Semi-sparse and Non-negative Long Short-Term Memory based Dictionary Representation for Throat Microphone Speech Enhancement

Dongjing Shan, Xiongwei Zhang and Chao Zhang

***Abstract*—Throat microphone (TM) speech is immune to noise but has some shortcomings such as severe loss of high-frequency components. As direct transformation is not sufficient to achieve satisfactory performance, we propose a dictionary representation based network model used for TM speech enhancement in this paper. Specifically, a magnitude spectrums dictionary of air-conducted speech is calculated via sparse non-negative matrix factorization (SNMF), and then it is used to represent the transformed speech in hidden layer. Meanwhile, a compensating dictionary is adopted to promote the representation accuracy. The Semi-sparse and Non-negative Long Short-Term Memory (SSN-LSTM) recurrent neural network is designed and deployed to generate the combination coefficients, with sparse fragment on the SNMF dictionary and non-sparse one on the other. Lastly, a three-layer neural network comprising SSN-LSTM is established as the enhancement model. In the experiments, our method and its variants are evaluated and compared with other direct enhancing models. Numerous criterions are adopted to measure the performance, the objective and subjective results can demonstrate the superiority of our proposed method.**

***Index Terms*—Speech enhancement, throat microphone speech, long short-term memory, non-negative matrix factorization.**

I. INTRODUCTION

RECENTLY, numerous enhancement methods have been developed to improve the quality and intelligibility of air-conducted (AC) speech by using single or multiple AC microphones. However, the improvement may be limited in case of non-stationary noise and strong background noise [1].

Throat microphone (TM) can cope with the problem efficiently because it is a skin-attached non-acoustic sensor and collects vibration signals directly from the throat skin. Though TM speech is immensely robust even in severely degraded environments [2], its intelligibility is lower than AC speech and the quality needs to be improved. In the first place, it faces severe loss of high-frequency components due to the attenuation of human body channel [3], [4]. Secondly, some phonemes like unvoiced fricatives, plosives and affricates are totally lost [5], which are usually generated by oral cavity rather than vocal cord. Moreover, the frequency characteristic and sound quality vary depending on the microphone location, which further increase the difficulty of TM speech enhancement.

In most cases, throat microphone plays an auxiliary role in improving the enhancement performance of AC speech in noisy environments [8], [10]; on the other hand, AC microphone is utilized to help enhance TM or other bone-conducted speech [11], [12]. Nevertheless, in some cases it is meaningful to enhance the TM speech independently, because AC speech can be completely unintelligible under some extreme conditions, such as strong noisy environments like driver cabins, or AC microphone is not convenient or secure to be worn in some special occasions like extreme sports.

Several approaches based on bandwidth extension, equalization and source-filter models have been proposed in this field. In bandwidth extension models, the low-frequency spectrum is expended directly to recover the high-frequency spectrum as they are assumed to have the same harmonic structures. Equalization methods aim to calculate inverse transformation function of the transmission channel. Shimamura [11] firstly proposed a linear-phase impulse response filter, which performed an inverse discrete Fourier transform on the ratio of AC and TM speech spectra. Kondo et al. [9] proposed a short-term DFT based method, which estimated a frame-by-frame equalization filter and calculate a mean estimate by averaging. Source-filter methods model speech as a combination of excitation and spectral envelope filter [13]. With the assumption that the excitation is unchanged between AC and TM speech, these approaches usually transform the Linear Predictive (LP) family parameters [14] by Gaussian mixture models (GMMs). 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 [15] has trained distinctive GMMs for different types of phones. Nevertheless, how to recognize phones effectively remains challenging.

In this paper, we propose a signal-based analysis/synthesis model, specifically it enhances the short-time spectral magnitude of TM speech via a three-layer neural network and then synthesizes enhanced speech with the phase information unchanged. The main issue in the first stage lies in how to transform the high-dimensional magnitude spectrums effectively. Unlike the direct transformation from TM magnitudes to AC ones, we calculate the transform coefficients of TM speech on dictionaries and then construct AC speech by linear combination. Concretely, we establish an AC speech dictionary

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

D. Shan and X. Zhang are with Laboratory of Intelligent Information Processing , Army Engineering University, NanJing, 210007, P. R. China. e-mail: shandongjing@pku.edu.cn.

C. Zhang is with Key Laboratory of Machine Perception (MOE), Peking University, Beijing, 100871, P. R. China.

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

Manuscript received December 10, 2018.

