IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 6, 2018, accepted September 26, 2018, date of publication October 4, 2018,
date of current version November 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2873728

A Novel Encoder-Decoder Model via NS-LSTM Used for Bone-Conducted Speech Enhancement

DONGJING SHAN[1], XIONGWEI ZHANG[1], CHAO ZHANG[2], AND LI LI[1]
[1]Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing 210007, China
[2]Key Laboratory of Machine Perception, Peking University, Beijing 100871, China

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

This work was supported in part by the National Natural Science Foundation of China under Grant 61671027 and Grant 61471394 and in part by the National Key Basic Research Project of China (973 Program) under Grant 2015CB352303.

ABSTRACT Bone-conducted (BC) speech can be used to communicate in a very high noise environment. In this paper, a method of improving the quality of BC 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 (LSTM) recurrent neural network 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 direct transformations via DNN and LSTM networks, and numerous criteria 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, bone-conducted speech, long short-term memory, attention mechanism, non-negative matrix factorization.

I. INTRODUCTION

Bone-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 activities, 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 [9], [11]. 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 [9], 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 and Tamiya [11] 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.* [10] proposed the short-term DFT magnitude ratio-based method, which estimated the equalization filter with a frame-by-frame basis approach, and then obtained a mean estimate by averaging. The source-filter model decomposes speech as a combination of excitation and spectral envelope filter [14]. 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) [15], [16] 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 [18] 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 bone-conducted 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. Firstly the sparse dictionary of AC speech is established by sparse non-negative matrix factorization (sparse NMF) [21]–[23], 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 network, 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 network's 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 [19] 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 (ARMA) [17] 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 network 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 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.

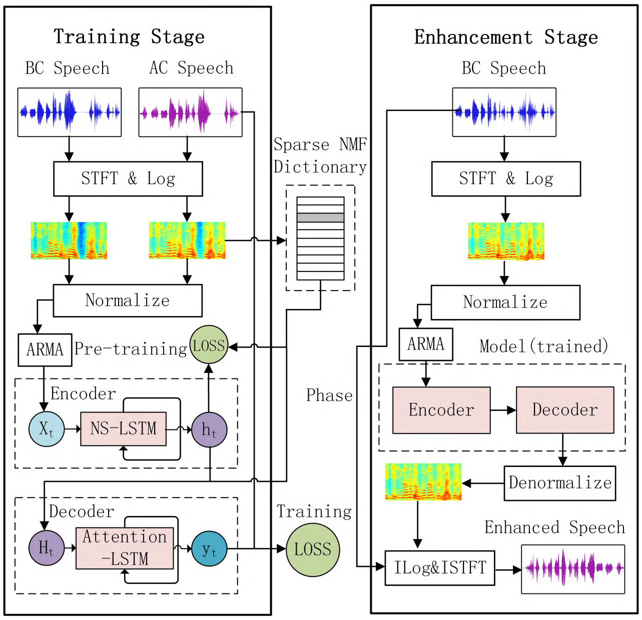


FIGURE 1. BC speech enhancement framework based on encoder-decoder model.

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, also ARMA filter is utilized sequentially. Next, the trained encoder-decoder model is used to enhance the feature vectors. The encoder transforms the features to representation coefficients on a dictionary, and the decoder generates spectral magnitudes approaching real AC speech based on the dictionary combination. In the end, the transformed 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 network 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 [24], 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:

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

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

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

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

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

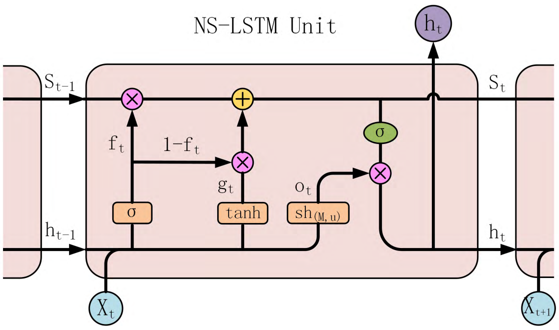


FIGURE 2. **The inner structure of NS-LSTM unit in the encoder.**

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 [24], 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 $\boldsymbol{c} = [c_1, c_2, \ldots, c_\tau]$ 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 [31], [32], 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.

