A Novel Throat Microphone Speech Enhancement Framework based on Deep BLSTM Recurrent Neural Networks

Changyan Zheng[1], Xiongwei Zhang[1], Meng Sun[1] , Jibin Yang[1], Yibo Xing[1]
1. Lab of Intelligent Information Processing, Army Engineering University, Nanjing, China
echoaimaomao@163.com

Abstract—**Body-conducted microphone (BCM) speech is immune to noise but has the shortcomings such as severe loss of high-frequency components. Direct enhancement about BCM speech is meaningful but few works have been done so far. Firstly, considering the lack of open datasets, a specific dataset of throat microphone (TM) speech is constructed by us and now is opened online. Secondly, we propose a novel speaker-dependent TM speech enhancement framework based on deep bidirectional recurrent neural networks using Long Short-Term Memory units (BLSTM-RNN). In this framework, magnitude spectrums are directly transformed to achieve speech enhancement, which is different from previous works based on the source-filter model. BLSTM-RNN is deployed to further improve the results of transformation. Objective and subjective results show that the quality of TM speech is substantially improved, where PESQ, STOI and MOS scores are improved 0.71, 0.21 and 1.36 respectively. Another important criterion, LSD decreases 0.63. Another important criterion, LSD decreases 0.63.**

Keywords- Body-conducted microphone; throat microphone; database; bidirectional recurrent neural networks; Long Short Term Memory

I. INTRODUCTION

Nowadays, researchers have investigated many different speech enhancement methods to improve the quality and intelligibility of speech using single or multiple air-conducted microphones (ACM). However, the improvement by using these microphones may be limited in case of non-stationary noise and strong background noise [1].

Body-conducted microphone (BCM) utilizes the vibration of the human body to conduct an electrical signal and is immune to background noise [2]. Though the essential characteristic of BCM is noise robustness, it faces severe loss of high-frequency components due to the attenuation of human body channel. Meanwhile, the energy of middle-frequency components is much larger than the normal ones, and this may be the reason why the speech sounds unbearable muffled [3]. At the same time, some phonemes, generated in the oral cavity not the vocal cords, like unvoiced fricatives, plosives and affricates are totally lost [4]. Moreover, some researchers [5] have proven that the frequency characteristic and sound quality varied depending on the microphone location, which further increase the difficulty of BCM speech enhancement.

Since all the above shortcomings, in most cases, BCM plays an auxiliary role for improving ACM speech enhancement performance in noise environments. For instance, [6] utilizes a BCM to acquire glottal source information. In [7], BCM speech is used to detect voice activity in low SNR conditions. In other cases, ACM is needed to help enhance BCM speech [8] [9]. Nevertheless, it is hard but meaningful to enhance the BCM speech directly, because ACM speech can be completely unintelligible in some occasions, especially in strong noise environments like driver cabins.

To date, some approaches have been proposed to enhance BCM speech directly. Almost all of these approaches are based on the source-filter model, which models speech as a combination of an excitation and a spectral envelope filter. With the assumption that the excitation is unchanged between ACM and BCM speech, these approaches usually transform the Linear Predictive (LP) family parameters, like Line Spectral Frequency (LSF), Linear Prediction Cepstrum Coefficient (LPCC) [10] [11], etc. Neural networks and Gaussian Mixture Models (GMMs) are often chosen as the transformation tools. However, the LP-based model has assumed the independence of source signal and filter, which may be problematic in some occasions. To overcome this problem, the latest related method [12] has trained distinctive GMMs for different types of phones. Meanwhile, the transformation of excitation features has also been taken into consideration. Nevertheless, how to recognize phones correctly and effectively remains challenging. Other methods [4] to alleviate this problem use STRAIGHT [13], which firstly extracts pitch, then extracts smoothed pitch adaptive spectrums. GMMs are deployed as the STRAIGHT spectrum transformation model.

In this paper, we focus on the throat microphone (TM) speech enhancement. Mohammadi [14] has pointed out that signal-based analysis/synthesis approaches usually have higher quality since they need not make any restrictive assumptions such as the independence of source signal and filter. Therefore, we transform the Short-Time Fourier Transformation (STFT) magnitude directly and synthesize enhanced speech with phase information unchanged. However, the high-dimensional magnitude is difficulty to learn.

