LETTER

Spectra Restoration of Bone-Conducted Speech via Attention-Based Contextual Information and Spectro-Temporal Structure Constraint

Changyan ZHENG[†*], **Tieyong CAO**[†*], **Jibin YANG**[†], **Xiongwei ZHANG**[†a)], ***Nonmembers***, ***and*** **Meng SUN**[†], ***Member***

SUMMARY Compared with acoustic microphone (AM) speech, bone-conducted microphone (BCM) speech is much immune to background noise, but suffers from severe loss of information due to the characteristics of the human-body transmission channel. In this letter, a new method for the speaker-dependent BCM speech enhancement is proposed, in which we focus our attention on the spectra restoration of the distorted speech. In order to better infer the missing components, an attention-based bidirectional Long Short-Term Memory (AB-BLSTM) is designed to optimize the use of contextual information to model the relationship between the spectra of BCM speech and its corresponding clean AM speech. Meanwhile, a structural error metric, Structural SIMilarity (SSIM) metric, originated from image processing is proposed to be the loss function, which provides the constraint of the spectro-temporal structures in recovering of the spectra. Experiments demonstrate that compared with approaches based on conventional DNN and mean square error (MSE), the proposed method can better recover the missing phonemes and obtain spectra with spectro-temporal structure more similar to the target one, which leads to great improvement on objective metrics.

key words: *bone-conducted microphone, speech enhancement, bidirectional long short-term memory, attention, Structural SIMilarity*

1. Introduction

Bone-conducted microphone (BCM) is a kind of skin-attached non-audible sensor and converts the vibration of the human body like throat and skull into electrical signal [1]. It is immune to ambient noise and can transmit speech signal even under severe environments, such as military field, air-craft and F1 racing, etc [2]. However, BCM speech does not sound natural and clear like conventional acoustic microphone (AM) speech. Due to the attenuation of human body channel, it faces severe loss of high-frequency components that are usually higher than 2 kHz [2]. Besides, some phonemes like unvoiced fricatives and plosives are totally lost, which are generated in the oral or nasal cavity rather than the vocal cord.

BCM is often used to improve the speech communication quality in noisy environments. In most cases, it plays an auxiliary role for improving the enhancement performance of AM speech. For example, in [3], BCM speech is utilized to estimate the speech present probability, and in [4], it is used to help trace the pitch. Other representative works include [5]–[9]. AM speech is indispensable in this kind of algorithms, but it is meaningful to enhance BCM speech directly, because AM speech can be completely unintelligible and become useless in some occasions.

Compared with fusion-based method, direct enhancement of BCM speech suffers from less information and is more challenging. The key of direct enhancement of BCM speech is to find the mapping relationship between the transmission channel functions of AM and BCM speech, because BCM speech can pick up the vibration of the glottis clearly and its excitation source can be approximately assumed to be the same as AM speech. In early algorithms [10]–[13], transmission channel functions were represented by low dimensional spectral envelope features, Gaussian Mixture Models (GMM) and shallow neural networks were often employed to learn the mapping relationship. Recently, deep neural networks (DNN) have been used to learn the complex nonlinear mapping relationship [14], [15], and some researchers start to use high-dimensional features to represent the difference of the two speech. For instance, deep denoising autoencoder is used to map the high dimensional Mel magnitude spectra of the two speech and achieves considerable improvements [16].

However, all of the methods mentioned above actually model the frame-based mapping relationship, even if multiple feature frames are concatenated as input like [16]. They can establish the correlation between the low-frequency and high-frequency components in the spectra, and are capable of recovering the lost high-frequency components from the spectra of BCM speech, but have little ability to model the sequential relationship required for inferring the lost phonemes. It is similar to a blank-filling game, in which human beings need the contextual information to guess the missing words reasonably.

In our previous work [17], we explore a BCM speech enhancement method based on bidirectional Long Short-Term Memory (BLSTM) [18], which can model the long-span context-dependencies of speech sequence effectively and benefit the inferring of missing phonemes greatly. Nevertheless, BLSTM sometimes fails to recover enough energy of lost phonemes as expected. It is likely that the restoration of missing phonemes require more attention since the extreme lack of necessary information. In this letter, in or-

Manuscript received September 5, 2019.

[†]The authors are with Army Engineering University, Nanjing, China.

