Speech Communication 104 (2018) 106–112

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Bone-conducted speech enhancement using deep denoising autoencoder

Hung-Ping Liu[a], Yu Tsao[b,*], Chiou-Shann Fuh[a]

[a] Graduate Institute of Biomedical Electronics and Bioinformatics, Nation Taiwan University, Taipei, Taiwan
[b] Research Center for Information Technology Innovation, Academia Sinica, Taiwan

ARTICLE INFO

Keywords:
Bone-conduction microphone
Deep denoising autoencoder

ABSTRACT

Bone-conduction microphones (BCMs) capture speech signals based on the vibrations of the speaker's skull and exhibit better noise-resistance capabilities than normal air-conduction microphones (ACMs) when transmitting speech signals. Because BCMs only capture the low-frequency portion of speech signals, their frequency response is quite different from that of ACMs. When replacing an ACM with a BCM, we may obtain satisfactory results with respect to noise suppression, but the speech quality and intelligibility may be degraded due to the nature of the solid vibration. The mismatched characteristics of BCM and ACM can also impact the automatic speech recognition (ASR) performance, and it is infeasible to recreate a new ASR system using the voice data from BCMs. In this study, we propose a novel deep-denoising autoencoder (DDAE) approach to bridge BCM and ACM in order to improve speech quality and intelligibility, and the current ASR could be employed directly without recreating a new system. Experimental results first demonstrated that the DDAE approach can effectively improve speech quality and intelligibility based on standardized evaluation metrics. Moreover, our proposed system can significantly improve the ASR performance by a notable 48.28% relative character error rate (CER) reduction (from 14.50% to 7.50%) under quiet conditions. In an actual noisy environment (sound pressure from 61.7 dBA to 73.9 dBA), our proposed system with a BCM outperforms an ACM, yielding an 84.46% reduction in the relative CER (proposed system: 9.13% and ACM: 58.75%).

1. Introduction

Nowadays, automatic speech recognition (ASR) is commonly used as a human-computer interface in phones, home robots, and navigation systems. It is convenient to use speech when interacting with devices and giving instructions to be carried out. Generally, ASR can achieve good recognition performance in quiet environments. However, the recognition accuracy may be degraded in the presence of background noise. Many approaches have been proposed to address this issue in order to obtain better ASR performance. One straightforward way is to adopt noise-resistant devices to collect speech signals. A bone-conduction microphone (BCM) records speech via bone conduction, and thus has the capability to suppress background noise. However, the recorded sounds are also distorted because only the low-frequency portions of speech signals are captured. For the speech recorded by BCMs, human intelligibility is not difficult, but ASR may not give optimal recognition results. It is possible to recreate a new ASR system using a voice database that is recorded by BCM. However, this approach is time-consuming and prohibitive. In this paper, we propose a novel deep learning-based method to bridge the acoustic difference between a BCM and a normal air-conduction microphone (ACM).

In the past, several approaches have been proposed to improve the quality of bone-conducted speech. Shimamura et al. proposed the use of a reconstruction filter, which is designed using the long-term spectra of the BCM and ACM speech, to transform the bone-conducted speech to normal air-conducted speech (Shimamura and Tomikura, 2005). Then, Shimamura et al. proposed to use a multi-layer perceptron model to form the reconstruction filter in order to more accurately model the transformation of BCM to ACM speech (Shimamura et al., 2006). Meanwhile, many approaches have been proposed to combine ACM and BCM using a single sound capture device (Zheng et al., 2003; Liu et al., 2004; Zhang et al., 2004). The information about the speech data from two microphones is integrated to recover the normal speech data. Graciarena et al. proposed the use of the probabilistic optimum filter (POF) mapping algorithm to combine standard and throat microphones for robust speech recognition (Graciarena et al., 2003). The experimental results demonstrate that the combined microphone approach significantly outperforms the single-microphone approach. Thang et al. proposed two types of models, namely the modulation transfer function (MTF) and linear prediction (LP), to restore normal speech from bone-conducted speech in noisy environments (Thang et al., 2006). The experimental results showed that both models can produce better restored

* Corresponding author.
E-mail addresses: howardliu1223@gmail.com (H.-P. Liu), yu.tsao@citi.sinica.edu.tw (Y. Tsao).

https://doi.org/10.1016/j.specom.2018.06.002
Received 7 September 2017; Received in revised form 18 April 2018; Accepted 15 June 2018
Available online 02 July 2018

speech than traditional methods in terms of human hearing and ASR systems. Dekens et al. proposed to use a throat microphone for accurate voice detection to improve speech recognition in noisy environments (Dekens et al., 2010). Kuo et al. showed that BCM can exhibit a higher ASR performance than ACM in different noisy environments (Kuo et al., 2015). Tajiri et al. proposed a noise-suppression method for body-conducted speech recorded by a nonaudible murmur microphone (Tajiri et al., 2017). The proposed algorithm is based on the non-negative tensor-factorization algorithm. The experimental results show that their proposed method is superior to conventional methods under real noisy environments.

