occluded ear can be used to capture noise-free speech via a noise blocking approach. However, the sound characteristics of the captured speech are different from those of normal speech. Therefore, an enhancement method for the IEM is required. To consider the phoneme-dependent nature of speech degradation, a learning method based on a neural network is proposed. This method corrects the mismatch in the spectral envelope by time-domain filtering whose function is estimated in a phoneme-dependent way using the neural network. From subjective and objective evaluations, it is confirmed that the speech enhanced using the proposed method sounds more like normal speech than that enhanced using conventional EQs.

## Acknowledgments

The work reported in this paper was conducted during the sabbatical year of Kwangwoon University in 2018.

### References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoustics, Speech, and Signal Processing, vol.32, no.6, pp.1109–1121, Dec. 1984.

[2] K. Kondo, T. Fujita, and K. Nakagawa, "On equalization of bone conducted speech for improved speech quality," IEEE Int. Symposium on Signal Processing and Information Technology, pp.426–431, 2006.

[3] A. Bernier and J. Voix, "Signal characterization of occluded in-ear versus free-air voice pickup on human subjects," Canadian Acoustics, vol.38, no.3, pp.78–79, 2010.

[4] R.E. Bouserhal, T.H. Falk, and J. Voix, "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," The Journal of the Acoustical Society of America, vol.141, no.3, pp.1321–1331, March 2017.

[5] A. Shahina and B. Yegnanarayana, "Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach," EURASIP Journal on Advances in Signal Processing, vol.2007.1:87219, 2007.

[6] M.A.T. Turan and E. Erzin, "Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra," Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp.7049–7053, 2013.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," IEEE Signal Process. Lett., vol.21, no.1, pp.65–68, 2014.

[8] K. Han, Y. Wang, D. Wang, W.S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol.23, no.6, pp.982–992, 2015.

[9] S.-Z. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," IEEE/ACM Trans. Audio, Speech and Language Processing, vol.26, no.9, pp.1570–1584, Sept. 2018.

[10] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," IEEE/ACM Trans. Audio, Speech, Language Process., vol.26, no.10, pp.1780–1892, Oct. 2018.

[11] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.5074–5078, 2018.

[12] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.5059–5063, 2018.

[13] [online] https://www.kickstarter.com/projects/ripplebuds/ripplebuds-noise-blocking-earbuds-with-an-in-ear-m

[14] R. Bristow-Johnson, "The equivalence of various methods of computing biquad coefficients for audio parametric equalizers," Audio Engineering Society Convention 97, 1994.

[15] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," Aug. 1996.

[16] K.K. Paliwal and B.S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," IEEE Trans. Speech and Audio Processing, vol.1, no.1, pp.3–14, Jan. 1993.

[17] A.W. Rix, J.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," Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp.749–752, 2001.