Thanks to DNN's strong ability of exploring in a much larger hypothesis space, it is possible for us to learn complex distribution of high-dimension spectra now [15]. In addition, based on the analysis of BCM speech spectra above, we think it is necessary to use the context to help infer the lost information, because unvoiced fricatives and plosives of ACM speech correspond to nothing in TM speech. We cannot expect the transformation model to infer the lost components, if we do not give it necessary information. The only way to recover this kind of components is to use the contextual information. Bidirectional recurrent neural networks with Long Short-Term Memory (BLSTM-RNN) [16] [17] can learn the contextual information both from the past and future, and it alleviates the exploding gradients problem in conventional RNN. Thus it is able to learn contextual information effective and is suitable for our task.

The rest of the paper is organized as follows. The TM speech enhancement framework is firstly described, then the parallel dataset of TM and ACM speech is introduced. Experimental settings and evaluations are presented in Section IV. Conclusions are provided in Section V

II. TM Speech Enhancement Framework

A. *The proposed Framework*

The overall framework is illustrated in Fig.1. In the training stage, spectral magnitudes of ACM and TM speech are computed by STFT firstly. The raw magnitude usually has very large dynamic range, thus we use the log compression to normalize. Dynamic Time Warping (DTW) is then performed to obtain aligned spectral features. To facilitate the training of neural networks, spectral features are further normalized to a standard normal distribution, and the mean and variance are recorded subsequently. A feature window with slide 1 frame is used to combine *N* TM frames, which means we use *N* neighborhood TM speech frames to estimate one central frame of ACM speech. Finally, the spectral features of TM and ACM are sent to BLSTM-RNN model for training.

In the enhancement stage, the magnitude and phase of TM speech are firstly computed, then the log spectral magnitude is normalized according to the recorded mean and variance of TM speech. Next, trained BLSTM-RNN model is used to transform the combined feature vectors and then the output is de-normalized by the mean and variance of ACM speech in the training stage. Finally, the de-normalized output and TM phase synthesize the enhanced TM speech by inverse STFT.

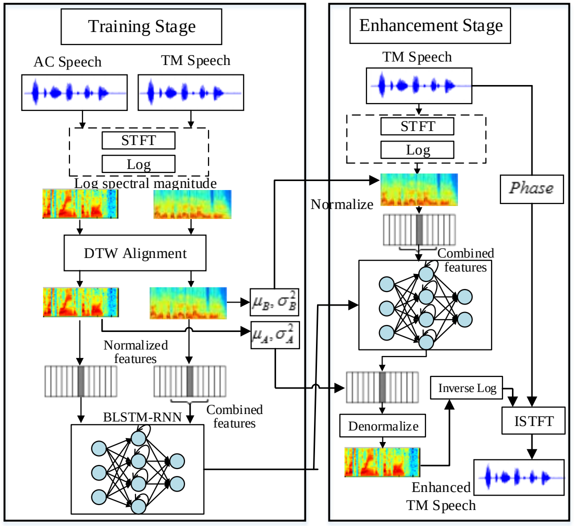

Fig. 1: The Overall TM Speech Enhancement Framework

B. *Bidirectional RNN with LSTM units*

A typical RNN is proven to be difficulty to learn long-time dependencies due to the gradients exploding and vanishing problem during training. LSTM units alleviate the gradients problem by introducing purpose-built gates to facilitate information flow and store in networks. Our networks is a combination of bidirectional RNN and LSTM memory block, which can learn long-range contextual in both forward and backward directions.

Bidirectional RNN computes the forward hidden sequence $\vec{h}$, the backward hidden sequence $\overleftarrow{h}$ and the output sequence y as followings.

$$\vec{h}_t = \sigma(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} \vec{h}_{t-1} + b_{\vec{h}}) \tag{1}$$

$$\overleftarrow{h}_t = \sigma(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \tag{2}$$

$$y_t = (W_{\vec{h}y} \vec{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y) \tag{3}$$

Where W 's are the weight matrices (e.g. W_{xh} is the input hidden weight matrix), the b's are the bias vectors (e.g. b_h is hidden bias vector) and δ denotes the hidden layer activation function.

A LSTM unit is composed of an input gate, an output gate, a forget gate, and a cell state vector. It computes the hidden layers according to the following equations:

$$i_t = \delta(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \tag{4}$$

$$f_t = \delta(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f) \tag{5}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \tag{6}$$

$$o_t = \delta(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \tag{7}$$

where i_t , f_t , c_t are input gate, forget gate, cell state and output gate respectively. By stacking multiple hidden layers, a deep LSTM-RNN architecture is built. The architecture is illustrated in Fig.2.