Recently, Lu et al. proposed a deep denoising autoencoder (DDAE) approach for speech enhancement (Lu et al., 20132014). The experimental results show that the DDAE model can more effectively reduce noise in speech, and thus improve the speech quality and signal-to-noise ratio (SNR) when compared with traditional speech-enhancement methods. Meanwhile, Xia and Bao proposed an integrated weighted DAE and noise-classification framework for noise reduction; the results in that study confirm the effectiveness of specifying weights for each feature dimension when training the DDAE model (Xia and Bao, 2014). Then, Lai et al. proposed the adoption of the DDAE model to improve the intelligibility of vocoded speech in cochlear implant (CI) simulations (Lai et al., 2017) and of real CI recipients (Lai et al., 2018). More recently, a new loss function was derived to incorporate the perceptual speech quality during DAE training to overcome the difficulty of speech enhancement at low SNR conditions (Shivakumar and Georgiou, 2016).

In this study, we adopt the DDAE approach to transform speech recorded by BCM to match that recorded by ACM. In addition, the speech intelligibility index (SII) weighted function is utilized to further improve the transformed speech in order to achieve better ASR results. Experimental results demonstrated that our proposed method can effectively improve the ASR performance by a notable relative CER reduction of 48.28% (from 14.50% to 7.50%) under quiet conditions. When tested in an actual noisy environment (sound pressure ranging from 61.7 dBA to 73.9 dBA), our proposed system enabled the BCM to outperform the ACM, yielding a relative CER reduction of 84.46% (our proposed system: 9.13% and ACM: 58.75%).

The rest of this paper is organized as follows: Section 2 presents our proposed bone-conducted speech enhancement approach, and Section 3 reports the experimental setup and results. Section 4 discusses the results and future directions, while Section 5 concludes the paper.

2. Proposed method

Our proposed bone-conducted speech-enhancement approach has two phases: the training phase and the testing phase, as shown in Fig. 1(a) and (b), respectively. Below, we present details about the procedure for these two phases.

2.1. Feature extraction

This section introduces the feature-extraction procedure that is used in our proposed bone-conducted speech-enhancement system. There are four integral stages:

(1) A speech waveform is first segmented into a series of frames, each of which consists of L samples, and has R overlapping with the next frame.

(2) Each frame is multiplied with a Hamming window to make the first and the last few points of the frame continuous. The coefficients of the Hamming window are computed by

$$w(l) = 0.54 - 0.46\cos\left(2\pi\frac{l}{L}\right), \quad 0 \le l \le L. \tag{1}$$

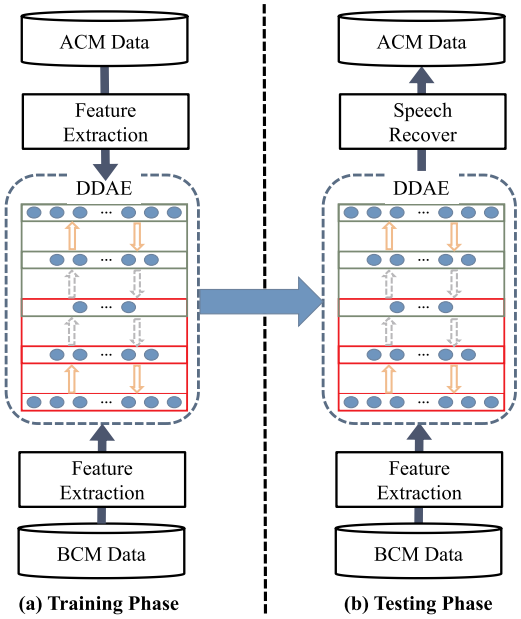

Fig. 1. Architecture of our proposed bone-conducted speech-enhancement system.

(3) The fast Fourier transform (FFT) is applied to the speech samples in each frame to transform data from the time domain to the spectral domain. We then obtain amplitude information by taking the square root of the spectral features, where the tth amplitude features for the BCM and ACM are denoted as $\boldsymbol{Y}_t^F$ and $\boldsymbol{X}_t^F$, respectively, and $\boldsymbol{Y}_t^F = [Y_t^F(1), \ldots Y_t^F(K)]'$ and $\boldsymbol{X}_t^F = [X_t^F(1), \ldots X_t^F(K)]'$ where K denotes the number of frequency bins. Next, we convert the amplitude feature into the Mel scale, which is defined based on the perceptual scale of pitches. The formula that is used to convert the kth hertz into the mth Mel is:

$$m = 2595 log_{10}\left(1 + \frac{k}{700}\right) \tag{2}$$

Based on Eq. (2), we further adopt a set of triangular filters on the Mel-scale features, and the generated data are called Fbank features. In the following discussion, we denote the tth Fbank features for the BCM and ACM as $\boldsymbol{Y}_t^{Mel}$ and $\boldsymbol{X}_t^{Mel}$, respectively, and $\boldsymbol{Y}_t^{Mel} = [Y_t^{Mel}(1), \ldots Y_t^{Mel}(M)]'$ and $\boldsymbol{X}_t^{Mel} = [X_t^{Mel}(1), \ldots X_t^{Mel}(M)]'$, where M denotes the number of Mel filter banks.

(4) We take the logarithm for each element of the Fbank features, and then apply the mean and variance normalization to the log-Fbank features. Moreover, we cascade the features that span multiple consecutive frames to form the input feature in order to incorporate the context information. The final input feature is: $\boldsymbol{Y}_t^E = [\log(Y_{t-\tau}^{Mel}(1)) \ldots \log(Y_{t-\tau}^{Mel}(M)) \ldots \log(Y_t^{Mel}(1)) \ldots \log(Y_t^{Mel}(M)) \ldots \log(Y_{t+\tau}^{Mel}(1)) \ldots \log(Y_{t+\tau}^{Mel}(M))]$, where τ is the length that is used to characterize the context information, and the output feature is $\boldsymbol{X}_t^E = [\log(X_t^{Mel}(1)) \ldots \log(X_t^{Mel}(M))]'$ for the DDAE model.

