Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Novel Encoder-Decoder Model Via NS-LSTM Used for Bone-conducted Speech Enhancement

DONGJING SHAN[1], XIONGWEI ZHANG[1], AND CHAO ZHANG.[2]

[1]Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing, 210007, P. R. China. (e-mail: shandongjing@pku.edu.cn)

[2]Key Laboratory of Machine Perception (MOE), Peking University, Beijing, 100871, P. R. China (e-mail: chzhang@cis.pku.edu.cn)

Corresponding author: Xiongwei Zhang (e-mail: xwzhang9898@163.com).

This work was supported by the National Natural Science Foundation of China under Grants 61471394.

ABSTRACT Bone-conducted speech can be used to communicate in a very high noise environment. In this paper, a method of improving the quality of bone-conducted speech is presented. The speech signal of a speaker is passed through a novel dictionary representation based encoder-decoder model. In the encoder, our designed Non-negative and Sparse Long Short-Term Memory (NS-LSTM) recurrent neural networks is deployed to generate combination coefficients on the dictionary established by sparse non-negative matrix factorization. Then the decoder is designed and utilized to enhance the dictionary representation based on local attention mechanism. Two optimizers are adopted when training the model as a whole and the encoder is pre-trained individually to make the convergence faster. In experiments, we compare the proposed method with a direct transformation via LSTM networks, and numerous criterions are used for evaluation. Objective and subjective results demonstrate that our method behaves better and achieves satisfactory performance even when coping with some challenging cases.

INDEX TERMS Speech enhancement, recurrent neural networks, long short-term memory, logic gates.

I. INTRODUCTION

BODY-conducted (BC) microphone utilizes the vibration of human body like throat [1], skull [2], the skin backed the ear [3] to conduct electrical signal, and then the speech is immensely robust even in severely degraded environments [4]. The BC microphone is widely used in the communication system of military equipments like tanks or helicopters, and also has well applications for civil activity, such as forestry, oil exploration and production, mine, special agent, emergency rescue and so on. It can satisfy the needs in special situations, nevertheless, its intelligibility is lower than air-conducted (AC) one as it faces severe degradation of high-frequency components due to the attenuation of human body channel [5], [6], or loses some phonemes like unvoiced fricatives, plosives and affricates [7], which are generated in human oral cavity.

In most cases, BC microphone plays an auxiliary role in improving the quality of AC speech in noise environments [12], [13]; on the other hand, AC microphone is needed to help enhance BC speech [14], [15]. But in a few cases, it is meaningful to enhance the BC speech independently, because the corresponding AC speech can be completely unintelligible in some extreme situations, such as in forge shops with strong noise. To our knowledge, the approaches for enhancement can be summarized into three categories: bandwidth extension, equalization method and source-filter model. In the first one, the high and low frequency components in the speech are regarded to have the same harmonic structures, so the low-frequency spectrum can be expended directly to recover the high-frequency structure. Specifically in [15], the speech generated from bone and tissue conduction captured using an in-ear microphone is enhanced using adaptive filtering and a non-linear bandwidth extension method. The equalization method aims to calculate the inverse transformation function of the transmission channel. It was firstly proposed by Shimamura [14] and a linear-phase impulse response filter was calculated by taking an inverse discrete Fourier transform of the ratio of long-term AC and BC speech spectra. Kondo et al. proposed the short-term DFT magnitude ratio-based method [10], which estimated the equalization filter with a frame-by-frame basis approach, and then obtain a mean estimate by averaging. The source-filter model de-

composes speech as a combination of excitation and spectral envelope filter [16]. Under the assumption of the identical excitation between BC and AC speech, these approaches usually transform the Linear Predictive (LP) family parameters like Line Spectral Frequency (LSF), Linear Prediction Cepstrum Coefficient (LPCC) [17], [18] by neural networks or Gaussian mixture models. However, the LP-based models assume the independence of source signal and filter, which may be problematic in some occasions. To overcome this problem, the method [20] has trained distinctive GMMs for different types of phones. Nevertheless, how to recognize phones correctly and effectively remains challenging.