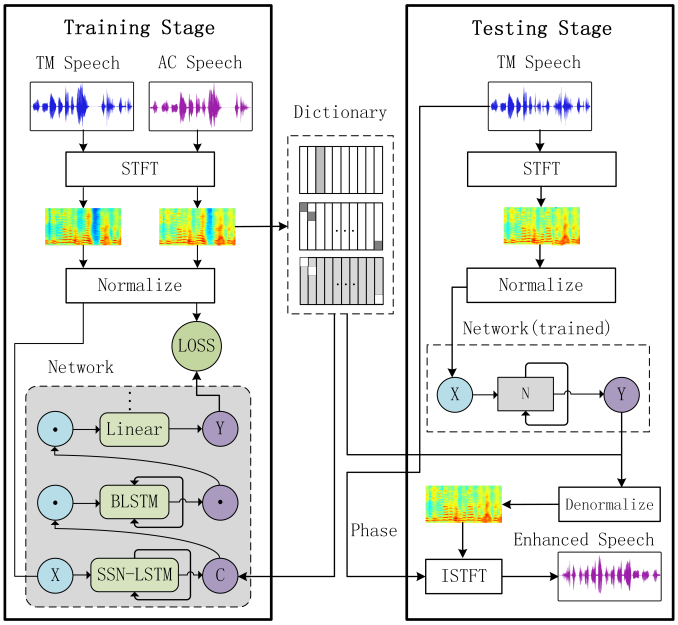

Fig. 1. TM speech enhancement framework based on dictionary representation. The network is depicted inside a gray bounding box in the left side of the figure, and simplified as a little box with "N" tag in the right side, in which X=$\{x_1, ..., x_t, ..., x_T\}$ represents the input spectral features of T frames, Y=$\{y_1, ..., y_t, ..., y_T\}$ indicates the corresponding output features, and C means the hidden states.

by sparse non-negative matrix factorization, then we design a LSTM-RNN structure to compute sparse and non-negative coefficients on this dictionary. Moreover, a compensating dictionary is appended to minimize the transformation error and make the representation more accurate. The coefficients on this dictionary is non-sparse and non-negative. So our proposed SLTM-RNN structure is named Semi-Sparse and Non-negative Long Short-Term Memory network (SSN-LSTM) for the reason that it generates sparse and non-sparse coefficients simultaneously on two different dictionaries. The combination results are passed forward to the next two network layers: a bidirectional LSTM (BLSTM) and a Linear layer. In the next synthesis stage, the phase information is extracted from TM speech without enhancement, because phase features cannot be enhanced by NMF dictionary (can be enhanced by network like LSTM). It is sufficient to evaluate the performance with other algorithms, when we keep the phase unchanged and utilize it in all the comparative methods.

The rest of the paper is organized as follows. The T-M speech enhancement framework is presented in the next section, then SSN-LSTM and the speech dictionaries are described in Section III. The parallel dataset of TM/AC speech and experiments are illustrated in Section IV. Nextly, a conclusion is drawn in Section V, and finally, the gradient deduction of SSN-LSTM is attached in the appendix.

II. TM Speech Enhancement Framework

Our proposed framework is illustrated in Fig.1. In the training stage, spectral magnitudes of AC and TM speech are computed frame by frame using short-term Fourier transform (STFT), and then the spectral features are normalized to a standard normal distribution to facilitate the training of neural network, the mean and variance are recorded subsequently. Meanwhile, an AC speech dictionary is computed by using sparse non-negative matrix factorization (SNMF) [18], [19], [20], and a compensating dictionary is established for more accurate representation. Next, the spectral features of TM are sent to the network for training, which is depicted in the gray bounding box of Fig.1. For limitation of the space, only a compressed network is presented here, and the network unfolded frame by frame is exhibited in Fig.2. The first layer SSN-LSTM generates coefficients on the dictionaries, the combination feature is then forwarded to the next two layers to get the final transformed feature. Huber loss [6] is adopted as loss function here, it indicates the difference between enhanced spectrums and the targets from AC speech, and is more robust than squared error loss when cope with outliers in data. The function is quadratic for small values of difference, and linear for large values, it can be defined piecewise as follows:

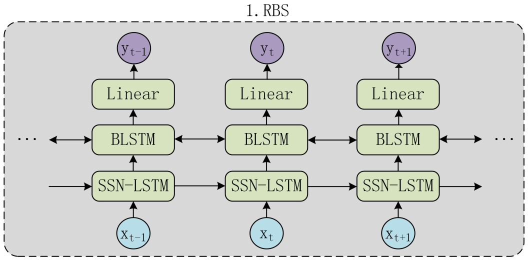

Fig. 2. The network used for spectral magnitudes enhancement is unfolded frame by frame through time. x_t symbolizes the spectral feature of the t-th frame in the TM speech, and y_t indicates the t-th feature in the enhanced speech. "RBS" symbolizes Linear BLSTM SSN-LSTM.

$$\text{Hloss}(y_{ti}, \widehat{y}_{ti}) = \begin{cases} \frac{1}{2}(y_{ti} - \widehat{y}_{ti})^2 & if \ |y_{ti} - \widehat{y}_{ti}| < 1 \\ |y_{ti} - \widehat{y}_{ti}| - \frac{1}{2}, & otherwise \end{cases} \tag{1}$$

where y_{ti} means the i-th element in the t-th frame of the output enahnced spectral feature, and $\widehat{\text{y}}_{ti}$ has the same meaning with regard to the target feature.