2.2. Training DDAE

A DDAE model consists of (1) one input layer: bone-conducted speech features, (2) multiple intermediate layers: layers of hidden neurons, and (3) one output layer: air-conducted speech features. For a DDAE model with J hidden layers, we have

$$
\begin{aligned}
h^1(\boldsymbol{Y}_t^E) &= g(\boldsymbol{W}^0\boldsymbol{Y}_t^E + \boldsymbol{B}^0),\\
&\vdots\\
h^J(\boldsymbol{Y}_t^E) &= g(\boldsymbol{W}^{(J-1)}h^{(J-1)}(\boldsymbol{Y}_t^E) + \boldsymbol{B}^{(J-1)}),\\
\hat{\boldsymbol{X}}_t^E &= \boldsymbol{W}^J h^J(\boldsymbol{Y}_t^E) + \boldsymbol{B}^J,
\end{aligned} \tag{3}
$$

where $\{\boldsymbol{W}^0 \ldots \boldsymbol{W}^J\}$ are the matrices of the connection weights, $\{\boldsymbol{B}^0 \ldots \boldsymbol{B}^J\}$ are the bias vectors, and $\hat{\boldsymbol{X}}_t^E$ represents the log-Fbank features of restored speech corresponding to the noisy counterpart $\boldsymbol{Y}_t^E$; $g(.)$ is an activation function, such as the sigmoid (Sigm), hyperbolic tangent (Tanh), and rectified linear unit (Relu) (Glorot et al., 2011). In this study, we adopt the sigmoid function, which can be defined as

$$
g(z) = 1/(1 + exp(-z)). \tag{4}
$$

The parameters are determined by optimizing the following objective function:

$$
\begin{aligned}
\theta^* &= \arg\min_\theta (F(\theta) + \eta^0 \|\boldsymbol{W}^0\|_F^2 + \cdots + \eta^J \|\boldsymbol{W}^J\|_F^2),\\
F(\theta) &= \frac{1}{T}\Sigma_{(t=1)}^T \left\| \boldsymbol{X}_t^E - \hat{\boldsymbol{X}}_t^E \right\|_2^2,
\end{aligned} \tag{5}
$$

where $\theta = \{\boldsymbol{W}^0 \ldots \boldsymbol{W}^J; \boldsymbol{B}^0 \ldots \boldsymbol{B}^J\}$ is the parameter set of the DDAE model, and T is the total number of training samples. In Eq. (5), $\{\eta^0 \ldots \eta^J\}$ controls the tradeoff between the reconstruction accuracy and the regularization of the weighting coefficients (we set $\eta^0 = \ldots = \eta^J = 0.0002$ in this study), and $\|.\|_F^2$ denotes the Frobenius norm. Eq. (5) can be optimized using any unconstrained optimization algorithm. In this study, we used a Hessian-free algorithm (Martens, 2010) to compute the parameters, θ.

2.3. Testing phase

The procedure of the testing phase is presented in Fig. 1(b). The DDAE model (prepared in the training phase) transforms the bone-conducted speech into enhanced speech. The output of the DDAE model is applied to a speech-recovery stage in order to convert spectral speech features to time-domain waveforms. The speech-recovery stage includes the following four steps, which are basically the reverse operations of those used in the feature-extraction stage: (1) the mean and variance de-normalizations are applied to process the output of the DDAE model; (2) the exponential transform is applied to the de-normalized features; (3) the Mel-to-spectrum transform is applied to obtain the amplitude features; and (4) the inverse fast Fourier transform (IFFT) and overlap-and-add operations are carried out to convert the spectral-domain features to time-domain waveforms.

2.4. Speech intelligibility index (SII-) based IE-ACM post-filter

Table 1 lists the critical band importance function (BIF), which refers to the American National Standards Institute (ANSI) S3.5: Higher BIF weights indicate more contributions to speech intelligibility for humans (ANSI, 1997). In total, there are 21 frequency bands, and the BIF score for each frequency band indicates its effect on the intelligibility. Based on Table 1, we design the IE filter, $V(\cdot)$, using:

$$
\begin{aligned}
\tilde{X}_t^F(k) &= V(X_t^F(k))\\
&= \sqrt{\frac{\sum_{k=1}^K \sum_{t=1}^T (X_t^F(k))^2}{\sum_{k=1}^K \sum_{t=1}^T (\rho(k) X_t^F(k))^2}}\, \rho(k) X_t^F(k),
\end{aligned} \tag{6}
$$

with

$$
\rho(k) = \beta^{(c)} \text{for} k \in \{f_L^{(c)}, f_H^{(c)}\} \tag{7}
$$

where T is the total number of frames in an utterance, $\rho(k)$ is the filter coefficient, $\tilde{X}_t^F(k)$ is the final output speech processed by IE, $f_L^{(c)}$ and $f_H^{(c)}$ are the lower and higher bounds of the cth frequency band, respectively; $\beta^{(c)}$ denotes the weights assigned to the cth frequency band.