In this paper, we propose a dictionary representation based encoder-decoder model for speech enhancement, specifically it transforms the short-time spectral magnitude of BC speech via an encoder-decoder and then synthesizes enhanced speech with the phase information unchanged. The AC speech sparse dictionary is established by sparse non-negative matrix factorization (sparse NMF) [23]–[25] firstly, and then the encoder transforms the spectral magnitudes into dictionary representation coefficients by using Non-negative and Sparse Long Short-Term Memory (NS-LSTM) recurrent neural networks, finally, the decoder with local attention mechanism is aimed to improve the quality and accuracy of the encoder outputs. In training stage, two optimizers are allocated to the encoder and decoder respectively and they are optimized as a whole, and also a pre-training with the encoder is adopted to provide initial parameters and accelerate the networks' convergence speed.

The rest of the paper is organized as follows. The encoder-decoder enhancement framework is presented in the next section, and then the NS-LSTM based encoder is described in Section III. After that, the decoder with local attention mechanism is illustrated in Section IV. Lastly, the parallel dataset of BC|AC speech and the experiment results are presented in Section V.

II. THE ENCODER-DECODER ENHANCEMENT FRAMEWORK

Our designed framework is illustrated in Fig.1 In the training stage, spectral magnitudes of AC and BC speech are computed by STFT firstly, and then, a log compression [21] is performed as the raw magnitude usually has very large dynamic range. 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. Next, Auto-Regressive and Moving Average Model (ARMR) [19] filter process is performed to make supplement of missing values and stabilize the signal. Meanwhile, an AC speech dictionary symbolized as D is computed by using sparse NMF. After that, the spectral features of BC are sent to the encoder networks for training and the outputs are the representation coefficients on the dictionary. In the pre-training stage, the loss function of the encoder is the difference between linear combination of the dictionary elements and the true AC speech. Finally, the decoder with

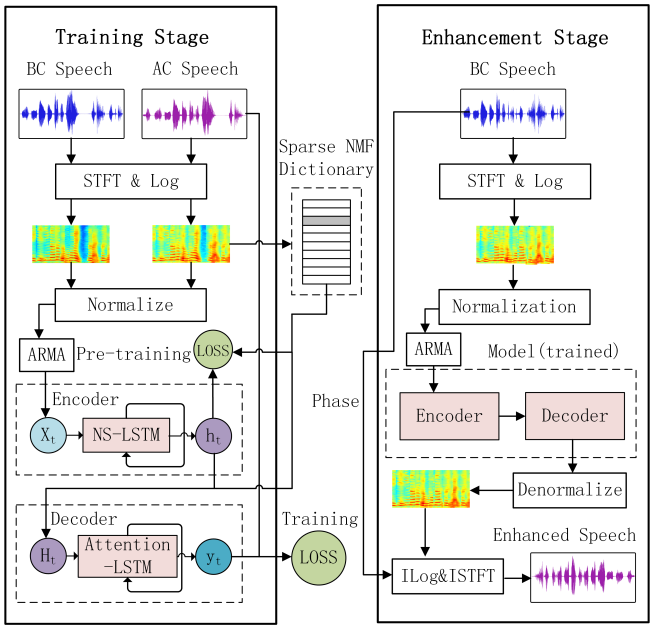

FIGURE 1. BC speech enhancement framework based on encoder-decoder model.

local attention mechanism is utilized to promote the accuracy of the encoder outputs, the encoder and decoder are training as a whole with optimizer for each of them. In this stage, the loss function is located in the end of the decoder with the error computed and back-propagated frame by frame.

In the enhancement stage, the magnitude and phase of BC speech are firstly computed, then the log spectral magnitudes are normalized according to the recorded mean and variance of BC speech. Next, trained encoder-decoder model is used to enhance the feature vectors through the intermediate state of dictionary coefficients. In the end, the generated spectral magnitudes are denormalized and used to synthesize the enhanced speech via inverse STFT together with the BC phase information.