In the testing stage, the magnitude and phase of TM speech are firstly computed, then the spectral magnitudes are normalized according to the recorded mean and variance of TM speech. Next, trained enhancement model is utilized to transform spectral features from TM speech to AC speech with the assistance of dictionaries. After that, the generated AC spectral magnitudes are denormalized by the mean and variance recorded in the training stage. Finally, the enhanced AC speech is synthesized based on the latest magnitudes and TM phase via inverse STFT (symbolized as ISTFT in Fig.1).

III. Semi-sparse and Non-negative LSTM

A. Speech Dictionary

As we know, TM speech losses high-frequency components severely. It causes that the generated AC speech degrades in high frequency region if direct transformation is adopted. Thus, we take use of indirect transform framework exhibited in Section II, in which the dictionaries $M = [F, E, -E]$ is

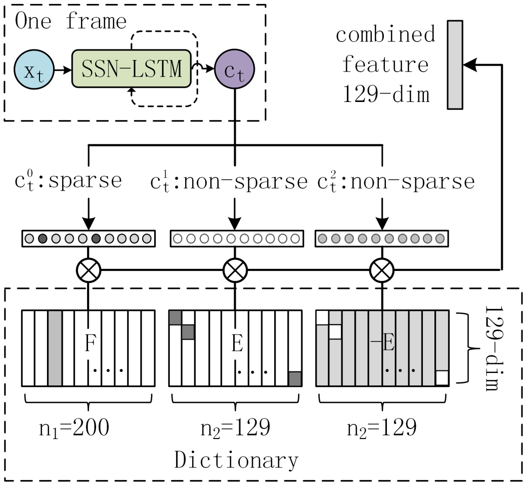

Fig. 3. The structure of dictionaries. One SNMF dictionary containing 200 elements in the experiments, one compensating dictionary and an inverse version. c_t is the generated coefficients, and "⊗" symbolizes element wise multiplication.

utilized for spectral feature synthesis. In Fig.3, F is a spectral feature dictionary comprising a number of 129-dimensional elements (also the upper one in the dictionary box of Fig.1), and it is computed by SNMF as mentioned before. The coefficients on F is constrained to be sparse and non-negative. E is a compensating dictionary (the middle one in the dictionary box of Fig.1 and Fig.3), in which each column has only one nonzero element (0.1) on the diagonal line and the column number is identical with the length of spectral feature. The coefficients on this dictionary is non-sparse, but is forced to be non-negative to remain consistent with the former ones. Therefore, a negative compensating dictionary -E is needed to provide inverse compensation. SSN-LSTM layer generates the desired coefficients, and then the dictionary representation via linear combination is passed forward to the next layer, in which an abundance of high-frequency remains for the reason of SNMF dictionary.

B. SSN-LSTM Network

The neural network is composed of three layers, the bottom one is a SSN-LSTM layer, the middle one original BLSTM and the top one Linear layer. SSN-LSTM aims to generate coefficients on the dictionary elements, the output should satisfy two conditions: semi-sparse and non-negative. We define the loss function as follows:

$$\begin{aligned}&\min_{C} L(C), \quad s.t.\ \ C \geq 0\\ &\mathrm{L(C)} = \mathrm{Hloss}(\psi_2(\psi_1(MC)), \widehat{Y}) + \lambda ||\mathrm{C}^0||_1^2\end{aligned} \tag{2}$$

where ψ_1 and ψ_2 represent the processes of BLSTM and Linear layers respectively. Our emphasis will place on SSN-LSTM network, so we use symbols represent other two layers. $C = [c_1, c_2, ..., c_T]$ is an output matrix comprising T coefficient vectors. From Fig.3 we can see that $c_t = [c_t^0, c_t^1, c_t^2]$, in which c_t^0 corresponds to the SNMF dictionary and the ones over T time points make up the matrix C^0 . $\widehat{\mathrm{Y}}$ is the spectral feature of corresponding AC speech. The Hloss term enforces the enhanced feature close to the ground truth, $||\cdot||_1$ symbolizes ℓ_1-norm and ensures the sparsity, C is constrained to be non-negative.

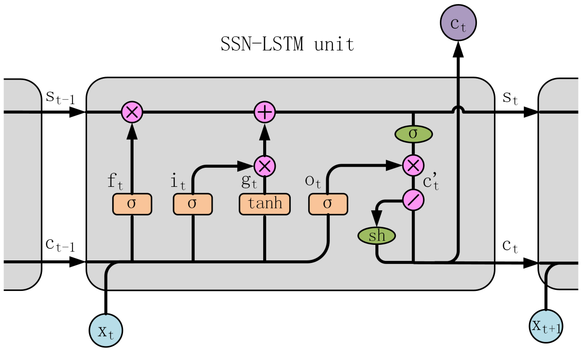

Fig. 4. The structure of SSN-LSTM unit. σ symbolises sigmoid activation function, σ in green elliptical box is used to guarantee non-negative constraint; the sh elliptical box represents an nonlinear shrinkage function, it is used to ensure sparsity of the coefficients on SNMF dictionay, and its bypass line transmits non-sparse and non-negative coefficients on the compensating dictionaries; ⊘ means the vector is divided into two branches.