The IE filter can be placed next to the DDAE process as a post-filter to further improve the intelligibility of the DDAE enhanced speech. With that setup, a two-stage processing step is required, thus increasing the online processing efforts. In this study, we propose to directly apply the IE filter in order to convert the clean training waveforms; the IE filtered waveforms are then used as the target in the DDAE training process. During testing, the DDAE model directly generates enhanced speech by incorporating the IE filtering effect. With this setup, we do not need an additional post-filtering process.

3. Experiment

3.1. Experimental setup

We recorded 320 utterances using a synchronized ACM and BCM recording system. The scripts that were used for recording were based on the Taiwan Mandarin hearing in noise test (Taiwan MHINT) (Huang, 2005), which contained 16 lists, each of which included 20 sentences. Each sentence consisted of 10 Chinese characters, and was designed to have similar phonemic characteristics among lists. The utterances were pronounced by a male native Chinese speaker in a quiet room, and recorded at a sampling rate of 44.1 KHz, which was further down-sampled to 16 KHz. Of these 320 utterances, 240 utterances were selected as the training data, and the remaining 80 utterances were used for testing. The DDAE model used in this study had three hidden layers, each of which included 300 neurons. We used 80 Mel filter banks to prepare the input and output features, with a context window of 11 frames. Thus, the overall DDAE structure was {880*300*300*300*80}. In the following discussion, the testing results that utilize speech processed by DDAE, the IE post-filter, and DDAE with the IE post-filter are denoted as DDAE, IE-ACM, and IE-DDAE, respectively.

3.2. Experimental results

In this section, we present the experimental results, which include three parts: (1) spectrogram and amplitude envelop comparisons, (2) objective speech quality and intelligibility evaluations, and (3) ASR performance.

3.2.1. Comparison of spectrogram and amplitude envelop

First, we qualitatively investigated the effect of DDAE and IE-DDAE by using spectrogram plots. A spectrogram plot is a popular tool to analyze the temporal-spectral characteristics of speech signals (Flanagan, 2013; Haykin, 1995). The x-axis and y-axis of a spectrogram

Table 1
Critical band importance function, reference to ANSI S3.5 (ANSI, 1997).

Frequency bands (Hz)	100–200	200–300	300–400	400–510	510–630	630–770	770–920
BIF	0.010	0.026	0.041	0.057	0.057	0.057	0.057
Frequency bands (Hz)	920–1080	1080–1270	1270–1480	1480–1720	1720–2000	2000–2320	2320–2700
BIF	0.057	0.057	0.057	0.057	0.057	0.057	0.057
Frequency bands (Hz)	2700–3150	3150–3700	3700–4400	4400–5300	5300–6400	6400–7700	7700–9500
BIF	0.057	0.057	0.057	0.046	0.034	0.023	0.011