III. NS-LSTM BASED ENCODER

The encoder networks consists of three layers: linear layer, original LSTM layer and our designed NS-LSTM layer, which are arranged in a bottom-up manner. The encoder is aimed to output non-negative and sparse coefficients on the dictionary elements generated by sparse NMF. Inspired by the work in [26], where a simple recurrent neural network was proposed to derive sparse coding, we exploit NS-LSTM layer to implement the above constrains and exhibit the unit structure in Fig.2. The layer's forward propagation process is formulated as follows:

$$f_t = \sigma(W_{fx}X_t + W_{fh}h_{t-1} + b_f) \quad (1)$$

$$g_t = \phi(W_{gx}X_t + W_{gh}h_{t-1} + b_g) \quad (2)$$

$$o_t = \mathrm{sh}_{(M,u)}(W_{ox}X_t + W_{oh}h_{t-1} + b_o) \quad (3)$$

$$S_t = S_{t-1} \odot f_t + g_t \odot (1 - f_t) \quad (4)$$

$$h_t = \sigma(S_t) \odot o_t \tag{5}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\phi(x) = \tanh(x)$; $\mathrm{sh}_{(M,u)}(x) = M(\tanh(x+u) + \tanh(x-u))$ is the so-called "double tanh" function [26], in which M is a trainable diagonal matrix and u a trainable vector. Apparently, the sigmoid function σ in Eqn.5 acts for vector compression and non-negative constraint, the shrinkage function $\mathrm{sh}_{(M,u)}$ is used to keep sparsity of the output gate, and as a result, the output of the unit satisfies the requirements of being Non-negative and Sparse.

The encoder is pre-trained individually to provide better initial parameters rather than random settings for the whole model training. The loss function in pre-training stage is presented in Eqn.6, and the parameters are updated to minimize the loss through backward propagation.

$$\begin{aligned} &\min_{\boldsymbol{c}} \mathrm{L}(\boldsymbol{c}), \quad s.t.\ \boldsymbol{c} \geq 0 \\ &\mathrm{L}(\boldsymbol{c}) = ||D\ \boldsymbol{c} - X_{ac}||_F^2 + \lambda||\boldsymbol{c}||_1^2 \end{aligned} \tag{6}$$

where $\boldsymbol{c} = [c_1, c_2, ..., c_\tau]$ is an output matrix comprising τ coefficient vectors, X_{ac} is the spectral features of one AC speech sentence, in which each column represents the feature of one speech frame. The coefficient c_t is the output of each NS-LSTM unit, it satisfies non-negative and sparse constraint, and is used to combine the sparse NMF dictionary D to approach the real speech frame. The regularization term ensures the sparsity and usually is relaxed to Frobenius norm to make it differentiable, $\lambda||\boldsymbol{c}||_1^2 \rightarrow \lambda||\boldsymbol{c}||_F^2$. Additionally, when training the model in experiments, speech sentences are fed in batch style to make the calculated gradient more stable, then the total loss will be the simple sum with regard to the sentences in a batch.

In the backpropagation process, the gradients of two hidden vectors are computed as a prerequisite:

$$\delta_h^{(t)} = \frac{\partial \mathrm{L}}{\partial h_t} = \frac{\partial \mathrm{L}}{\partial c_t}\frac{\partial c_t}{\partial h_t} = 2D^T(Dc_t - X_{ac}^{(t)}) + 2\lambda c_t \tag{7}$$

$$\begin{aligned} \delta_S^{(t)} &= \frac{\partial \mathrm{L}}{\partial S_t} = \frac{\partial \mathrm{L}}{\partial S_{t+1}}\frac{\partial S_{t+1}}{\partial S_t} + \frac{\partial \mathrm{L}}{\partial h_t}\frac{\partial h_t}{\partial S_t} \\ &= \delta_S^{(t+1)} \odot f_{t+1} + \delta_h^{(t)} \odot \sigma(S_t)(1-\sigma(S_t)) \odot o_t \end{aligned} \tag{8}$$

The parameters' gradients can be calculated based on the above ones, we list one of them below and deduce the rest in the appendixes.

$$\begin{aligned} \frac{\partial \mathrm{L}}{\partial W_{fh}} &= \sum_{t=1}^{\tau} \frac{\partial \mathrm{L}}{\partial S_t}\frac{\partial S_t}{\partial f_t}\frac{\partial f_t}{\partial W_{fh}} \\ &= \sum_{t=1}^{\tau} \delta_S^{(t)} \odot S_{t-1} \odot f_t(1-f_t)(h_{t-1})^T \end{aligned} \tag{9}$$

IV. LOCAL ATTENTION MECHANISM BASED DECODER

The encoder-decoder model is widely used in Sequence to Sequence machine translation [33], [34], and to the best of my knowledge, we are the first to introduce this model to speech enhancement. Based on the encoder illustrated in the section above, we redesign and utilize the decoder with local attention mechanism to improve the quality and accuracy of the dictionary representation, and achieve better performance when coping with the silent frames in BC speech (no salient in AC speech) or the ones accompanied by noise. Our designed decoder structure is depicted in Fig.3.