Inspired by the work in [21], where a simple recurrent neural network was proposed to derive sparse coding, we design a SSN-LSTM network to impose the above constrains on the coefficients and present the unit structure in Fig.4. The network's forward propagation can be formulated as follows:

$$f_t = \sigma(W_{fx}x_t + W_{fc}c_{t-1} + b_f) \tag{3}$$

$$i_t = \sigma(W_{ix}x_t + W_{ic}c_{t-1} + b_i) \tag{4}$$

$$g_t = \phi(W_{gx}x_t + W_{gc}c_{t-1} + b_g) \tag{5}$$

$$o_t = \sigma(W_{ox}x_t + W_{oc}c_{t-1} + b_o) \tag{6}$$

$$s_t = g_t \odot i_t + s_{t-1} \odot f_t \tag{7}$$

$$c'_t = \sigma(s_t) \odot o_t \tag{8}$$

$$c_t = \left[sh_{(D,u)}(c'_t[1:n_1]), c'_t[n_1+1:n_1+2n_2]\right]^{\mathrm{T}} \tag{9}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$, $\phi(x) = \tanh(x)$; c'_t satisfies non-negative constraint; n_1, n_2 corresponds to the element numbers of SNMF and compensating dictionaries respectively (see Fig.3). $sh_{(D,u)}(x) = D(tanh(x+u) + tanh(x-u))$ is the so-called "double tanh" function [21] (see Fig.5), in which D is a trainable diagonal matrix and initialized as $\frac{1}{2} \times \mathrm{I}$ (a $n_1 \times n_1$ identity matrix) in the experiments, when updating the network parameters in training process, D maintains its diagonal property. u is a trainable $n_1 \times 1$ vector with all the elements initialized as 2. This shrinkage activation function is used to keep sparsity of the coefficients.

The parameters are updated through backward propagation to minimize the loss. In order to make the loss function differentiable with the matrix C, we relax it to the following Equations:

$$\mathrm{L(C)} = \mathrm{Hloss}(\psi_2(\psi_1(MC)), \widehat{Y}) + \lambda ||\mathrm{C}^0||_F^2 \tag{10}$$

In the backpropagation process, the gradients are calculated based on the chain rule, and we put our emphasis on the derivation of SSN-LSTM parameters. The gradients of two hidden vectors are computed as a prerequisite:

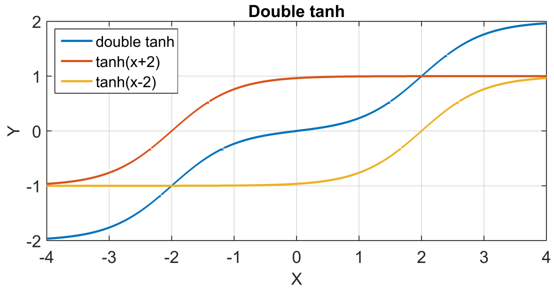

Fig. 5. A plot of double tanh function with one-dimensional variables, in which D=1, u=2.

1) $\delta_c^{(t)}$:

$$\delta_c^{(t)} = \frac{\partial L}{\partial c_t} = \left[\frac{\partial L}{\partial c_t^0}, \frac{\partial L}{\partial c_t^1}, \frac{\partial L}{\partial c_t^2}\right]^{\mathrm{T}} \tag{11}$$

$$\begin{aligned} \tfrac{\partial L}{\partial c_t^0} &= F^{\mathrm{T}} Hloss'(\psi_2\psi_1(MC))\psi'_2(\psi_1(MC))\psi'_1(MC) \\ &\qquad +2\lambda c_t^0 \\ \tfrac{\partial L}{\partial c_t^1} &= E^{\mathrm{T}} Hloss'(\psi_2\psi_1(MC))\psi'_2(\psi_1(MC))\psi'_1(MC) \\ \tfrac{\partial L}{\partial c_t^2} &= -\tfrac{\partial L}{\partial c_t^1} \end{aligned}$$