[*]Authors contribute equally.

a) E-mail: xwzhang9898@163.com (Corresponding author)

DOI: 10.1587/transfun.E102.A.2001

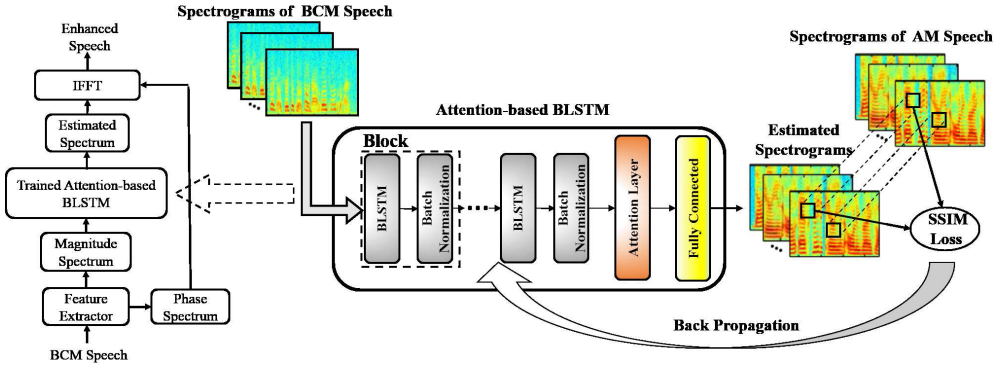

Fig. 1 The framework of the proposed method.

der to learn the importance and irrelevance of sequential information actively, we introduce the attention mechanism [19] in our model and design an attention-based BLSTM (AB-BLSTM) to further optimize the use of contextual information. Preliminary results show that the designed AB-BLSTM can achieve better results in recovering the missing components. Meanwhile, to make AB-BLSTM pay more attention on recovering the structures of harmonics, which are of great importance to human auditory system [20], a new loss function, the Structural SIMilarity (SSIM) [21] loss function is introduced to train our model. Different from traditional mean square error (MSE) loss function, it fuses the structural information in the error measurement and can provide constraint of the spectro-temproal structures in the recovering of spectra. Since SSIM loss function is originated from image processing, we also analyze the effect of hyper-parameter of it and provide an optimal choice for application in spectrogram image.

The rest of this letter is organized as follows. The overall framework and the details of the proposed method is introduced in Sect. 2. A set of evaluation experiments to assess the performance of proposed method are provided in Sect. 3. Finally, we conclude the letter in Sect. 4.

2. The Proposed Method

2.1 The Overall Framework

The overall framework of the proposed method is depicted in Fig. 1. The high-dimensional Short-Time Fourier Transformed (STFT) magnitude is selected as the spectral feature. The log compression of the feature is conducted to reduce the dynamic range and global mean-variance normalization is applied to make the training amenable. For brevity, the data pre-processing is not shown in Fig. 1.

The right part of Fig. 1 shows the training stage, in which an AB-BLSTM is built as the spectra mapping model between the BCM and AM speech and is trained using the SSIM loss function. The left part of Fig. 1 demonstrates the enhancement stage. The spectrum is firstly extracted from BCM speech, and is then fed into the well trained AB-BLSTM model to estimate the spectrum of corresponding

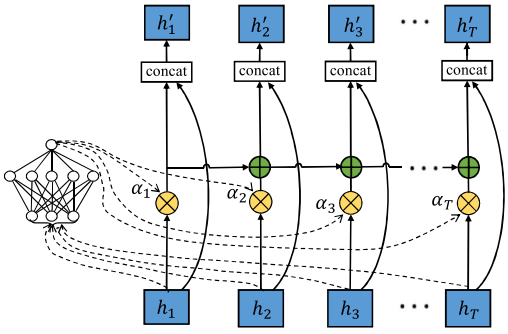

Fig. 2 The attention layer.

AM speech. The time domain waveform of the enhanced speech is finally reconstructed based on the inverse STFT of the estimated magnitude spectrum along with the original phase of the BCM speech.

2.2 The Attention Based BLSTM

The AB-BLSTM architecture we designed is composed of three blocks, one attentive layer and one fully connected layer. Each block consists of one BLSTM layer and one batch normalization (BN) [22] layer. BLSTM layer extends the unidirectional LSTM network by introducing a second layer, in which hidden connections flow in reverse time order. Therefore, the model is able to take advantage of past and future information. LSTM can overcome the gradients vanishing problem of the standard recurrent neural networks (RNN) by some some purpose-built gates. Details of BLSTM can be found in [23]. BN layer is used to alleviate the gradient dispersion problem in deep network by adjusting the deviation of the data statistical parameters.