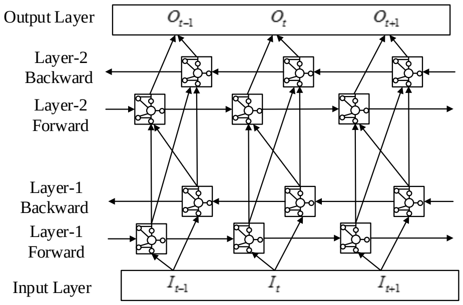

Fig. 2: The deep BLSTM-RNN model

III. Speech Dataset

To the best of our knowledge, there is no public BCM speech dataset online, hence we make our own TM speech dataset. The corpus is scratched from Chinese newspapers, magazines and articles, etc. Most Chinese Mandarin pronunciations are included. Phonetically-balanced sentences are also constructed. At last, 1000 sentences, each of which is around 3-5*s*, are selected for our corpus. In the current version of the dataset, eight male speakers are involved in our recordings, and each speaker reads 200 different sentences, which are divided into 160 sentences and 40 sentences for trainset and testset, respectively. No overlap between the two sets.

TM and ACM speech are recorded at 32-kHz sampling rate. Whereas they are recorded synchronously, there is still a 1-20*ms* deviation in each sentence due to the mismatch of recording devices. To alleviate potential alignment error in later experiments, we mark start and end points of every TM and ACM wave files in a manual way, thus, the corresponding speech is with the same content but different lengths. DTW is performed to solve the time alignment problem and as is seen in Fig.1.

We have to point out that the quality of TM speech recordings are not equally good, some are clear but many others are very unclear. We conjecture the quality variation attributes to different speakers' ways of pronunciation and biological features, which may affect the vibration of sensors in TM.

Our dataset is now open on the web site: https://github.com/echoaimaomao/TM-Speech-Dataset. In the future work, we will promote to make more high quality data of BCM speech.

IV. EXPERIMENTS

A. *Experimental Setup*

In our experiments, we train an enhance model for each speaker. The duration of training speech is about 11 minutes, which contains 52174 speech frames (32ms with the frame shift 10ms) in average, while the testing data is about 3 minutes, and 10% of the training data is used for model validation. Both the training and testing data are down-sampled to 8kHz. 127-dimensional spectral magnitudes are extracted, where a feature window of 23 frames (11 to the left, 11 to the right) are used. Time step is set to 23 frames either, which means we force the hidden layers of BLSTM-RNN to memory the contextual information of 23 frames. The number of units in each layer of BLSTM-RNN is [129 512 512 129] respectively, where the hidden layer activation function is *tanh* and the output layer is linear. The dropout ratio 0.2 is set to all hidden layers. We train the networks with mean square error (MSE) cost function with min-batch size of 128. The root mean square propagation (RMSProp) is chosen as the optimizer. The initial global learning rate is set to 0.01 which is reduced by half once the validation loss is not reduced. The best model is chosen by the least validation loss.

Three metrics including Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI) and Log-spectral Distance (LSD) are used to evaluate speech quality objectively. PESQ score measures the overall speech quality, STOI score measures the speech intelligibility, while LSD measures the log-spectral distance between two signals.

Mean Opinion Score (MOS) test is also conducted to gain the subjective evaluation. In the MOS test, listeners are asked to score the speech using a 5-point scale lied in [0 5], which is corresponding to PESQ score [-0.5 4.5]. Twelve native Chinese volunteers (six males and six females) are involved in the test.

B. *Results and Analysis*

1) Comparisons with RNN: Table I is the objective evaluation results about RNN and BLSTM-RNN models, where RNN is set with the same architecture and training scheme as BLSTM-RNN. Fig.3, Fig.4 are the spectrograms of two different utterances. The ORIGIN is the evaluation between ACM and TM speech. We can see that majorities of PESQ scores are under 2 and STOI are under 0.55, which indicates the rather low quality of TM speech. From Fig.3, Fig.4 (b), severe high-frequency components (2-4kHz) loss can be observed, and the consonant ***xi*** (sounds like /si:/ in English) almost disappears. Also, energy of the middle-frequency (1-2kHz) circled in solid line is higher than the corresponding components of ACM speech. These can explain why the TM speech sounds muffled and unclear. A great restoration of high-frequency components can be seen in Fig.3, Fig.4 (c)、(d) and ***xi*** has also been restored, which indicates the effectiveness of both the RNN and BLSTM-RNN models. The average PESQ and STOI scores have been improved to 2.624 and 0.756 by BLSTM-RNN model, which means the TM speech can be understood after enhancing. BLSTM-RNN model scores much better than RNN model among all the three metrics. From the figures marked by rectangular boxes we can see that, RNN model seems incapable of inferring the missing parts as well as the BLSTM-RNN model does. Instead, RNN fills the blank with much noise.