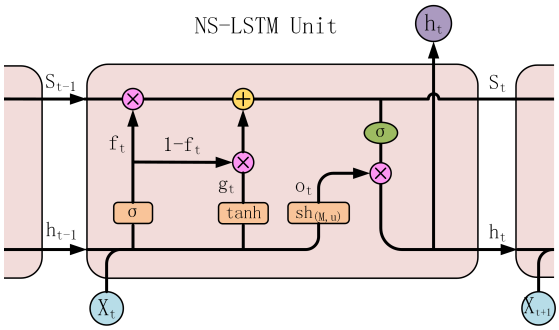

FIGURE 2. The inner structure of NS-LSTM unit in the encoder.

The output of the attention layer is calculated by weighted linear combination of the local encoder outputs:

$$a_i = \sum_{j \in N(i)} \omega_{ij} e_j \tag{10}$$

where $N(i)$ is neighbors of the j-th encoder output e_j, we adopt 10 neighbors with half on each side in the experiments, ω_{ij} is the combination weight.

$$\omega_{ij} = \frac{\exp(score_{ij})}{\sum_{j \in N(i)} \exp(score_{ij})}, \quad score_{ij} = d_{i-1}^T W_a e_j \tag{11}$$

where d_{i-1} is the decoder output in the last time frame, and W_a is the linear layer matrix of the attention mechanism. The local attention a_i is concatenated with the corresponding e_i as input of the decoder.

The mean square error (MSE) is calculated between d_i and the ground truth, and the decoder is optimized frame by frame. Back-propagation is used to update the parameters of the decoder, and meanwhile the optimizer of encoder is activated to update its own parameters.

V. EXPERIMENTS

A. TM SPEECH DATASET

One thousand of Chinese Mandarin sentences are selected as corpus, and each of them lasts for 3 to 5 seconds. Eight male speakers are required to read 200 sentences selected from the corpus randomly, and the speeches are recorded by air-conducted microphone and throat-conducted microphone simultaneously. For each person, 160 sentences are used for training and the rest for testing. All the speech are recorded at 32-kHz sampling rate, and the dataset is open on the web site: https://github.com/cvcoding/BC-Speech-Dataset.

B. EXPERIMENTAL SETUP

In our experiments, we train an enhance model for each speaker. The duration of training speech is about 11 minutes, while the testing data is about 3 minutes. Both of the

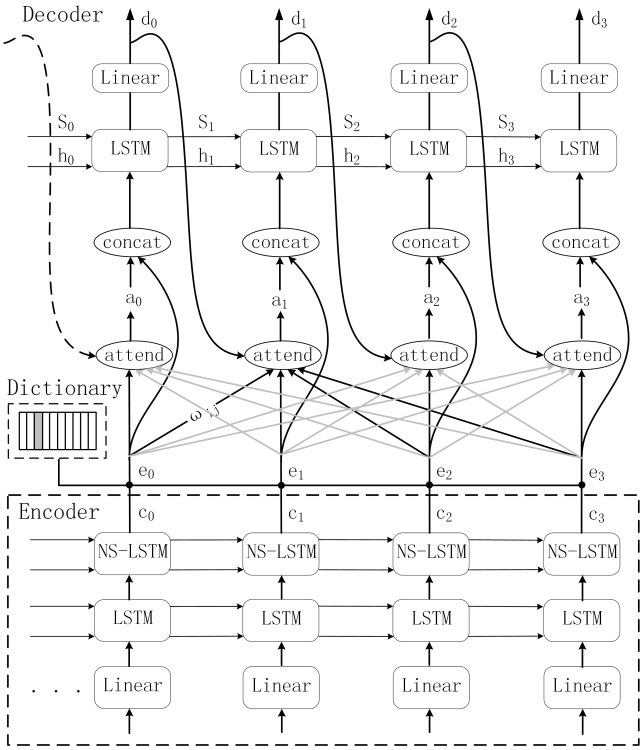

FIGURE 3. The structure of the decoder networks.