The attention layer we designed is shown in Fig. 2, which refers to one of the state-of-the-art attention mechanism [24].

In [24], a learnable function $a(\cdot)$ defined by a feed-forward net layer is used to generate weighted values for each time state, the value represents the importance of the state. Then the weighted states are summed to form a super vector as the final output which can distinguish the important and less important information. Like [24], we acquire

the weights according to the following equation:

$$e_t = a(h_t) \tag{1}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} exp(e_k)} \tag{2}$$

where t is the current time step, e_t denotes the intermediate vector. Equation (2) is the normalization of the weights using the softmax function, k is the index of time step, T is the sequence length and α_t represents the value of the weight vectors.

Different from the classification problem which only needs a final output in [24], the enhancement task is a regression problem, and the output of the attention layer in each time state should be affected by all the previous attentive information. So we propose a super vector composed of all the previous attentive information for each time state to focus on recovering the missing components. Experimentally, we find out that it is better to concatenate the super vector with the original state vector h_t than to use either of the information alone. We infer that the super vector tends to capture local attentive context while the original hidden state is inclined to represent global context, thus concatenating them together is able to keep the information complete and achieve better performance.

Therefore, the final output is defined as following:

$$h'_t = [f(\sum_{t=1}^{k\leq t} \alpha_k h_k); h_t] \tag{3}$$

where k is the index of time step, t is the current time step. The final output h'_t is then sent to the fully connected layer to predict the frame of spectrograms.

2.3 SSIM Loss Function

In image processing, SSIM metric comprehensively takes *illumination*, *contrast* and *structure* of the **local** image patch into consideration. As the spectra of speech can be viewed as an image format [25], [26], we think SSIM metric is also suitable to be applied on the spectrograms.

Suppose two local spectrogram patches centered at the spectro-temporal point x and y, typically square patches, the three components combined together to yield an overall similarity measure and are defined respectively as following:

$$\begin{aligned} SSIM(x,y) &= L(x,y)C(x,y)S(x,y) \\ &= \frac{2\mu_x\mu_y + C_1}{{\mu_x}^2 + {\mu_y}^2 + C_1} \cdot \frac{2\delta_{xy} + C_2}{{\delta_x}^2 + {\delta_y}^2 + C_2} \end{aligned} \tag{4}$$

where μ_x, μ_y, δ_x and δ_y denote mean spectral magnitude and standard deviation of spectral magnitude in the patch, δ_{xy} is the covariance coefficient. C_1 and C_2 are small constants introduced to avoid numerical instability, and are computed as below:

$$C_1 = (k_1 L)^2, C_2 = (k_2 L)^2 \tag{5}$$

In image processing, L is the dynamic range of the pixel values (255 for 8-bit grayscale images). Since the range of the spectral magnitude is much less than 255, we set L to 7, which achieves better results in the subsequent experiments. k_1 and k_2 are predefined constants and are set as 0.01 and 0.03 respectively according to [21]. We have to point out that since SSIM metric is usually applied on non-negative signals, we should first transform the log spectral magnitude back to raw magnitude to compute it.

From (4) we can clearly note that SSIM metric is very different from MSE. The former includes the local statistics such as μ_x, δ_x and δ_{xy}, while the latter only relates with single spectro-temporal point.

A Gaussian filter with standard deviation δ_G is used to compute the means and standard deviations, which can resist the undesirable "blocking" artifacts [21]. In fact, δ_G is an important parameter which controls the size of filter, and in the following section, we will experiment the choice of this hyper-parameter. As the filter moves point-by-point over the entire spectrogram to compute the SSIM score, a map of SSIM scores based on per-point can be formed.

In the model training, our objective is to maximize the similarity between the estimated spectrogram $f_\theta(X_n)$ and target spectrogram Y_n, where f_θ represents the mapping model f with parameter θ. The SSIM loss function can be defined as following, and the parameter is updated by minimizing it:

$$J_{SSIM}(\theta) = -\frac{1}{N}\sum_{n=1}^{N} SSIM(f_\theta(X_n), Y_n) \tag{6}$$

3. Experiments and Results Analysis

3.1 Data Collection and Evaluation Metrics

The speech data is collected in an anechoic chamber, where the clean speech signal can be acquired. In order to ensure the synchronization of BCM and AM speech, a stereo sound card is used, BCM is connected to the left channel and AM is connected to the right channel. 5 male and 5 female are required to read 200 different phoneme-balanced sentences in Mandarin respectively, each of which lasts about 3-5s. The recording sampling rate is set to 32 kHz.