By analyzing the PESQ results improved by BLSTM-RNN model, it can be noticed that Fig.4 which has lower original speech quality than Fig.3 has achieved relatively greater improvement, and both the original utterances have been improved to around 2.6. This implies the BLSTM-RNN model has reached to a saturated state. The small random noise which can be seen in (c) of the two figures may result in this problem. We guess that the mean square error loss (MSE) of the neural networks may responsible for this, because researchers [18] have proven the Euclidean distance is likely to introduce average noise.

2) Comparisons with STRAIGHT Model: In this section, we explore the performance of the two different speech analysis/synthesis models, the magnitude-phase model (abbreviate to MP) and STRAIGHT model. STRAIGHT decomposes speech into pitch, aperiodicity and STRAIGHT spectrums, and we just transform spectrums as well. STRAIGHT spectrums are firstly compressed to log spectrums, because we have found this operation can obtain better results. After then, log STRAIGHT spectrums are processed according to Fig.1. The same BLSTM-RNN is deployed.

Table II presents the objective results. Table III shows the subjective results. Comparing Table II with Table I, we

TABLE I: Objective Evaluation Results of RNN and BLSTM-RNN

Person	PESQ			STOI			LSD		
	ORIGIN	RNN	BLSTM	ORIGIN	RNN	BLSTM	ORIGIN	RNN	BLSTM
male1	1.775	2.416	2.882	0.561	0.726	0.765	1.752	1.081	1.046
male2	1.809	2.192	2.463	0.506	0.674	0.703	1.691	1.254	1.212
male3	2.285	2.567	2.801	0.547	0.736	0.752	1.786	1.131	1.116
male4	2.119	2.591	2.931	0.609	0.769	0.795	1.599	1.147	1.119
male5	1.988	2.207	2.428	0.549	0.688	0.788	2.215	1.121	1.096
male6	1.675	2.097	2.492	0.532	0.714	0.757	1.741	1.198	1.131
male7	1.839	2.197	2.428	0.498	0.695	0.722	1.633	1.154	1.133
smale8	1.798	2.214	2.564	0.537	0.725	0.764	1.542	1.142	1.085
Average	1.911	2.310	2.624	0.542	0.716	0.756	1.745	1.154	1.117

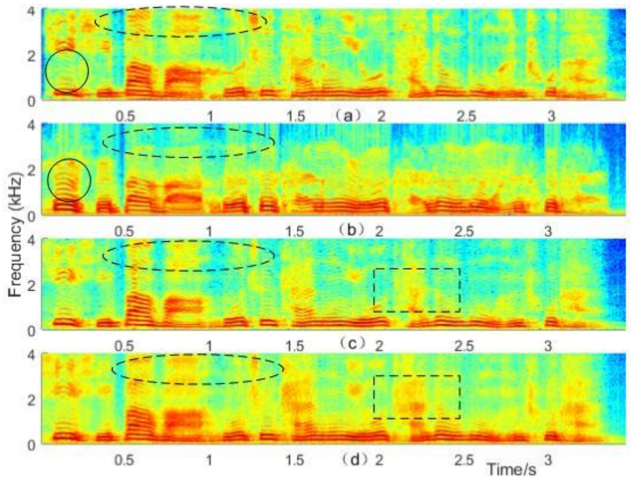

Fig. 3: Spectrograms of an utterance (a) AC speech (PESQ=4.5) (b) TM speech (PESQ=2.167) (c) BLSTM Enhanced Speech (PESQ=2.671) (d) RNN Enhanced Speech (PESQ=2.408)

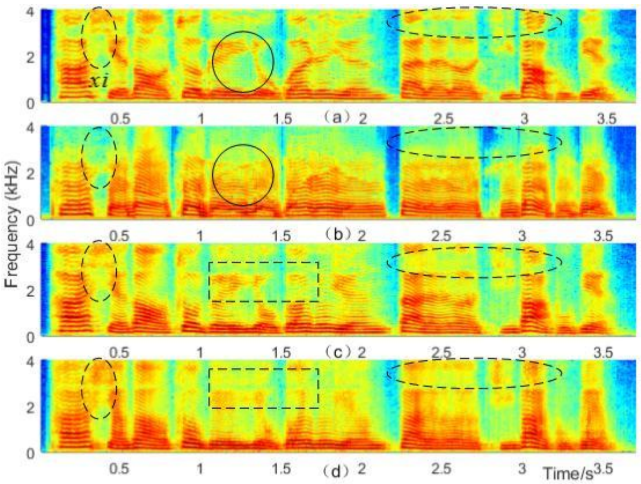

Fig. 4: Spectrograms of an utterance (a) AC speech (PESQ=4.5) (b) BC speech (PESQ=1.577) (c) BLSTM Enhanced Speech (PESQ=2.578) (d) RNN Enhanced Speech (PESQ=1.967)