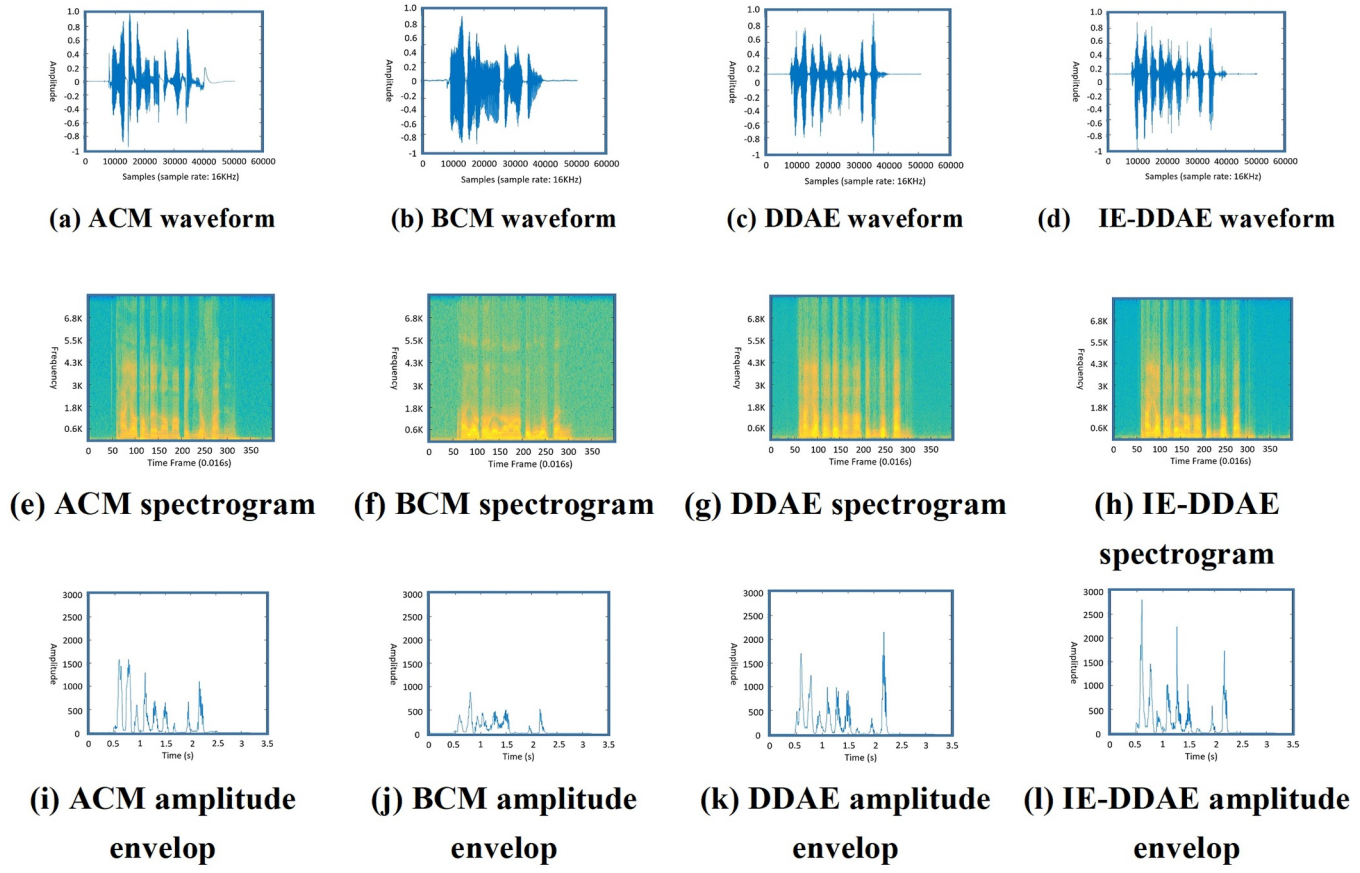

Fig. 2. Speech waveforms, spectrograms, and amplitude envelops of ACM, BCM, DDAE, and IE-DDAE.

represent the time index and frequency bins, respectively, with the intensity represented by color (red corresponds to high intensities, and blue corresponds to low intensities). For the speech of ACM, BCM, DDAE, and IE-DDAE, the waveforms are presented in Fig. 2(a), (b), (c), and (d), respectively, and the corresponding spectrograms are presented in Fig. 2(e), (f), (g), and (h), respectively. From the figures, we can note that although the differences between the ACM, BCM, DDAE, and IE-DDAE waveforms are not significant, the spectrograms show quite different characteristics, including: (1) power in the middle- and high-frequency parts are weaker in speech recorded by BCM, suggesting that the high-frequency components (when compared with low-frequency components) are decayed by the nature of the solid vibration. (2) DDAE can effectively transform the speech, so that the difference between ACM and DDAE spectrograms are considerably smaller than that between ACM and BCM.

We adopted another qualitative analysis tool, i.e., amplitude envelope plots, to compare the speeches of ACM, BCM, DDAE, and IE-DDAE. Previous studies showed that the modulation depth of an amplitude envelope is closely related to human-speech perception (van Hoesel et al., 2005, Lai et al., 2015). A higher modulation depth corresponds to better speech intelligibility. Here, we followed a previous study (Lai et al., 2015) and applied the eight-channel tone vocoder to extract amplitude envelopes. In (ANSI, 1997), it was reported that the middle-frequency band is relatively important for speech intelligibility; therefore, in this study, the amplitude envelopes of the fifth channel (1158–1790 Hz) were adopted for comparison purposes. Fig. 2(i), (j), (k), and (l) show the amplitude envelops of ACM, BCM, DDAE, and IE-DDAE, respectively. From these plots, we can see that the amplitude envelopes of DDAE and IE-DDAE have larger modulation depth values than BCM, and resemble that of ACM, indicating that the speech intelligibilities of DDAE and IE-DDAE are better than that of BCM, and approach that of ACM.

3.2.2. Speech quality and intelligibility

Next, we use two standardized objective evaluation matrices, namely the perceptual evaluation of speech quality (PESQ) (Loizou, 2007) and short-time objective intelligibility (STOI) scores (Taal et al., 2011), respectively, to measure the speech quality and speech intelligibility of BCM, DDAE, and IE-DDAE. The PESQ score ranges from −0.5 to 4.5, which corresponds to low to high speech quality, and the STOI score ranges from 0.0 to 1.0, which corresponds to low to high speech intelligibility. To compute the PESQ and STOI results, the ACM and IE-ACM speech utterances are used as the references, and the results are shown in Table 2. In the table, BCM/ACM and DDAE/ACM denote the results obtained when using ACM as the reference, and BCM and DDAE are the target results to be compared, respectively. Similarly, BCM/(IE-ACM) and (IE-DDAE)/(IE-ACM) denote the results obtained when using IE-ACM as the reference, and BCM and IE-DDAE are the respective target results to be compared. From Table 2, we can note that our proposed DDAE and IE-DDAE methods could improve PESQ scores by 9.09% [= (2.6920–2.4678)/2.4678] and 9.38% [= (2.6420–2.4155)/2.4155] and STOI scores with 14.52% [= (0.8313–0.7259)/0.7259] and 16.31% [= (0.8149–0.7006)/0.7006], respectively.

3.2.3. Automatic speech recognition

We also tested the ASR performance by employing the speech of

Table 2
PESQ and STOI results for different cases.

	BCM/ACM	DDAE/ACM	BCM/(IE-ACM)	(IE-DDAE)/(IE-ACM)
PESQ	2.4678	2.6920	2.4155	2.6420
STOI	0.7259	0.8313	0.7006	0.8149

Table 3
CERs for different cases.