Perceptual Evaluation of Speech Quality (PESQ) [27] and Log-spectral Distance (LSD) [28] are employed to evaluate the performance of the proposed method objectively. PESQ score ranges from −0.5 to 4.5. It measures the overall speech quality and has high correlation with subjective evaluation scores. LSD is used to measure the spectral distortion between the referenced speech and enhanced speech. Smaller LSD score is better. For evaluating the performance of the proposed method more thoroughly, we conduct the subjective listening test and the mean opinion score (MOS) [29] is adopted.

3.2 Experimental Setup and Compared Methods

In our experiments, the 200 sentences of each speaker are

divided into 160 sentences for training and the rest for testing. Currently, we consider enhancing the BCM speech of 8 kHz sampling rate, because in telecommunications, 8 kHz is still the mainstream sampling rate. In addition, the effective spectral component of BCM speech is about 2 kHz, and it is difficult to recover the lost components to very high bands. Thus, all the speech data is firstly down-sampled to 8 kHz, and then the spectra are extracted using 256-point STFT with a hop size of 64. Since the spectrum of the real signal is symmetric, the magnitude in the first 129 frequency bins are selected as the processed features.

The overall architecture of AB-BLSTM is introduced in Sect. 3. It should be pointed out that using the non-linear function ReLU (Rectified Linear Units) [30] to compute the weighted values in attention layer can get slight better result than sigmoid function that is used in [24]. Experiments are also performed on BLSTM for comparison. In addition, the deep neural network (DNN) proposed in [15] and the deep denoising autoencoder (AE) proposed in [16] are also conducted as comparative methods. The DNN consists of 3 hidden layers, each with 1024 units, and ReLU is selected as the activation function. The AE contains 4 hidden layers, 512 hidden units per layer, and the encoder and decoder structures are symmetrical. All the neural networks are trained with MSE and SSIM loss function respectively.

An individual enhancement model is developed for each speaker. The Adam [31] algorithm with an initial global learning rate of 0.002 is used to optimize the neural networks.

3.3 Results and Discussions

(1) The Optimal Hyper-parameter δ_G

We first explore the optimal value of hyper-parameter δ_G in SSIM loss function. In this experiment, AB-BLSTM is trained using SSIM loss function with different δ_G values. The speech quality is evaluated by comparing with the referenced clean AM speech, and the PESQ and LSD results of male1 and female1 are shown in Fig. 3. The results of AB-BLSTM trained using MSE loss function is also presented as baseline.

It can be noted that when δ_G is set to 0.5, the model can get the best results on both the PESQ and LSD score. We infer that it is because when $\delta_G = 0.5$, the Gaussian kernel approximately covers about 3×3 region. One harmonic texture covers about 2 to 4 frequency bins in our spectrogram image, so if the size of filter is chosen around 2 to 4, the mapped spectra tends to preserve better structural details. Therefore, we conjecture that δ_G value should be chosen according to the characteristics of the harmonic, and if the resolution of the spectra or the sampling rate of the speech is changed, the δ_G value should be reconsidered.

When δ_G is set to 0.01, SSIM loss function is actually retreated to the Cosine Distance (CD) loss function according to (4). The results show that CD loss function performs slightly better than MSE loss function. It is interesting that when $\delta_G = 1.5$, the SSIM loss function achieves worse LSD

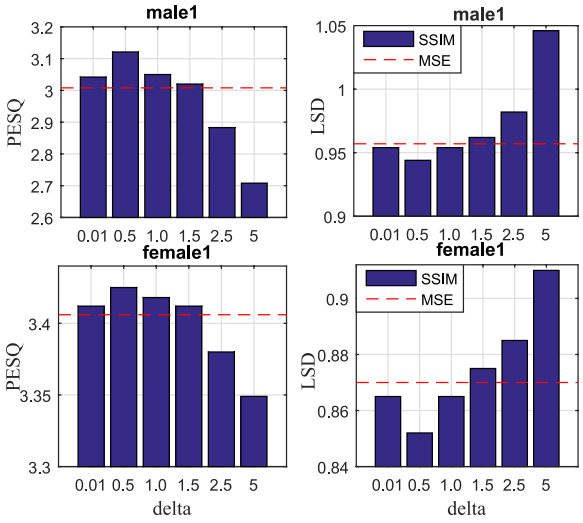

Fig. 3 Objective results of AB-BLSTM trianed using SSIM loss function with different δ_G values.