TABLE II : Objective Results of STRAIGHT

	m001	m002	m003	m004	m005	m006	m007	m008	Average
PESQ	2.473	2.135	2.292	2.492	2.145	2.213	2.219	2.277	2.281
STOI	0.709	0.657	0.716	0.746	0.653	0.691	0.639	0.705	0.690
LSD	1.161	1.302	1.172	1.228	1.184	1.222	1.201	1.171	1.205

TABLE III: Subjective Results of MP and STRAIGHT

	m001	m002	m003	m004	m005	m006	m007	m008	Average
TM	2.017	1.263	1.402	1.538	1.113	1.175	1.050	1.788	1.418
MP	3.467	2.450	2.713	2.963	2.075	3.275	2.263	2.975	2.773
STRAIGHT	3.334	2.378	2.756	2.948	2.063	3.125	2.346	2.912	2.732

can notice that although using the same transformation model, STRAIGHT has poorer performance than MP in all the three metrics. In particular, its PESQ score is lower about 0.4 than MP, which means a larger spectrum reconstruction distortion. In Table III, it can be noticed that the MOS score of MP model is higher about 0.2 than STRAIGHT. From all the objective and subjective results above, we can conclude that in the case of just performing spectrum transformation, the MP model can obtain better results than STRAIGHT in TM speech enhancement. By the way, we find out that despite having rather low original MOS score like m006, m005 and m007 are not improved as greatly as the former. This may relate with the difference of speakers, which we have mentioned in Section III.

V. CONCLUSION

In this paper, we propose a novel speaker-dependent TM speech enhancement framework based on BLSTM-RNN. Speech enhancement is achieved by directly transforming the magnitude spectrums. Meanwhile, BLSTM-RNN is deployed as the transformation model. Objective and subjective evaluation results show that the TM speech quality is substantially improved. Future works include the combination of more advanced models like generative adversarial networks and the adaptation techniques for speaker-independent enhancement.

REFERENCES

[1] Loizou, Speech Enhancement: Theory and Practice, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[2] H. S. Shin, H. G. Kang, and T. Fingscheidt. Survey of speech enhancement supported by a bone conduction microphone. Proceedings of Speech Communication Symposium, 2012, 1–4.

[3] Jing Li. Speech reconstruction technology based on bone conduction signal, Master's thesis, Northwestern Poly-technical University, China, 2004.

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

[5] M. McBride, P. Tran, T. Letowski, and R. Patrick. The effect of bone conduction microphone locations on speech intelligibility and sound quality, Applied ergonomics, vol. 42, no. 3, pp. 495–502, 2011.

[6] M. S. Rahman and T. Shimamura. Pitch characteristics of bone conducted speech. 18th IEEE European Signal Processing Conference, 2010, pp. 795–799.

[7] M. Zhu, H. Ji, F. Luo, and W. Chen. A robust speech enhancement scheme on the basis of bone-conductive microphones, Signal Design and Its Applications in Communications, 2007, pp. 353–355.

[8] T. Shimamura and T. Tamiya. A reconstruction filter for bone-conducted speech, 48th Midwest Symposium on Circuits and Systems, 2005, pp. 1847–18

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

[10] 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, no. 2, pp. 10–10, 2007.

[11] T.Vu, M. Unoki, and M. Akagi. An lp-based blind model for restoring bone-conducted speech. IEEE Second International Conference on Communications and Electronics, 2008, pp. 212–217.

[12] M. T. Turan and E. Erzin. Source and filter estimation for throat microphone speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 2, pp. 265–275, 2016.

[13] H. Kawahara, J. Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, 2001.

[14] S. H. Mohammadi and A. Kain. An overview of voice conversion systems. Speech Communication, 2017.

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

[16] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, vol. 18, no. 5, pp. 602–610, 2005.

[17] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory . Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] Isola, Phillip and Zhu, Jun-Yan. Image-to-image translation with conditional adversarial networks. arXiv preprint.

AUTHORS' BACKGROUND

Your Name	Title*	Research Field	Personal website
Changyan Zheng	Phd candidate	Speech enhancement and deep learning	
Xiongwei Zhang	full professor	multimedia information processing	
Meng Sun	associate professor	Signal processing	
Jibin Yang	Associate Professor	Speech enhancement, Signal processing	
Yibo Xing	Master candidate	Speech enhancement	