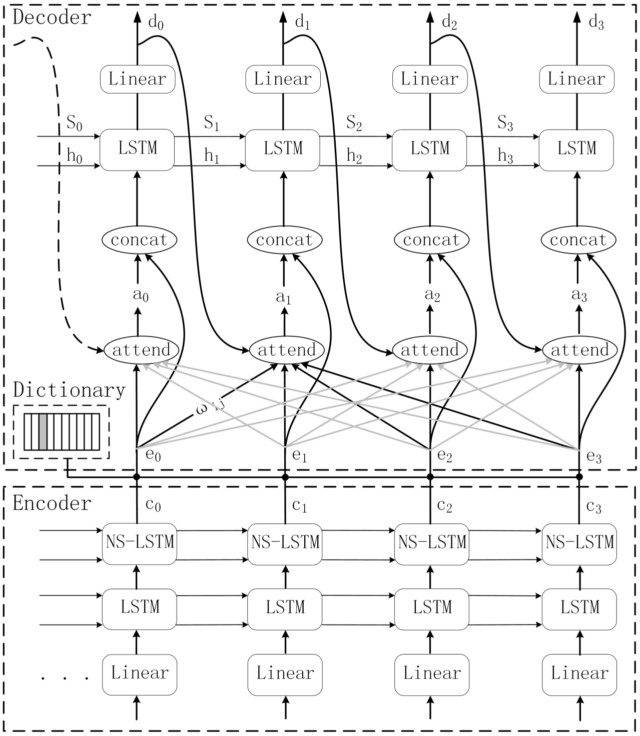


FIGURE 3. **The structure of the decoder network.**

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's LSTM layer.

The mean square error (MSE) is calculated between d_i and the ground truth, and the decoder is optimized frame by frame. Specifically, with regard to one speech frame in a sentence, loss (i.e. MSE) is back propagated to compute gradients of the model's parameters, and then, optimizers of the decoder and encoder steps respectively to update their own parameters. The process should be executed sequentially for each frame. After that, all the sentences in dataset are used for training sequentially. The training repeats a number of epoches until convergence.

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 and eight female speakers are required to read 200 sentences selected from the corpus randomly, and the speech materials 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 utterances are recorded at 32-kHz sampling rate, and the corpus is open on the web site: https://github.com/cvcoding/BC-Speech-Dataset.

B. EXPERIMENTAL SETUP

In our experiments, we train an enhancement model for each speaker. The duration of training speech is about 11 minutes, while the testing data is about 3 minutes. Both of the 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 network is trained by using Adaptive Moment Estimation (Adam) optimizer, the dropout [25] 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. In the next stage the decoder network is 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 magnitude 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) [27] and LPC (13-dimensional) [26] are normalized by feature scaling respectively, then they are concatenated to formulate a 155-dimensional features, lastly normalization is performed and the results are used as the networks' input, the networks' output are also spectral magnitude features which are compared to the ground truth to calculate loss. Additionally, we use the model to construct the relationship of spectrum parts 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) [28], Short-Time Objective Intelligibility [29] (STOI) and Log-Spectral Distance [30] (LSD) are used to measure 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. Moreover, ABX preference test is utilized to evaluate the results subjectively.

C. RESULTS AND ANALYSIS

1) RESULTS WITH SPECTRAL MAGNITUDE FEATURES

Table 1 is the objective evaluation results about DNN network, LSTM network 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 comprising male and female speech.