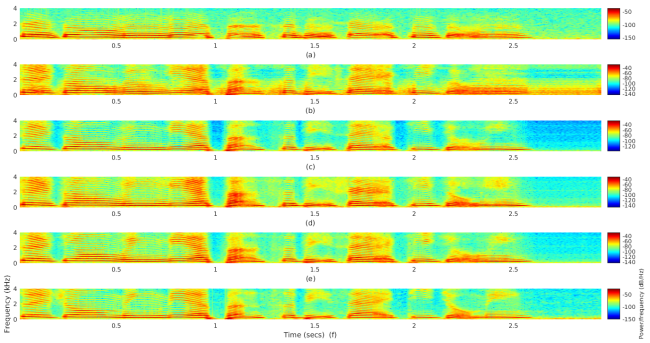

FIGURE 4. Spectrograms of one utterance. (a) BC speech, (b)speech enhanced by DNN, (c) speech enhanced by LSTM, (d) speech enhanced by our method, (e) speech enhanced by our method with multiple feature, (f) AC speech.

training and testing data are down sampled to 8 kHz, and 129-dimensional spectral magnitudes are extracted, where a feature window of 23 frames (11 frames to each side of the current frame) are used.

In the pre-training stage, the encoder networks are trained by using Adaptive Moment Estimation (Adam) optimizer, the dropout [27] ratio is set to 0.2 with regard to all hidden layers, the initial global learning rate is set to 0.01 which is reduced by half once the validation loss is not reduced. The training sentences are fed in batch style and the batch size is set to 8. The best model is chosen to initialize the encoder in the combined training stage according to the least validation loss.

Subsequently, in the next stage the decoder networks are

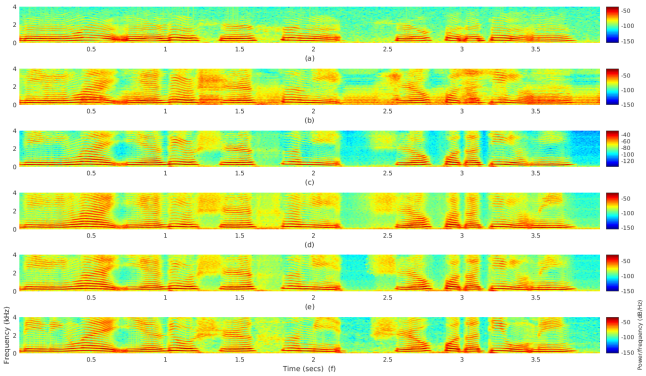

FIGURE 5. Spectrograms of the other sample. (a) BC speech, (b)speech enhanced by DNN, (c) speech enhanced by LSTM, (d) speech enhanced by our method with single feature, (e) speech enhanced by our method with multiple feature, (f) AC speech.

trained by another Adam optimizer, with the dropout ratio 0.2 and the learning rate 0.001. The encoder is updated from the initial state together with the decoder, and its learning rate is set to 0.0005 now.

In the model, only spectral magnitudes feature is used for training. Here we also adopt another two features to test the influence on performance. At first, the spectral magnitude (129-dimensional), MFCC (13-dimensional) [29] and LPC (13-dimensional) [28] are normalized by feature scaling respectively, then they are concatenated to formulate a 155-dimensional features, lastly normalization is performed and the result are used as the networks' input, the networks' output is also spectral magnitude feature, which is compared to the ground truth to calculate the loss. Additionally, we use the model to construct the relationship of spectrum magnitude part between BC and AC, and the phase part is assumed to be unchanged. In the experiment, we try to transform the phase of BC to the phase of AC by using the encoder-decoder model but without dictionary, and then synthesize enhanced speech according to the estimated phase.

Three metrics including Perceptual Evaluation of Speech Quality (PESQ) [30], Short-Time Objective Intelligibility [31] (STOI) and Log-spectral Distance [32] (LSD) are used to evaluate the 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.