score than MSE loss function, while acquires better PESQ scores. It is possible that SSIM metric which concentrates on the constraint of the spectro-temporal patterns is more consistent with the human auditory system. The spectra distance may get worse, but the perceptual quality of speech can be still improved. With the increasing of delta values, the scores of PESQ and LSD deteriorate rapidly. It may result from the too large Gaussian filter which leads to many blurred blocks in the spectrograms.

Therefore, we can conclude that 0.5 is the optimal δ_G value in our experiment setting, and in the following experiments, it is set as the default δ_G value in SSIM loss function to train other models.

(2) Objective Evaluation of Speech Quality

The PESQ and LSD results of different methods are shown in Table 1 and Table 2 respectively.

Firstly, we evaluate the effectiveness of neural networks with different architectures and analyze the results of neural networks trained with MSE loss functions. It can be seen that the average PESQ scores of BCM speech are about 2.1 for both male and female speakers, which indicate that the original speech quality is quite low and unsatisfactory.

It can be noticed that DNN improves the average PESQ score about 0.4 and decreases the LSD score about 0.35, which correspond to 20% and 25% improvement compared with the original scores of BCM speech. The results of DNN and AE are very close. In fact, the two architectures are similar. The difference lies in the training strategy and AE is trained by greedy algorithm per layer. It may be that the application of special activation functions like ReLU and optimizing technology like dropout overcome the difficulty of DNN training, resulting in no big difference between the two architectures.

Compared to DNN and AE, we can note that BLSTM can further improve PESQ score to a considerable extent, and the average scores of male and female speakers have

Table 1 Perceptual speech quality evaluation with PESQ on BCM and enhanced speech.

Person	BCM	DNN		AE		BLSTM		AB-BLSTM	
		MSE	SSIM	MSE	SSIM	MSE	SSIM	MSE	SSIM
male1	2.277	2.719	2.802	2.723	2.810	2.913	3.056	3.008	**3.121**
male2	1.963	2.257	2.322	2.260	2.331	2.739	2.877	2.808	**2.951**
male3	1.931	2.324	2.417	2.328	2.410	2.580	2.726	2.664	**2.772**
male4	2.281	2.762	2.861	2.756	2.855	3.058	3.203	3.130	**3.253**
male5	2.102	2.403	2.489	2.418	2.492	2.661	2.840	2.752	**2.916**
Average	2.111	2.493	2.578	2.497	2.580	2.790	2.940	2.872	**3.003**
female1	2.508	3.139	3.181	3.132	3.176	3.336	3.399	3.406	**3.425**
female2	2.023	2.533	2.627	2.541	2.635	2.832	2.994	2.949	**3.077**
female3	2.078	2.469	2.561	2.463	2.565	2.716	2.811	2.762	**2.864**
female4	2.294	2.611	2.708	2.620	2.711	2.849	2.941	2.915	**3.022**
female5	1.847	2.214	2.312	2.212	2.327	2.394	2.495	2.476	**2.558**
Average	2.150	2.593	2.678	2.594	2.683	2.825	2.928	2.902	**2.989**

Table 2 Perceptual speech quality evaluation with LSD on BCM and enhanced speech.