2) $\delta_s^{(t)}$:

$$\begin{aligned} \delta_s^{(t)} &= \tfrac{\partial L}{\partial \mathrm{s}_t} = \tfrac{\partial L}{\partial s_{t+1}} \tfrac{\partial s_{t+1}}{\partial s_t} + \tfrac{\partial L}{\partial c_t} \tfrac{\partial c_t}{\partial \mathrm{s}_t} \\ &= \delta_s^{(t+1)} \odot f_{t+1} + \delta_c^{(t)} \tfrac{\partial c_t}{\partial \mathrm{s}_t} \end{aligned} \tag{12}$$

$$\begin{aligned} (i) \qquad & s_t^0 = s_t[1:n_1],\ s_t^1 = s_t[n_1+1:n_1+n_2], \\ & s_t^2 = s_t[n_1+n_2+1:n_1+2n_2] \\ & o_t^0 = o_t[1:n_1],\ o_t^1 = o_t[n_1+1:n_1+n_2], \\ & o_t^2 = o_t[n_1+n_2+1:n_1+2n_2] \\ (ii) \qquad & \frac{\partial \mathrm{c}_t}{\partial s_t} = \left[\frac{\partial \mathrm{c}_t^0}{\partial s_t^0}, \frac{\partial \mathrm{c}_t^1}{\partial s_t^1}, \frac{\partial \mathrm{c}_t^2}{\partial s_t^2}\right]^{\mathrm{T}} \\ & \frac{\partial \mathrm{c}_t^0}{\partial s_t^0} = sh'(\sigma(s_t^0) \odot o_t^0)\left[\sigma(s_t^0)(1-\sigma(s_t^0)) \odot o_t^0\right]^{\mathrm{T}} \\ & sh'(x) = D(2 - \tanh^2(x+u) - \tanh^2(x-u)) \\ & \frac{\partial \mathrm{c}_t^1}{\partial s_t^1} = \sigma(s_t^1)(1-\sigma(s_t^1)) \odot o_t^1 \\ & \frac{\partial \mathrm{c}_t^2}{\partial s_t^2} = \sigma(s_t^2)(1-\sigma(s_t^2)) \odot o_t^2 \end{aligned}$$

where $\odot$ means Hadamard product [26]. The parameters' gradients can be calculated based on the above ones. We list one of them below and present others in the appendix.

$$\begin{aligned} \tfrac{\partial L}{\partial W_{fx}} &= \sum_{t=1}^{T} \tfrac{\partial L}{\partial s_t} \tfrac{\partial s_t}{\partial f_t} \tfrac{\partial f_t}{\partial W_{fx}} \\ &= \sum_{t=1}^{T} \delta_s^{(t)} \odot s_{t-1} \odot f_t \odot (1-f_t) x_t^{\mathrm{T}} \end{aligned} \tag{13}$$

IV. Experiments

A. TM Speech Dataset

We use a thousand sentences as corpus, and each of them lasts for about 3 to 5 seconds. Ten male and ten female speakers are required to read 200 different sentences selected from the corpus randomly, the speeches are recorded at 32-kHz sampling rate by air-conducted and throat-conducted microphones simultaneously, among which, 160 sentences of each person are used for training and the rest for testing. Our dataset combined with the code is freely available on the web site: https://github.com/cvcoding/TM.

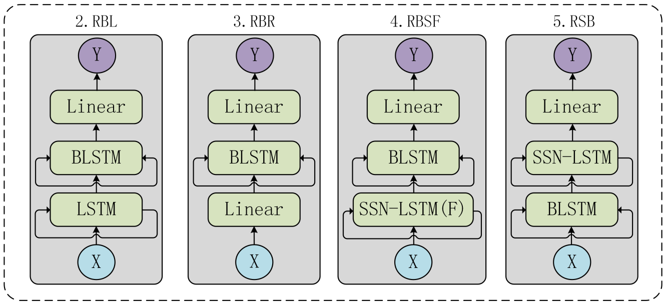

Fig. 6. Four models used for comparison, R: Linear, L: LSTM, B: BLSTM, S: SSN-LSTM. RBSF indicates the model trained with only SNMF dictionary "F".

B. Experimental Setup

In our experiments, we train an enhancement model of the spectral magnitude feature for each speaker. A speech sentence is divided into 32ms speech frames with 10ms frame shift. Both the training and testing data are down sampled to 8 kHz, and then a 129-dimensional spectral magnitude is extracted as the feature with regard to each frame.

Four models (Fig.6) are used for comparison. In RBL and RBR, the SSN-LSTM layer of our model (RBS in Fig.2) is replaced with LSTM and Linear respectively; in RBSF, the SSN-LSTM layer is trained with only SNMF dictionary and the compensating one is abandoned; in RSB, the SSN-LSTM and BLSTM layers are exchanged. All the networks are trained by using adaptive moment estimation (Adam) [27] optimizer, the dropout [22] 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 for 3 times. The best model parameters are chosen from the epoches according to the least validation loss. Additionally, in our SSN-LSTM model, the number of SNMF dictionary elements (n_1) is set to 200, and λ in Eqn.10 is assigned as 0.01.

Three metrics including Perceptual Evaluation of Speech Quality (PESQ) [23], Short-Time Objective Intelligibility [24] (STOI) and Log-spectral Distance [25] (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. Moreover, ABX preference test is utilized to evaluate the enhanced results subjectively.