C. RESULTS AND ANALYSIS

1) Results with spectral magnitude features

Table 1 is the objective evaluation results about DNN networks, LSTM networks and our model, where DNN and LSTM comprise two hidden layers and connect with a linear layer. The same training scheme as our encoder is used for DNN and LSTM. Fig.4, Fig.5 are two samples of speech spectrograms.

"BC" column is the evaluation of BC compared with AC speech. We can see that majorities of PESQ scores are

under 2 and STOI are under 0.60, which indicates the low intelligibility and quality of BC speech. From Fig.4, Fig.5 (a), severe high-frequency components (2-4kHz) loss can be observed, and the energy of the middle-frequency is higher than the corresponding components of AC speech.

The restoration of high-frequency components can be seen in Fig.4, Fig.5 (b),(c) and (d), which indicates the effectiveness of the three models. The average PESQ and STOI scores have been improved significantly, which means the enhanced BC speech can be understood. Among three models, our proposed one scores much better than others on the three metrics. From the figures we can see that, DNN and LSTM model seems incapable of inferring the missing parts while our model can fill the blank with the help of dictionary. The code is publicly available on the web site: https://github.com/cvcoding/BC-Speech-Code.

2) Results with combined features

The generated spectrograms of our model with multiple features (spectral magnitude, MFCC and LPC) are depicted in Fig.4(e) and Fig.5(e). Compared with the spectrograms of single feature, we can find that there is no apparent progress by using multiple features as input. The networks can extract useful features by large-scale training, input of multiple features will not exert obvious influence on the generated spectral magnitudes. The objective evaluation results are exhibited in Table.1. The average metrics indicate almost the same performance.

3) Results with estimated phase

In this section, we use the encoder-decoder model to enhance the speech log spectral magnitude part, and then, we use the model with Minor changes to enhance the phase part. The modified model discards the speech dictionary and replace NS-LSTM units by conventional LSTM units. The networks input is 129-dimensional BC phase feature and the output is enhanced phase. Finally, the enhanced spectral magnitude and phase are combined to synthesize the enhanced speech. The results are exhibited in Table.1.

VI. CONCLUSION

In this paper, we propose an encoder-decoder based speech enhancement framework, in which the encoder via non-negative and sparse LSTM networks is used to generate the representation coefficients, and the decoder with local attention mechanism is combined to further improve the speech quality. In the experiments, we adopted two methods for comparison, and the results demonstrate that our method behaves well when reconstructing the high-band signals. Nevertheless, our work is based on specific speaker, in the future work, we would like to propose a framework that realizes speaker-independent effect, and meanwhile, the loss function can be exploited and re-designed to improve the enhancement performance.

.

Appendices for Section III, the gradients are listed as follows:

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{fx}} &= \sum_{t=1}^{\tau} \frac{\partial \mathrm{L}}{\partial S_t}\frac{\partial S_t}{\partial f_t}\frac{\partial f_t}{\partial W_{fx}} \\ &= \sum_{t=1}^{\tau} \delta_S^{(t)} \odot S_{t-1} \odot f_t(1-f_t)(X_t)^T\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{gh}} &= \sum_{t=1}^{\tau} \frac{\partial \mathrm{L}}{\partial S_t}\frac{\partial S_t}{\partial f_t}\frac{\partial f_t}{\partial W_{gh}} \\ &= \sum_{t=1}^{\tau} \delta_S^{(t)} \odot (1-f_t) \odot g_t(1-g_t)(h_{t-1})^T\end{aligned} \tag{13}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{gx}} &= \sum_{t=1}^{\tau} \frac{\partial \mathrm{L}}{\partial S_t}\frac{\partial S_t}{\partial f_t}\frac{\partial f_t}{\partial W_{gx}} \\ &= \sum_{t=1}^{\tau} \delta_S^{(t)} \odot (1-f_t) \odot g_t(1-g_t)(X_t)^T\end{aligned} \tag{14}$$