Person	BCM	DNN		AE		BLSTM		AB-BLSTM	
		MSE	SSIM	MSE	SSIM	MSE	SSIM	MSE	SSIM
male1	1.482	1.047	1.017	1.053	1.032	0.961	0.949	0.957	**0.944**
male2	1.480	0.991	0.963	0.987	0.956	0.899	0.877	0.887	**0.871**
male3	1.455	1.061	1.035	1.062	1.031	0.961	0.936	0.951	**0.935**
male4	1.353	0.981	0.952	0.976	0.946	0.857	0.843	0.851	**0.836**
male5	1.440	1.014	0.984	1.010	0.981	0.955	0.930	0.956	**0.927**
Average	1.442	1.019	0.990	1.018	0.989	0.927	0.907	0.920	**0.903**
female1	1.369	0.912	0.889	0.911	0.906	0.869	0.856	0.870	**0.852**
female2	1.305	0.962	0.934	0.960	0.929	0.924	0.904	0.919	**0.902**
female3	1.427	1.133	1.107	1.127	1.102	1.070	0.992	1.018	**0.985**
female4	1.239	0.978	0.960	0.972	0.956	0.953	0.927	0.954	**0.926**
female5	1.389	1.047	1.002	1.042	1.004	0.995	0.966	0.995	**0.964**
Average	1.346	1.006	0.978	1.002	0.979	0.962	0.929	0.951	**0.926**

continued to increase by about 0.3 and 0.25 respectively, which proves the great advantage of BLSTM in BCM speech enhancement. However, the LSD scores decrease about 0.09 and 0.04, which is not a big improvement. It is because that BLSTM can help recover lost phonemes, such as nasal and fricative, which correspond to frequency components with very low energy, so the difference in spectral energy measured by LSD is not very noticeable. Compared with BLSTM, AB-BLSTM achieves better PESQ scores, but the LSD scores still do not decrease much. It may be because that AB-BLSTM further emphasizes the recovering of the lost low energy components, which has impact on the perceptual quality rather than spectral difference.

The results also clearly show that SSIM loss function has an obvious advantage over MSE loss function when training the same architecture, regardless of DNN, AE or BLSTM. The improvement are very stable.

Overall, compared with the original scores of BCM speech, the proposed method improves the average PESQ score about 0.89 and 0.84, and LSD score about 0.54 and 0.42 for male and female speakers respectively, corresponding to 42.3%, 39.1% and 37.4%, 31.2% improvement.

The spectrograms of an utterance are shown in Fig. 4, and it is best viewed by zooming in. As can be seen from the rectangular box in Fig. 4(a) and Fig. 4(b), frequency components above 2 kHz disappear in BCM speech, and some

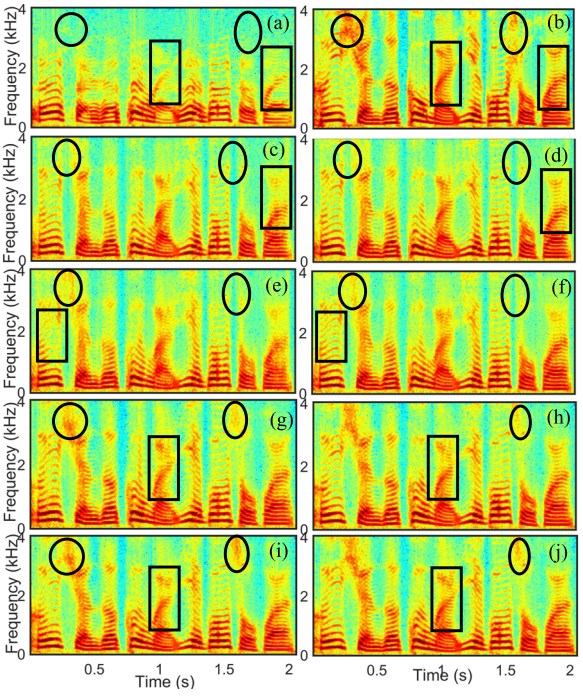

Fig. 4 Spectrograms of an utterance, (a) BCM speech, (b) AM speech, (c) DNN-MSE enhanced, (d) DNN-SSIM enhanced, (e) AE-MSE enhanced, (f) AE-SSIM enhanced, (g) BLSTM-MSE enhanced, (h) BLSTM-SSIM enhanced, (i) AB-BLSTM-MSE enhanced, (j) AB-BLSTM-SSIM enhanced.

harmonics below 2 kHz are also incomplete. In the oval and circular boxes, the missing phonemes can be noticed. In Fig. 4(c)–(f), we see that DNN and AE is able to recover considerable high-frequency components, but has the difficulty to infer the lost phonemes, which can be seen in the oval boxes. BLSTM and AB-BLSTM show great advantages in recovering the missing components, as can be seen from the oval and circular boxes in Fig. 4(g)–(j), which clearly demonstrates the importance of using contextual information in BCM speech enhancement. AB-BLSTM achieves similar spectrograms to BLSTM, but recovers more missing energy of phonemes, which can be seen in oval boxes. This indicates that the attention-layer can help strengthen the use of contextual information. The advantage of SSIM loss function can be clearly noticed when comparing the rectangular boxes from the left to the right spectrograms at the same row. In particular, SSIM loss function helps recover very tiny harmonics, as shown in Fig. 4(h) and Fig. 4(j), proving its ability to constrain the temporal-spectral structures.