C. Results and Analysis

Table I and Table II present the objective evaluation results with four male and four female speakers. The PESQ scores are exhibited in Table I and the other two metrics are in Table II, among which, the values rank the best are written in bold font and the second ones are identified using underlines. Fig.7, Fig.8 are the spectrograms of one male and one female speech utterances respectively. "TM" items in the two tables are the evaluation between TM and AC speech. We can see that the majorities of PESQ scores are under 2.1 and STOI scores under 0.70, which indicates the low quality and poor intelligibility of TM speech. From Fig.7(a) and Fig.8 (a), severe high-frequency components (2-4kHz) loss in TM speech can be observed, and also, its middle-frequency (1-2kHz) energy is larger than the corresponding components in AC speech. These can explain why the TM speech sounds muffled and unclear.

The restoration of high-frequency components can be seen in Fig.7 and Fig.8 obviously, which indicates the effectiveness of the models, and our designed model behaves well when it restores spectra in red cycles, while RBL and RBR models seem incapable of inferring the missing parts without auxiliary of the dictionaries. In Table I and Table II, as the low and middle-frequency components account for more proportion in male speeches than female ones, the average scores of males in three metrics are higher than females. Apparently, we can see that our model (RBS) achieves the best while our model without compensating dictionary (RBSF) ranks the second. For instance, our model (RBS) improves the average PESQ score of female TM speech from 2.10 to 2.42 and performs better than others.

In the ABX preference test, twenty persons (ten males and ten females) are asked to judge which one sounds more similar to X from A and B, if they cannot distinguish them, no preference (N/P) can be chosen. Forty sentences are selected for testing and the results are depicted in Fig.9. We conduct four pairs of comparative experiments: 1.RBS (our designed model) with 2.RBL, 1.RBS with 3.RBR, 1.RBS with 4.RBSF (our designed model with only SNMF dictionary), and 1.RBS with 5.RSB (our designed model with the layers in different order). From the first two bars, we can see that our model behaves much better than conventional recurrent networks. The third bar shows that the compensating dictionary in our model contributes to the whole performance, and the fourth bar illustrates our model behaves much better with the layers in a correct order. P-values in the caption are used to determine the significance of the results, the small p-value indicates large significance and vice versa.

TABLE I
OBJECTIVE EVALUATION RESULTS OF THE NETWORKS (PESQ)

Speakers	PESQ					
	TM	2.RBL	3.RBR	4.RBSF	5.RSB	1.RBS
male1	2.2780	2.7379	2.7242	2.7212	2.7101	**2.7531**
male2	2.3610	2.6695	2.6768	**2.6821**	2.6749	2.6781
male3	1.7140	2.1336	**2.1855**	2.0742	2.0968	2.1793
male4	1.9623	2.6046	2.6475	2.6412	2.5827	**2.6527**
AVG1	2.0788	2.5364	2.5585	2.5297	2.5161	**2.5658**
female1	2.5049	2.9702	2.9696	3.0006	3.0251	**3.0544**
female2	2.0748	2.3381	2.4767	2.4903	2.3708	**2.5045**
female3	1.7904	1.8934	1.9183	1.8250	1.8102	**1.9392**
female4	2.0357	2.1690	2.1771	**2.1820**	2.1770	2.1786
AVG2	2.1015	2.3427	2.3854	2.3745	2.3458	**2.4192**

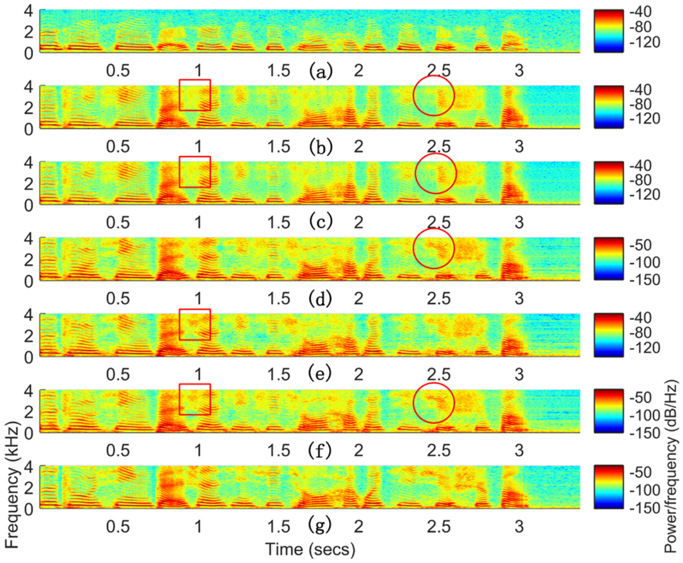

Fig. 7. Spectrograms of one male utterance. (a) TM speech. Spectrograms enhanced by: (b) RBL, (c) RBR, (d) RBSF, (e) RSB, (f) RBS. (g) AC speech.