$$\begin{aligned}&\frac{\partial \mathrm{L}}{\partial W_{oh}} = \sum_{t=1}^{\tau} \frac{\partial \mathrm{L}}{\partial h_t}\frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial W_{oh}} \\ &= \sum_{t=1}^{\tau} \delta_h^{(t)} \odot \sigma(S_t) \odot o'_t(W_{ox}X_t + W_{oh}h_{t-1} + b_o)(h_{t-1})^T \\ &\quad o'_t(x) = M(2 - \tanh^2(x+u) - \tanh^2(x-u))\end{aligned} \tag{15}$$

$$\begin{aligned}&\frac{\partial \mathrm{L}}{\partial W_{ox}} = \sum_{t=1}^{\tau} \frac{\partial \mathrm{L}}{\partial h_t}\frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial W_{ox}} \\ &= \sum_{t=1}^{\tau} \delta_h^{(t)} \odot \sigma(S_t) \odot o'_t(W_{ox}X_t + W_{oh}h_{t-1} + b_o)(X_t)^T\end{aligned} \tag{16}$$

$$\begin{aligned}&\frac{\partial \mathrm{L}}{\partial M} = \sum_{t=1}^{\tau} \frac{\partial \mathrm{L}}{\partial h_t}\frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial M} \\ &= \sum_{t=1}^{\tau} \delta_h^{(t)} \odot \sigma(S_t) \odot (\tanh(\Delta+u) + \tanh(\Delta-u)) \\ &\quad \Delta = W_{ox}X_t + W_{oh}h_{t-1} + b_o\end{aligned} \tag{17}$$

$$\begin{aligned}&\frac{\partial \mathrm{L}}{\partial u} = \sum_{t=1}^{\tau} \frac{\partial \mathrm{L}}{\partial h_t}\frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial u} \\ &= \sum_{t=1}^{\tau} \delta_h^{(t)} \odot \sigma(S_t) \odot M(\tanh^2(\Delta-u) - \tanh^2(\Delta+u)) \\ &\quad \Delta = W_{ox}X_t + W_{oh}h_{t-1} + b_o\end{aligned} \tag{18}$$

REFERENCES

[1] E. Erzin, "Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings," IEEE Transactions on Audio Speech & Language Processing, vol. 17, no. 7, pp. 1316–1324, 2009.

[2] T. Ito, "Bone conduction thresholds and skull vibration measured on the teeth during stimulation at different sites on the human head," Audiology & Neurotology, vol. 16, no. 1, pp. 12–22, 2011.

[3] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," Proceedings of Icassp April, vol. 5, no. 21, pp. V–708–11 vol.5, 2003.

[4] H. S. Shin, H. G. Kang, and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphone," in Speech Communication; 10. ITG Symposium; Proceedings of, 2012, pp. 1–4.

[5] Y. Zheng, Z. Liu, Z. Zhang, and M. Sinclair, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," Proc Asru, vol. 19, no. 19, pp. 249–254, 2003.

[6] M. S. Rahman and T. Shimamura, "Intelligibility enhancement of bone conducted speech by an analysis-synthesis method," Midwest Symposium on Circuits and Systems, vol. 47, no. 10, pp. 1–4, 2011.

TABLE 1. Objective Evaluation Results of DNN, LSTM, our model with single feature (Ours1) and our model with multiple features combined (Ours2).

Person	PESQ					STOI					LSD				
	BC	DNN	LSTM	Ours1	Ours2	BC	DNN	LSTM	Ours1	Ours2	BC	DNN	LSTM	Ours1	Ours2
male1	1.775	2.412	2.513	2.872	2.865	0.561	0.722	0.753	0.786	0.784	1.752	1.081	1.058	1.035	1.037
male2	1.809	2.189	2.212	2.492	2.501	0.506	0.663	0.714	0.743	0.752	1.691	1.254	1.224	1.187	1.179
male3	2.285	2.566	2.807	2.877	2.859	0.547	0.725	0.743	0.778	0.781	1.786	1.131	1.127	1.096	1.091
male4	2.119	2.589	2.615	2.945	2.930	0.609	0.745	0.769	0.819	0.816	1.599	1.147	1.138	1.098	1.102
male5	1.988	2.201	2.297	2.447	2.442	0.549	0.672	0.708	0.831	0.829	2.215	1.136	1.121	1.081	1.074
male6	1.675	2.095	2.297	2.567	2.545	0.532	0.701	0.734	0.787	0.795	1.741	1.315	1.298	1.093	1.095
male7	1.839	2.201	2.293	2.483	2.476	0.498	0.686	0.698	0.805	0.816	1.633	1.154	1.144	1.118	1.112
male8	1.798	2.213	2.324	2.581	2.569	0.537	0.717	0.735	0.815	0.822	1.542	1.182	1.162	1.069	1.071
Average	1.911	2.308	2.420	2.658	2.648	0.542	0.704	0.732	0.795	0.799	1.745	1.175	1.159	1.097	1.095