	ACM	IE-ACM	BCM	DDAE	IE-DDAE
CER	1.00%	1.00%	14.50%	8.38%	7.50%

ACM, IE-ACM, BCM, DDAE, and IE-DDAE using Google ASR (Google, 2017). The results are reported with respect to the character error rate (CER), and the average CERs of 80 testing utterances of these five approaches are reported in Table 3. From the table, we note that the CER results of ACM and IE-ACM are low (both achieve 1.00% CER), indicating that the ACM used in this study can exhibit good ASR performance in noise-free conditions. However, the CER result of BCM is 14.50%, indicating that the speech recorded by BCM has a serious mismatch when compared with those used in Google ASR. We can also note that from the results of DDAE and IE-DDAE enhancement for BCM, the ASR performance can be significantly improved by CER reductions of 42.21% [= (14.50–8.38)/14.50] and 48.28% [= (14.50–7.50)/14.50], indicating that the use of DDAE can effectively reduce the mismatch between BCM and speech data used in Google ASR.

Next, we tested the ASR performance with the ACM speech under noisy conditions. We simulated noisy speech signals by adding two-men-talking (2T) noise to ACM speech at 0 dB, 10 dB, and 20 dB SNR levels. The same Google ASR was used to test recognition, and the results for over 80 testing utterances are listed in Table 4.

From Table 4, we can note that the ASR performance is severely degraded when noise is involved. When comparing the results of Tables 3 and 4, we observe that although BCM performs worse than ACM in the 20 dB SNR condition, the ASR performance can be significantly enhanced by IE-DDAE.

From the results in Tables 3 and 4, we can confirm the effectiveness of using DDAE and IE-DDAE to transform BCM and ACM in order to obtain better ASR results using simulated noisy data (the noisy speech utterances were prepared by contaminating clean speech with noise signals at specific SNR levels). In this section, we further investigate the ASR performance of IE-DDAE with speech data recorded in an actual noisy environment. To this end, we recorded another 80 testing utterances with ACM and BCM using the same scripts as those that were used in the previous experiments. The 2T noise was played by a loud speaker during recording. The sound-pressure level without playing 2T noise was about 35.3 dBA. When playing 2T noise, the sound-pressure level ranged from 61.7 dBA to 73.9 dBA. The results obtained for ACM and BCM with IE-DDAE are given in Table 5.

From Table 5, we first note that the recognition performance of ACM in a real noisy environment degraded significantly (from 1.00% in Table 3 to 58.75% in Table 5). Based on the results in Tables 4 and 5, we can infer that the SNR noise level of this real-world environment is lower than 10 dB. However, we observed that the CER of IE-DDAE (9.13%) is considerably lower than ACM (58.75%) under noisy conditions. The above experiments confirm that our proposed IE-DDAE method could bridge the channel mismatch issue between BCM and ACM, and the speech processed by IE-DDAE can yield better speech-recognition results than ACM in noisy conditions.

In the previous experiments, we tested recognition on the proposed IE-DDAE using training and testing data recorded from the same speaker. The promising results indicate that the IE-DDAE system can operate well in personal devices, such as smartphones, in-car voice control systems, and personal computers. Although our proposed system can achieve satisfactory performance in such user-specific scenarios, we are interested in investigating whether the IE-DDAE system can still perform well when the test utterances are from a different speaker. To this end, we recorded 320 sets of paired ACM and BCM speech data from another two male speakers using the same scripts as those used in the previous experiments. We used 240 utterances of one speaker to train the IE-DDAE model. Another 80 utterances from the second speaker were used to form the testing set. Note that different scripts were used to prepare the 240 training and 80 testing utterances. The average CER results are listed in Table 6.

Table 4
CERs in ACM with different noise levels (noise type: 2T).

	Clean	20 dB	10 dB	0 dB
CER	1.00%	13.88%	39.25%	99.13%

Table 5
CERs of ACM and BCM with IE-DDAE in a real noisy environment.

	ACM	BCM with IE-DDAE
CER	58.75%	9.13%

Table 6
CERs of established training model with a different speaker.

	BCM	BCM with IE-DDAE
CER	13.25%	31.63%

From Table 6, the BCM with IE-DDAE performs worse than BCM, indicating that the IE-DDAE model cannot perform well with a different speaker. In the literature, the limited generalization capability is a common issue in many speech-related applications, such as ASR (Li et al., 2014; Huang et al., 2015), speaker recognition (Kolbœk et al., 2016), and lip-reading (Wand and Schmidhuber, 2017) systems. In the past, many model-adaptation approaches have been proposed to address the training-testing mismatch issue. More recently, the domain-adversarial training criterion has been used for model adaptation, and has been proven to provide satisfactory performance in many tasks (Meng et al., 2018; Wang et al., 2018; Wand and Schmidhuber, 2017). In the future, we will also explore the adoption of the domain-adversarial training criterion to adapt our proposed bone-conducted speech-enhancement system in order to increase its applicability.

4. Discussion

Based on the qualitative analysis results obtained, our proposed IE-DDAE method effectively enhances the high-frequency components and notably increases the modulation depth of the BCM speech. Based on the quantitative analysis results, our proposed approach yields relative improvements of 9.38% and 16.31% in terms of PESQ and STOI, respectively, when compared to ACM speech, confirming the effectiveness of the proposed approach for improving speech quality and intelligibility. Finally, from the ASR results, we observed that the performance of ACM speech is significantly reduced under noisy conditions, while BCM can maintain similar recognition performance. Moreover, our proposed IE-DDAE provides notable CER reductions of 14.50% (BCM) to 7.50% (BCM + IE-DDAE), indicating that our proposed method can bridge the acoustic mismatch of BCM speech and the speech data used to train ASR.