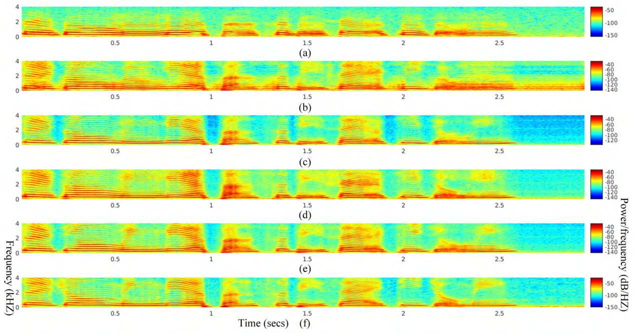


FIGURE 4. **Spectrograms of one male utterance. (a) BC speech, (b) speech enhanced by DNN, (c) speech enhanced by LSTM, (d) speech enhanced by our method, (e) speech enhanced by our method with multiple feature, (f) AC speech.**

"BC" column is the evaluation of BC compared with AC speech. We can see that majorities of PESQ scores are under 2.2 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. From the figures

TABLE 1. Objective Evaluation Results of DNN, LSTM, NS-LSTM model with single feature (Ours1), NS-LSTM model with combined features (Ours2) and NS-LSTM single feature model with estimated phase (Ours3).

Person	PESQ						STOI						LSD					
	BC	DNN	LSTM	Ours1	Ours2	Ours3	BC	DNN	LSTM	Ours1	Ours2	Ours3	BC	DNN	LSTM	Ours1	Ours2	Ours3
male1	2.119	2.589	2.615	2.945	2.930	2.932	0.609	0.745	0.769	0.819	0.821	0.823	1.599	1.147	1.138	1.098	1.102	1.095
male2	1.988	2.201	2.297	2.447	2.442	2.440	0.549	0.672	0.708	0.831	0.832	0.836	2.215	1.136	1.121	1.081	1.074	1.076
male3	1.839	2.201	2.293	2.483	2.476	2.491	0.498	0.686	0.698	0.805	0.816	0.812	1.633	1.154	1.144	1.118	1.112	1.107
male4	2.285	2.566	2.807	2.877	2.859	2.862	0.547	0.725	0.743	0.778	0.785	0.790	1.786	1.131	1.127	1.096	1.091	1.087
Average1	2.058	2.389	2.503	2.688	2.677	2.681	0.551	0.707	0.730	0.808	0.814	0.815	1.808	1.142	1.133	1.098	1.095	1.091
female1	1.786	2.413	2.515	2.876	2.869	2.891	0.572	0.734	0.761	0.786	0.791	0.782	1.763	1.093	1.072	1.045	1.048	1.041
female2	1.824	2.211	2.226	2.508	2.515	2.524	0.521	0.675	0.728	0.758	0.767	0.764	1.706	1.269	1.241	1.203	1.195	1.205
female3	1.907	2.314	2.427	2.684	2.689	2.690	0.640	0.817	0.836	0.913	0.921	0.918	1.645	1.288	1.264	1.173	1.169	1.181
female4	1.762	2.176	2.364	2.622	2.632	2.628	0.619	0.781	0.824	0.874	0.883	0.891	1.828	1.402	1.375	1.178	1.182	1.186
Average2	1.820	2.279	2.383	2.673	2.676	2.683	0.588	0.752	0.787	0.833	0.841	0.839	1.736	1.263	1.238	1.150	1.149	1.153

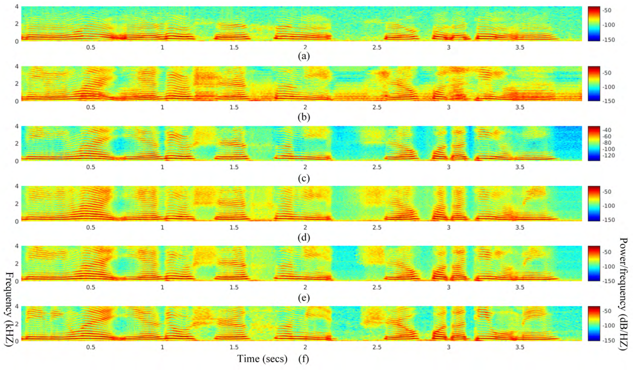


FIGURE 5. Spectrograms of one female utterance. (a) BC speech, (b)speech enhanced by DNN, (c) speech enhanced by LSTM, (d) speech enhanced by our method with single feature, (e) speech enhanced by our method with multiple feature, (f) AC speech.