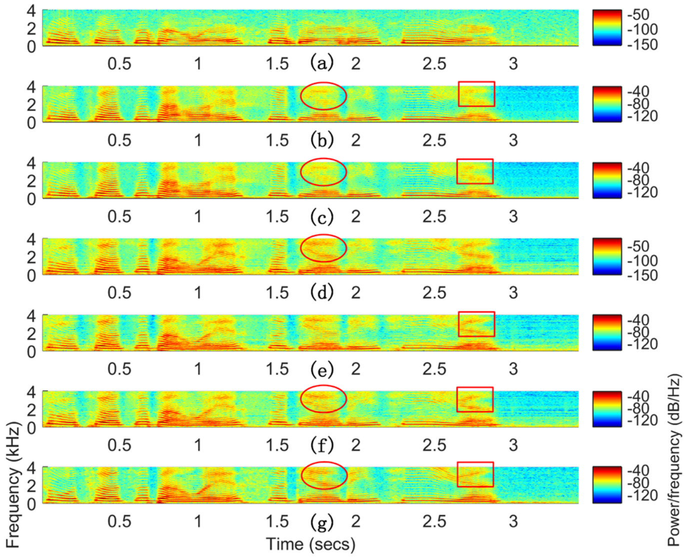

Fig. 8. Spectrograms of one female utterance. (a) TM speech. Spectrograms enhanced by: (b) RBL, (c) RBR, (d) RBSF, (e) RSB, (f) RBS. (g) AC speech.

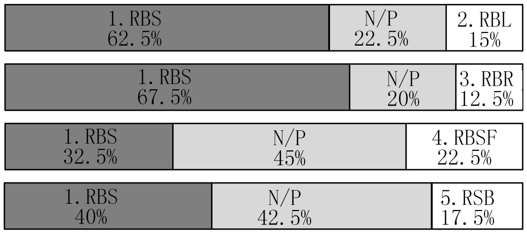

Fig. 9. ABX preference test results, The p-values of the four pairs are 2.7212×10^{-4}, 1.8880×10^{-5}, 0.4006 and 0.0596.

TABLE II
OBJECTIVE EVALUATION RESULTS OF THE NETWORKS (LSD AND STOI)

Speakers	LSD						STOI					
	TM	2.RBL	3.RBR	4.RBSF	5.RSB	1.RBS	TM	2.RBL	3.RBR	4.RBSF	5.RSB	1.RBS
male1	1.4816	$\underline{1.0496}$	1.0625	1.0770	1.0721	**1.0493**	0.6277	0.8699	$\underline{0.8707}$	0.8692	0.8679	**0.8732**
male2	1.3286	**0.9882**	1.0097	1.0198	1.0210	$\underline{1.0085}$	0.7158	0.8856	0.8886	0.8866	0.8858	**0.8927**
male3	1.4725	1.1094	$\underline{1.0986}$	1.1010	1.1064	**1.0856**	0.6724	0.8243	**0.8378**	0.8191	0.8273	$\underline{0.8290}$
male4	1.4797	0.9768	0.9706	$\underline{0.9701}$	0.9941	**0.9691**	0.5760	0.8390	$\underline{0.8451}$	0.8408	0.8423	**0.8516**
AVG3	1.4406	$\underline{1.0310}$	1.0354	1.0420	1.0484	**1.0281**	0.6480	0.8547	$\underline{0.8606}$	0.8539	0.8558	**0.8616**
female1	1.3691	0.9517	0.9675	$\underline{0.9432}$	0.9500	**0.9332**	0.6747	0.8906	0.8919	$\underline{0.8946}$	0.8930	**0.9001**
female2	1.4805	1.2600	1.2401	$\underline{1.1644}$	1.2588	**1.1210**	0.6640	0.8242	$\underline{0.8245}$	0.8228	0.8225	**0.8257**
female3	1.5793	$\underline{1.2110}$	1.2432	1.2501	1.2546	**1.1955**	0.6734	0.7669	0.7719	$\underline{0.7723}$	0.7678	**0.7767**
female4	1.5128	1.0769	**1.0755**	1.0909	1.0819	$\underline{1.0762}$	0.6560	0.7722	0.7744	$\underline{0.7789}$	0.7728	**0.7822**
AVG4	1.4854	1.1249	1.1316	$\underline{1.1122}$	1.1363	**1.0815**	0.6670	0.8135	0.8157	$\underline{0.8172}$	0.8140	**0.8212**

V. CONCLUSION

In this paper, we propose a dictionary based speech enhancement framework, in which a semi-sparse and non-negative LSTM network has been designed to generate the combination coefficients, and the dictionaries are used to provide structural information for speech enhancement. In the experiments, we adopt four methods for comparison and the results demonstrate that our method works the best. In the future, we aim to record a large scale of dataset and extend the networks much deeply, which may result in more excellent performance, and also, we would like to design a post-processing method base on boosting, which aims to place more emphasis on the challenging samples in training stage and can improve the robustness of our method, finally, phase enhancement can be performed simultaneously with the spectral magnitudes.