(3) Subjective Listening Test

We conduct MOS tests on the proposed method AB-BLSTM-SSIM and two typical methods including DNN-MSE and BLSTM-MSE. The original BCM speech is also evaluated for comparison. Thus totally 4 kinds of samples are to be evaluated. As mentioned in the experimental setup, the test set contains 10 speakers, 40 testing utterances for each speaker. In the listening test, we select

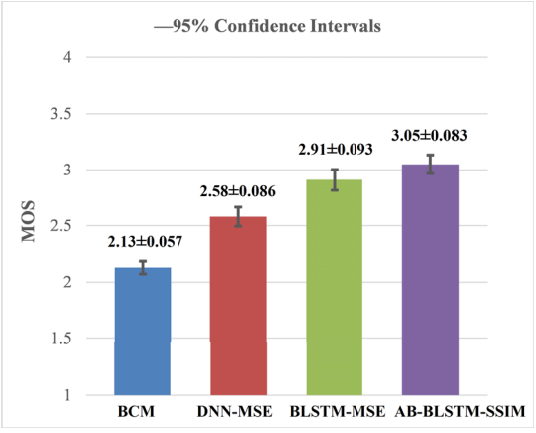

Fig. 5 MOS of different enhancement methods with 95% confidence intervals.

3 utterances from each speaker randomly, so there are total $3 \times 10 \times 4 = 120$ stimuli for each listener. We invite 5 male and 5 female native Chinese to participate in the evaluation and the participants are asked to rate the given stimuli on a scale from 1 to 5 with 1 point increments. Finally, the subjective MOS of each kind of sample is calculated. In our test, the participants listen to the stimuli over headphones in sound-treated booths and each one spends about 50 minutes completing the evaluation in average. The MOS results are shown with different colored bars in Fig. 5. The 95% confidence intervals are marked with I-bar and the specific values are also presented.

As can be seen from Fig. 5, the MOS of the BCM speech is around 2.1 and the 95% confidence interval is very small, which means the perceived quality of the BCM speech is very low and cannot meet the requirements in real applications. The AB-BLSTM-SSIM method achieves better MOS than the other two methods at a high confidence level, which demonstrates that the proposed method can also obtain better subjective speech quality. At the same time, we can see that the subjective results show a good match with the PESQ scores in Table 1.

4. Conclusion

In this letter, we propose to build an attention-based BLSTM trained with SSIM loss function to model the spectra mapping relationship for BCM speech enhancement. The proposed method can utilize attention-based long-span contextual information and provide spectro-temproal constraint on the restored spectra. Experimental results show that the proposed method is able to recover the missing components and obtain spectra with better harmonic structures, which makes great improvement in the quality of BCM speech. In the future work, we would like to address the phase-estimation problem by introducing waveform modelling method [32] and investigate the speaker adaptation technology [33], [34] for speaker-independent BCM speech enhancement to meet the requirements in real applications.

Some speech demos of different enhancement methods are presented in follwing website https://github.com/echoaimaomao/Demos-for-Attention-based-BLSTM-trained-with-SSIM-loss-function.

Acknowledgments

This work is partially supported by NSF of China (Grant No. 61471394) and NSF of Jiangsu Province for Excellent Young Scholars (Grant No. BK20180080).