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. Among three models, our proposed one scores much better than others on the three metrics. DNN ranks last because it stacks linear layers simply and ignores the inherent sequential relationship of the sentences. LSTM is a kind of recurrent neural network and capable of utilizing the former information. Our model ranks the first owe to its recurrent structure, dictionary representation and local attention mechanism. Additionally, the male speech scores better than female because female voice is more challenging for the existence of more high-frequency components. Our model still achieves the best with the female speech data. 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 network 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 replaces NS-LSTM units by conventional LSTM units. The network 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.

4) ABX PREFERENCE TEST

In the ABX preference test, twenty listeners (ten males and ten females) are asked to choose which sample (A or B) sounds more similar to X, if they can not distinguish between the two, no preference (N/P) can be selected. Forty testing sentences are evaluated and the results are depicted in Fig. 6. We conduct four sets of comparative experiments: DNN with Ours1 (NS-LSTM with single feature), LSTM with Ours1, Ours1 with Ours2 (NS-LSTM with combined features), and Ours1 with Ours3 (NS-LSTM with estimated phase). P-values are used to determine the significance of the results, the small p-value indicates large significance and vice versa.

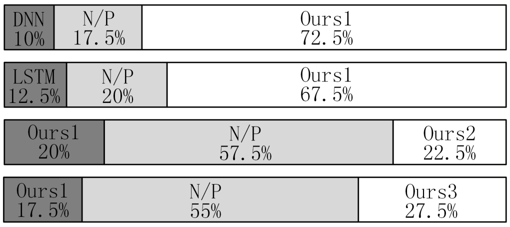


FIGURE 6. ABX preference test results, The p-values of the four pairs are 6.64×10^{-7}, 1.89×10^{-5}, 0.8118 and 0.35.

From the first two bars, we can see that our model behaves much better than DNN and LSTM. The third bar shows that our model works at the similar levels with single feature or combined features. The fourth bar indicates that our model performs a little better when the phase transformation is utilized.

VI. CONCLUSION

In this paper, we propose an encoder-decoder based bone-conducted speech enhancement framework, in which the encoder via non-negative and sparse LSTM network 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 components. 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:

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

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

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

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

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

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

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

REFERENCES

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

[2] T. Ito, C. Röösli, C. J. Kim, J. H. Sim, A. M. Huber, and R. Probst, "Bone conduction thresholds and skull vibration measured on the teeth during stimulation at different sites on the human head," *Audiol. Neurotol.*, 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," in *Proc. ICASSP*, vol. 5, Apr. 2003, pp. V-708–V-711.

[4] H. S. Shin, H.-G. Kang, and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphone," in *Proc. 10th ITG Symp. Speech Commun.*, 2012, pp. 1–4.

[5] Y. Zheng *et al.*, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *Proc. ASRU*, 2003, pp. 249–254.

[6] M. S. Rahman and T. Shimamura, "Intelligibility enhancement of bone conducted speech by an analysis-synthesis method," in *Proc. Midwest Symp. Circuits Syst.*, 2011, pp. 1–4.

[7] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 9, pp. 2505–2517, Nov. 2012.

[8] M. Mcbride, P. Tran, T. Letowski, and R. Patrick, "The effect of bone conduction microphone locations on speech intelligibility and sound quality," *Appl. Ergonom.*, vol. 42, no. 3, pp. 495–502, 2011.