[7] 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.

[8] 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.

[9] Bouserhal R E, Falk T H, Voix J. "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," Journal of the Acoustical Society of America, 2017, 141(3):1321.

[10] K. Kondo, T. Fujita and K. Nakagawa, "On Equalization of Bone Conducted Speech for Improved Speech Quality," 2006 IEEE International Symposium on Signal Processing and Information Technology, Vancouver, BC, 2006, pp. 426-431. doi: 10.1109/ISSPIT.2006.270839

[11] Shimamura T, Tamiya T. "A reconstruction filter for bone-conducted speech," Circuits and Systems, 2005. Midwest Symposium on. IEEE, 2005:1847-1850 Vol. 2.

[12] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," Signal Processing Letters IEEE, vol. 10, no. 3, pp. 72–74, 2003.

[13] Z. Zhang, Z. Liu, M. Sinclair, and A. Acero, "Multi-sensory microphones for robust speech detection, enhancement and recognition," in IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings, 2004, pp. iii–781–4 vol.3.

[14] T. Shimamura and T. Tamiya, "A reconstruction filter for bone-conducted speech," in Circuits and Systems, 2005. Midwest Symposium on, 2005, pp. 1847–1850 Vol. 2.

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

[16] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," Speech Communication, vol. 88, pp. 65–82, 2017.

[17] A. Shahina and B. Yegnanarayana, Mapping speech spectra from throat microphone to close-speaking microphone: a neural network approach. Hindawi Publishing Corp., 2007.

[18] T. T. Vu, M. Unoki, and M. Akagi, "A study on an lp-based model for restoring bone-conducted speech," in International Conference on Communications and Electronics, 2007, pp. 212–217.

[19] Xu J, Wang X L. "A Structural Identification Method Based on Recurrent Neural Network and Auto-Regressive and Moving Average Model," in Applied Mechanics & Materials, 2013, 256-259(3):2261-2265.

[20] M. A. 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.

[21] 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, 2013.

[22] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," Knowledge and Information Systems, vol. 7, no. 3, pp. 358–386, 2005.

[23] D. D. Lee, "Algorithm for non-negative matrix factorization," Proc Nips, 2001.

[24] J. Eggert and E. Korner, "Sparse coding and nmf," in IEEE International Joint Conference on Neural Networks, 2004. Proceedings, 2004, pp. 2529–2533 vol.4.

[25] M. N. Schmidt, "Speech separation using nonnegative features and sparse nonnegative matrix factorization," Computer Speech and Language, 2007.

[26] K. Gregor and Y. Lecun, "Learning fast approximations of sparse coding," in International Conference on International Conference on Machine Learning, 2010, pp. 399–406.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] Atal B S, David N. "On finding the optimum excitation for LPC speech synthesis,"[J]. Journal of the Acoustical Society of America, 1978, 63(S1):79.

[29] Tiwari V. "MFCC and its applications in speaker recognition,"[J]. International Journal on Emerging Technologies Issn, 2010.

[30] 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 Proceedings, vol. 2, pp. 749–752 vol.2, 2001.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timécfrequency weighted noisy speech," IEEE Transactions on Audio Speech and Language Processing, vol. 19, no. 7, pp. 2125–2136, 2011.

[32] J. Gray, A. and J. Markel, "Distance measures for speech processing," IEEE Trans Acoustics Speech and Signal Processing, vol. 34, no. 5, pp. 380–391, 1976.

[33] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," Eprint Arxiv, pp. 2773–2781, 2014.

[34] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," vol. 4, pp. 3104–3112, 2014.