In addition to using DDAEs (Lu et al., 20132014) and deep neural networks (DNNs) (Wang and Wang, 2012; Wang and Chen, 2017; Wang et al., 2014; Xu et al., 2014), many neural network models have been developed and used for speech-enhancement applications. Successful approaches include the extreme learning machine (Hussain et al., 2017; Odelowo and Anderson, 2017), recurrent neural networks (RNNs) (Erdogan et al., 2015), long short-term memory (Chen et al., 2015; Weninger et al., 2015; Sun et al., 2017), convolutional neural networks (Fu et al., 2016; Fu et al., 2017), and fully convolutional networks (Fu et al., 2018). Based on the structures of these models, the enhancement can be performed in raw-waveform, (complex) spectrogram, or

logarithm amplitudes. Meanwhile, different objective functions have been derived to optimize the parameters, such as those based on the minimum mean-square error (MMSE) (Xu et al., 2015), maximum-likelihood (Chai et al., 2017), and maximizing STOI score (Fu et al., 2018), as well as those based on the generative adversarial network training criterion (Santiago et al., 2017; Daniel and Tan, 2017; Donahue et al., 2017; Mimura et al., 2017). In this paper, the DDAE model is an encoder-decoder structure, and the model parameters are estimated based on the MMSE criterion. In the future, we will explore the adoption of advanced deep-learning models and learning criteria for the BCM speech-enhancement task. Moreover, in this study, the phase information of the BCM speech was used for the enhanced speech. This approach may not be perfect because the BCM speech does not cover high-frequency regions, and the phase may therefore not be optimal for use in the enhanced speech. Recently, some approaches have been proposed to perform speech enhancement in the waveform domain (Fu et al., 2018). We will also explore to perform BCM speech enhancement using these approaches to address the imperfect phase-estimation issue.

5. Conclusion

In this paper, we propose a DDAE approach to perform BCM speech enhancement. An SII-based post-filter is also derived to improve the speech intelligibility of enhanced speech. The major contributions of this study are as follows. (1) we proposed an IE-DDAE algorithm that further improves speech intelligibility over DDAE alone by incorporating the information from SII, and (2) we confirmed that based on Google ASR, our proposed algorithm can enhance BCM speech to achieve higher recognition accuracies. The above two contributions enable our proposed algorithm to serve as a simple and effective solution to bridging the acoustic mismatch of BCM and speech recognizers. In the future, we will attempt to reduce the computation cost of the IE-DDAE algorithm to further improve its applicability. Moreover, we will investigate potential model adaptation approaches to address the training-testing mismatch issue.

Acknowledgments

This research was supported by the Ministry of Science and Technology of Taiwan, R.O.C., under grants MOST 104-2221-E-002-133-MY2 and MOST 106-2221-E-002-220, and by Test Research, Jorjin Technologies, III, Egistec, D8AI, Lite-on, and NeoVictory Technology Co., Ltd.