[9] R. E. Bouserhal, T. H. Falk, and J. Voix, "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," *J. Acoust. Soc. Amer.*, vol. 141, no. 3, p. 1321, 2017.

[10] K. Kondo, T. Fujita, and K. Nakagawa, "On Equalization of Bone Conducted Speech for Improved Speech Quality," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Vancouver, BC, Canada, Aug. 2006, pp. 426–431, doi: 10.1109/ISSPIT.2006.270839.

[11] T. Shimamura and T. Tamiya, "A reconstruction filter for bone-conducted speech," in *Proc. 48th Midwest Symp. Circuits Syst.*, vol. 2, 2005, pp. 1847–1850.

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

[13] Z. Zhang, Z. Liu, M. Sinclair, and A. Acero, "Multi-sensory microphones for robust speech detection, enhancement and recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, May 2004, pp. III-781–III-784.

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

[15] 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. 2007, p. 087219, Dec. 2007.

[16] T. T. Vu, M. Unoki, and M. Akagi, "A study on an LP-based model for restoring bone-conducted speech," in *Proc. Int. Conf. Commun. Electron.*, 2007, pp. 212–217.

[17] J. Xu and X. L. Wang, "A structural identification method based on recurrent neural network and auto-regressive and moving average model," *Appl. Mech. Mater.*, vols. 256–259, no. 3, pp. 2261–2265, 2013.

[18] M. A. T. Turan and E. Erzin, "Source and filter estimation for throat-microphone speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 2, pp. 265–275, Feb. 2016.

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

[20] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, 2005.

[21] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2001, pp. 535–541.

[22] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 4, Jul. 2004, pp. 2529–2533.

[23] M. N. Schmidt, "Speech separation using non-negative features and sparse non-negative matrix factorization," *Interspeech*, pp. 19–33, Jun. 2007.

[24] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 399–406.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] B. S. Atal and N. David, "On finding the optimum excitation for LPC speech synthesis," *J. Acoust. Soc. Amer.*, vol. 63, no. S1, p. 79, 1978.

[27] V. Tiwari, "MFCC and its applications in speaker recognition," *Int. J. Emerg. Technol.*, vol. 1, no. 1, pp. 19–22, 2010.

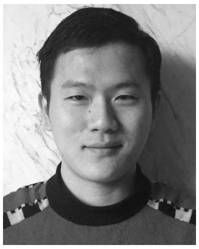
[28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[30] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. ASLP-24, no. 5, pp. 380–391, Oct. 1976.

[31] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2773–2781.

[32] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, 2014, pp. 3104–3112.



DONGJING SHAN received the B.Eng. degree in computer science from the Beijing Institute of Technology, Beijing, China, in 2008, and the M.S. degree in information science and technology from Peking University, Beijing, in 2013.

He is currently pursuing the Ph.D. degree in signal processing with the Laboratory of Intelligent Information Processing, Army Engineering University. His research interests include speech recognition, image processing, and machine learning.

XIONGWEI ZHANG received the Ph.D. degree in signal and information processing from the Nanjing Institute of Communications Engineering, Nanjing, China, in 1992.

He is currently a Professor with the Laboratory of Intelligent Information Processing, Army Engineering University. His research interests include speech signal processing, machine learning, and pattern recognition.

CHAO ZHANG received the Ph.D. degree in electrical engineering from Beijing Jiaotong University, Beijing, China, in 1995.

He is currently an Associate Professor with the Department of Machine Intelligence, School of Electronics Engineering and Computer Science, Peking University. His research interests include computer vision, image processing, and pattern recognition.

LI LI received the M.S. degree in signal processing from the PLA University of Science and Technology, Nanjing, China, in 2003.

She is currently a Researcher with the Laboratory of Intelligent Information Processing, Army Engineering University. Her research interests include speech signal and image processing, pattern recognition, and machine learning.