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

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{fc}} &= \sum_{t=1}^{T} \frac{\partial L}{\partial s_t}\frac{\partial s_t}{\partial f_t}\frac{\partial f_t}{\partial W_{fc}} \\ &= \sum_{t=1}^{T} \delta_s^{(t)} \odot s_{t-1} \odot f_t \odot (1-f_t)(c_{t-1})^{\mathrm{T}}\end{aligned} \tag{14}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{ix}} &= \sum_{t=1}^{T} \frac{\partial L}{\partial s_t}\frac{\partial s_t}{\partial i_t}\frac{\partial i_t}{\partial W_{ix}} \\ &= \sum_{t=1}^{T} \delta_s^{(t)} \odot g_t \odot i_t \odot (1-i_t)x_t^{\mathrm{T}}\end{aligned} \tag{15}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{ic}} &= \sum_{t=1}^{T} \frac{\partial L}{\partial s_t}\frac{\partial s_t}{\partial i_t}\frac{\partial i_t}{\partial W_{ic}} \\ &= \sum_{t=1}^{T} \delta_s^{(t)} \odot g_t \odot i_t \odot (1-i_t)(c_{t-1})^{\mathrm{T}}\end{aligned} \tag{16}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{gx}} &= \sum_{t=1}^{T} \frac{\partial L}{\partial s_t}\frac{\partial s_t}{\partial g_t}\frac{\partial g_t}{\partial W_{gx}} \\ &= \sum_{t=1}^{T} \delta_s^{(t)} \odot i_t \odot (1-g_t \odot g_t)x_t^{\mathrm{T}}\end{aligned} \tag{17}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{gc}} &= \sum_{t=1}^{T} \frac{\partial L}{\partial s_t}\frac{\partial s_t}{\partial g_t}\frac{\partial g_t}{\partial W_{gc}} \\ &= \sum_{t=1}^{T} \delta_s^{(t)} \odot i_t \odot (1-g_t \odot g_t)(c_{t-1})^{\mathrm{T}}\end{aligned} \tag{18}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{ox}} &= \sum_{t=1}^{T} \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial o_t}\frac{\partial o_t}{\partial W_{ox}} \\ &= \sum_{t=1}^{T} \delta_c^{(t)} \odot \frac{\partial c_t}{\partial o_t} \odot o_t \odot (1-o_t)x_t^{\mathrm{T}} \\ \frac{\partial c_t}{\partial o_t} &= \left[\frac{\partial sh_{(D,u)}(\delta(s_t^0)\odot o_t^0)}{\partial o_t^0}, \frac{\partial(\delta(s_t^1)\odot o_t^1)}{\partial o_t^1}, \frac{\partial(\delta(s_t^2)\odot o_t^2)}{\partial o_t^2}\right] \\ &= \left[sh'(\delta(s_t^0)\odot o_t^0)\odot\delta(s_t^0), \delta(s_t^1), \delta(s_t^2)\right]\end{aligned} \tag{19}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial W_{oc}} &= \sum_{t=1}^{T} \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial o_t}\frac{\partial o_t}{\partial W_{oc}} \\ &= \sum_{t=1}^{T} \delta_c^{(t)} \odot \frac{\partial c_t}{\partial o_t} \odot o_t \odot (1-o_t)(c_{t-1})^{\mathrm{T}}\end{aligned} \tag{20}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial D} &= \sum_{t=1}^{T} \frac{\partial L}{\partial c_t^0}\frac{\partial c_t^0}{\partial D} = \sum_{t=1}^{T} \frac{\partial L}{\partial c_t^0}\left[\tanh(\Delta+u)+\tanh(\Delta-u)\right] \\ \Delta &= \delta(s_t^0)\odot o_t^0\end{aligned} \tag{21}$$

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial u} &= \sum_{t=1}^{T} \frac{\partial L}{\partial c_t^0}\frac{\partial c_t^0}{\partial u} = \\ &\sum_{t=1}^{T} \frac{\partial L}{\partial c_t^0}\left[D\left((\tanh(\Delta-u))^2-(\tanh(\Delta+u))^2\right)\right] \\ \Delta &= \delta(s_t^0)\odot o_t^0\end{aligned} \tag{22}$$

REFERENCES

[1] Loizou P C . “Speech Enhancement: Theory and Practice[M],” in *CRC Press*, Inc. 2007.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Huber, Peter J, “Robust Estimation of a Location Parameter,” *Annals of Statistics*, 53 (1): 73C101. doi:10.1214/aoms/1177703732. JSTOR 2238020.

[7] 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.

[8] 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.

[9] 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

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

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

[19] 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.

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

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

[22] 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.

[23] 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.

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

[25] 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.

[26] Styan, George P. H. , "Hadamard products and multivariate statistical analysis," *Linear Algebra & Its Applications* 6(1973):217-240.

[27] Kingma D P, Ba J. Adam, "A method for stochastic optimization[J]," *arXiv preprint* arXiv:1412.6980, 2014.

[28] Wax M, Kailath T, "Detection of signals by information theoretic criteria[J]," *IEEE Trans on Acoustics Speech & Signal Processing*, 1985, 33(2):387-392.