References

ANSI, 1997. American National Standard: Methods for Calculation of the Speech Intelligibility Index: Acoustical Society of America.
Chen, Z., Watanabe, S., Erdogan, H., Hershey, J.R., 2015. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In: Proceedings of the Interspeech, pp. 3274–3278.
Chai, L., Du, J., Wang, Y.-N., 2017. Gaussian density guided deep neural network for single-channel speech enhancement. In: IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), 2017. IEEE.
Daniel, M., Tan, Z.-H., 2017. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In: Proceedings of the Interspeech, pp. 2008–2012.
Dekens, T., Verhelst, W., Capman, F., Beaugendre, F., 2010. Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection. In: Proceedings of the EUSIPCO, pp. 1978–1982.
Donahue, C., Li, B., Prabhavalkar, R., 2017. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In: . arXiv:1711.05747.
Erdogan, H., Hershey, J.R., Watanabe, S., Le Roux, J., 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: Proceedings of the ICASSP, pp. 708–712.
Flanagan, J.L., 2013. Speech Analysis Synthesis and Perception, Ed. 3. Springer Science and Business Media, Germany Berlin.
Fu, S.W., Tsao, Y., Lu, X., 2016. SNR-aware convolutional neural network modeling for speech enhancement. In: Proceedings of the Interspeech, pp. 3768–3772.
Fu, S.W., Hu, T.Y., Tsao, Y., Lu, X., 2017. Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: Proceedings of the MLSP.
Fu, S.W., Wang, T.W., Tsao, Y., Lu, X., Kawai, H., 2018. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 26 (9), 1570–1584.
Google 2017. Cloud Speech API, https://cloud.google.com/speech/.
Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: Proceedings of the AISTATS, pp. 315–323.
Graciarena, M., Franco, H., Sonmez, K., Bratt, H., 2003. Combining standard and throat microphones for robust speech recognition. IEEE Signal Process. Lett. 10 (3), 72–74.
Haykin, S., 1995. Advances in Spectrum Analysis and Array Processing, 3. Prentice-Hall, NJ Upper Saddle River.
Huang, M.W., 2005. Development of Taiwan Mandarin hearing in noise test. Master thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health Sciences.
Huang, Z., Li, J., Siniscalchi, S.M., Chen, I.F., Wu, J., Lee, C.H., 2015. Rapid adaptation for deep neural networks through multi-task learning. In: Proceedings of the Interspeech 2015.
Hussain, T., Siniscalchi, S.M., Lee, C.C., Wang, S.S., Tsao, Y., Liao, W.H., 2017. Experimental study on extreme learning machine applications for speech enhancement. IEEE Access 5, 25542–25554.
Kolbœk, M., Tan, Z.H., Jensen, J., 2016. Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification. In: Proceedings of the SLT. 2016. pp. 305–311.
Kuo, H.H., Yu, Y.Y., Yan, J.J., 2015. The bone conduction microphone parameter measurement architecture and its speech recognition performance analysis. In: Proceedings of the JIMET, pp. 137–140.
Lai, Y.H., et al., 2015. Effects of adaptation rate and noise suppression on the intelligibility of compressed-envelope based speech. PloS One 10, e0133519.
Lai, Y.H., et al., 2017. A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation. IEEE Trans. Biomed. Eng. 64 (7), 1568–1578.
Lai, Y.H., et al., 2018. Deep learning-based noise reduction approach to improve speech intelligibility for cochlear implant recipients. Ear Hear.
Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (4), 745–777.
Liu, Z., Zhang, Z., Acero, A., Droppo, J., Huang, X.D., 2004. Direct filtering for air- and bone-conductive microphones. In: Proceedings of the MMSP, pp. 363–366.
Loizou, P.C., 2007. Speech Enhancement: Theory and Practice. CRC Press, Florida Boca Raton.
Lu, X., Tsao, Y., Matsuda, S., Hori, C., 2014. Ensemble modeling of denoising autoencoder for speech spectrum restoration. In: Proceedings of the Interspeech, pp. 885–889.
Lu, X., Tsao, Y., Matsuda, S., Hori, C., 2013. Speech enhancement based on deep denoising autoencoder. In: Proceedings of the Interspeech, pp. 436–440.
Martens, J., 2010. Deep learning via Hessian-free optimization. In: Proceedings of the ICML, pp. 735–742.
Meng, Z., Li, J., Gong, Y., Juang, B.-H., 2018. Adversarial teacher-student learning for unsupervised domain adaptation. In: Proceedings of the ICASSP.
Mimura, M., Sakai, S., Kawahara, T., 2017. Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks. In: Proceedings of the ASRU.
Odelowo, B.O., Anderson, D.V., 2017. Speech enhancement using extreme learning machines. In: Proceedings of the WASPAA, pp. 200–204.
Santiago, P., Antonio, B., Joan, S., 2017. SEGAN: Speech enhancement generative adversarial network. In: Interspeech, pp. 3642–3646.
Shimamura, T., Tomikura, T., 2005. Quality improvement of bone-conducted speech. In: Proceedings of the ECCTD, pp. 1–4.
Shimamura, T., Mamiya, J., Tamiya, T., 2006. Improving bone-conducted speech quality via neural network. In: Proceedings of the ISSPIT, pp. 628–632.
Shivakumar, P.G., Georgiou, P.G., 2016. Perception optimized deep denoising autoencoders for speech enhancement. In: Proceedings of the Interspeech, pp. 3743–3747.
Sun, L., Du, J., Dai, L.-R., Lee, C.-H., 2017. Multiple-target deep learning for LSTM-RNN based speech enhancement. In: Proceedings of the HSCMA, pp. 136–140.
Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136.
Tajiri, Y., Kameoka, H., Toda, T., 2017. A noise suppression method for body-conducted soft speech based on non-negative tensor factorization of air- and body-conducted signals. In: Proceeding of the ICASSP, pp. 4960–4964.
Thang, T.V., Kimura, K., Unoki, M., Akagi, M., 2006. A study on restoration of bone-conducted speech with MTF-based and LP-based models. J. Signal Process. 407–417.
van Hoesel, R., et al., 2005. Amplitude-mapping effects on speech intelligibility with unilateral and bilateral cochlear implants. Ear Hear. 26, 381–388.
Wand, M., Schmidhuber, J., 2017. Improving Speaker-Independent Lipreading with Domain-Adversarial Training. arXiv:1708.01565.
Wang, D., Chen, J., 2017. Supervised Speech Separation Based on Deep Learning: An Overview. arXiv:1708.07524.
Wang, Q., Rao, W., Sun, S., Xie, L., Chng, E.S., Li, H., 2018. Unsupervised domain adaptation via domain adversarial training for speaker recognition. In: Proceedings of the ICASSP..
Wang, Y., Wang, D., 2012. Cocktail party processing via structured prediction. In: Proceedings of the NIPS, pp. 224–232.
Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process 22 (12), 1849–1858.
Weninger, F., Erdogan, H., Watanabe, S., et al., 2015. , Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: Proceedings of the LVA/ICA, pp. 91–99.

Xia, B., Bao, C., 2014. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Commun. 60, 13–29.
Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2014. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. 21 (1), 65–68.
Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2015. A regression approach to speech enhancement based on deep neural networks,. IEEE/ACM Trans. Audio Speech Lang. Process 23, 7–19.
Zhang, Z., Liu, Z., Sinclair, M., Acero, A., Deng, L., Droppo, J., Huang, X.D., Zheng, Y., 2004. Multi-sensory microphones for robust speech detection, enhancement, and recognition. In: Proceedings of the ICASSP, pp. 781–784.
Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., Acero, A., Huang, X.D., 2003. Air- and bone-conductive inte-grated microphones for robust speech detection and enhancement. In: Proceedings of the ASRU, pp. 249–254.