References

[1] H. Ono, "Device for picking up bone-conducted sound in external auditory meatus and communication device using the same," U.S. Patent 5,295,193[P], 1994-3-15, 1994.
[2] H. Shin, G. Kang, and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphone," Proc. Speech Communication Symposium, pp.1–4, 2012.
[3] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.249–254, 2003.
[4] M.S. Rahman and T. Shimamura, "Pitch characteristics of bone conducted speech," Proc. 18th European Signal Processing Conference (EURASIP), pp.795–799, 2010.
[5] L. Deng, Z. Liu, Z. Zhang, and A. Acero, "Nonlinear information fusion in multi-sensor processing-extracting and exploiting hidden dynamics of speech captured by a bone-conductive microphone," Proc. 6th Workshop on Multimedia Signal Processing, pp.19–22, 2004.
[6] T. Dekens and W. Verhelst, "Body conducted speech enhancement by equalization and signal fusion," IEEE/ACM Trans. Audio, Speech, Language Process., vol.21, no.12, pp.2481–2492, 2013.
[7] E. Erzin, "Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings," IEEE/ACM Trans. Audio, Speech, Language Process., vol.17, no.7, pp.1316–1324, 2009.
[8] H.W Park, A.R. Khil, and M.J. Bae, "The improvement of mobile phone voice quality by bone-conduction device," Proc. IEEE International Conference on Consumer Electronics (ICCE), pp.397–398, 2013.
[9] R. Xiao, Y. Xiao, H. Wei, and K. Hasegawa, "Speech enhancement using bone-and air-conducted signals and adaptive GFLANN filter," Proc. 8th International Conference on Wireless Communications Signal Processing (WCSP), pp.1–5, 2016.
[10] A. Shahina and B. Yegnanarayana, "Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach," EURASIP J. Adv. Signal Process., vol.1, no.2, pp.1–10, 2007.
[11] P.N. Trung, M. Unoki, and M. Akagi, "A study on restoration of bone-conducted speech in noisy environments with LP-based model and gaussian mixture model," J. Signal Process., vol.16, no.5, pp.409–417.
[12] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," IEEE/ACM Trans. Audio, Speech, Language Process., vol.20, no.9, pp.2505–2517, 2012.
[13] K. Vijayan and R. Murty, "Comparative study of spectral mapping techniques for enhancement of throat microphone speech," Proc. 20th IEEE National Conference on Communications (NCC), pp.1–5, 2014.
[14] B. Huang, Y. Gong, J. Sun, and Y. Shen, "A wearable bone-conducted speech enhancement system for strong background noises," Proc. 18th IEEE International Conference on Electronic

Packaging Technology (ICEPT), pp.1682–1684, 2017.
[15] D. Watanabe, Y. Sugiura, T. Shimamura, and H. Makinae, "Speech enhancement for bone-conducted speech based on low-order cepstrum restoration," Proc. IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp.212–216, 2017.
[16] H.P. Liu, Y. Tsao, and C.S. Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," Speech Commun., vol.104, pp.106–112, 2018.
[17] 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," Proc. IEEE International Conference on Computer and Communications (ICCC), 2018.
[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol.9, no.8, pp.1735–1780, 1997.
[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint, arXiv:1409.0473, 2014.
[20] X. Wang, "The harmonic organization of auditory cortex," Front. Syst. Neurosci., vol.7, no.114, pp.1–11, 2013.
[21] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp.600–612, 2004.
[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint, arXiv:1502.03167, 2015.
[23] L. Sun, S. Kang, and K. Li, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4869–4873, 2015.
[24] C. Raffel and D.P.W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," arXiv preprint, arXiv:1512.08756, 2015.
[25] X. Mankun, P. Xijian, L. Tianyun, and X. Mantian, "A new time-frequency spectrogram analysis of FH signals by image enhancement and mathematical morphology," Proc. Fourth IEEE International Conference on Image and Graphics (ICIG 2007), pp.610–615, 2007.
[26] J. Dennis, H.D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," IEEE Signal Process. Lett., vol.18, no.2, pp.130–133, 2011.
[27] 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," IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP), pp.749–752, 2001.
[28] A. Gray and J. Markel, "Distance measures for speech processing," IEEE Trans. Acoust., Speech, Signal Process., vol.24, no.5, pp.380–391, 1976.
[29] P.C. Loizou, Speech Enhancement: Theory and Practice, CRC Press, Boca Raton, FL, USA, 2007.
[30] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," The 27th International Conference on Machine Learning (ICML), pp.807–814, 2010.
[31] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint, arXiv:1412.6980, 2014.
[32] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint, arXiv:1609.03499, 2016.
[33] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," IEEE/ACM Trans. Audio, Speech, Language Process., vol.22, no.12, pp.1713–1725, 2014.
[34] H.T. Luong, S. Takaki, G.E. Henter, and J. Yamagishi, H.T. Luong, S. Takaki, Henter G E, et al, "Adapting and controlling DNN-based speech synthesis using input codes," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4905–4909, 2017.