Improving the Spectra Recovering of Bone-Conducted Speech via Structural SIMilarity Loss Function

Changyan Zheng* Jibin Yang* Xiongwei Zhang*† Meng Sun* and Kun Yao*
* Army Engineering University, Nanjing, China
† The Corresponding Author, E-mail:xwzhang9898@163.com

***Abstract*—Bone-conducted (BC) speech is immune to background noise, but suffers from low speech quality due to the severe loss of high-frequency components. The key to BC speech enhancement is to restore the missing parts in the spectra. However, even with advanced deep neural networks (DNN), some of the recovered components still lack expected spectro-temproal structures. Mean Square Error loss function (MSE) is the typical choice for supervised DNN training, but it can only measure the distance of the spectro-temporal points and is not able to evaluate the similarity of structures. In this paper, Structural SIMilarity loss function (SSIM) originated from image quality assessment is proposed to train the spectral mapping model in BC speech enhancement, and to our best knowledge, it is the first time that SSIM is deployed in DNN-based speech signal processing tasks. Experimental results show that compared with MSE, SSIM can acquire better objective results and obtain spectra with spectro-temporal structures more similar to the target one. Some adjustments of hyper-parameters in SSIM are made due to the difference between natural image and magnitude spectrogram, and the optimal choice of them are suggested. In addition, the effects of three components in SSIM are analyzed individually, aiming to help further study on the applications of this loss function in other speech signal processing tasks.**

***Index Terms*—Bone-conducted speech enhancement, DNN, Mean Square Error, Structural SIMilarity (SSIM), loss function**

I. Introduction

Bone-conducted (BC) microphone is a kind of skin-attached non-audible sensor and converts the vibration around the skull or throat into electrical signal [1]. Though BC speech is extremely robust in adverse environments, its intelligibility is lower than conventional air-conducted (AC) speech [2]. Due to the attenuation of human body channel, it faces severe loss of high-frequency components. Besides, some phonemes like unvoiced fricatives and plosives are totally lost, which are generated in the oral or nasal cavity rather than the vocal cord.

Since BC speech is immune to the acoustic noise, it is often used to improve the speech communication quality in noisy environments. In most cases, BC speech plays an auxiliary role for improving AC speech enhancement performance [3] [4]. That is to say, AC speech is indispensable in many related algorithms. Nevertheless, it is meaningful to enhance BC speech independently, because AC speech can be completely unintelligible and become totally useless in some occasions.

The basic idea of direct BC speech enhancement is to find the mapping relationship from BC speech to its corresponding clean AC speech. Most of the methods are based on the source-filter model. Since BC microphone can pick up the vibration of the glottis clearly, the source signal is often assumed unchanged between the two speech. Therefore, the key of these approaches is to find the mapping relationships of different vocal tract filters, which are often represented by spectral envelope features like Linear Predictive (LP) family parameters or Mel-frequency cepstral coefficients (MFCC). Neural networks or Gaussian Mixture Models (GMM) are often adopted as the mapping model [5] [6] [7]. However, the quality of the enhanced speech is sensitive to the distortion of the mapped low-dimensional spectral envelope [8] and thus cannot be guaranteed robust.

Due to DNN's strong ability of exploring in a much larger hypothesis space, it is possible for us to learn the complex distribution of high-dimensional spectra now [9] [10]. In our previous work [11], we explored an effective mapping model based on Bidirectional Long Short-Term memory Recurrent Neural Networks (BLSTM-RNN) [12], which maps the log Short-Time Fourier transformation (STFT) magnitude of BC speech to that of AC speech. With BLSTM-RNN's strong ability of modeling the temporal relationship, a great improvement is achieved when compared with RBM-DNN based model. However, the relationship of adjacent frequency bins is still not emphasized well, and the distortion in some spectral structures are still obvious in the mapped spectra. Note that mean squared error (MSE) is the most-used loss function for training DNN in speech processing, since it merely measures the distance of the spectro-temporal points, we think it cannot constrain the structures of recovered spectra effectively. In fact, MSE has already aroused wide discussion for its inherent shortcomings [13], especially when dealing with perceptually important signals such as speech and images.

Motivated by this thought, we would like to adopt a kind of error metric which can also measure the difference of spectro-temporal structures. Structural SIMilarity (SSIM) [14] is one of the most popular metrics in image quality assessment. It comprehensively takes illumination, contrast and structure of the local patch into consideration, which is more appealing to human visual system than pixel-based assessment. Most

recently, it is integrated as a perceptual loss function for training DNN [15] and achieves significant improvement in image restoration tasks. Inspired by the research [16] that SSIM metric can be deployed for quantifying the similarity between noise and speech, we consider it is also possible to deploy SSIM loss function for training the spectral mapping model in BC speech enhancement and expect it to guide better recovering of the lost components in spectra. Meanwhile, due to the difference between the spectrogram of speech and the natural image, we explore the adjustments of the hyper-parameters in the SSIM loss function and analyze the influence of the three components in it on the mapped spectra individually.

The rest of this paper are organized as follows. The details of proposed method is introduced in Section II. A set of evaluation experiments to assess the performance of proposed method are provided in Section III. Finally, the conclusion is made in Section IV.

II. The Proposed Method

In this section, we first describe the problem formulation of BC speech enhancement. Then, we introduce the proposed SSIM loss function in detail.

A. Problem Formulation

Formally, let $X \in \mathbb{R}_{+}^{d\times T}$ which is composed of multiple feature frames $\{\cdots, x_{t-1}, x_t, x_{t+1}, \cdots\}$ be the BC speech magnitude spectrogram, and $Y \in \mathbb{R}_{+}^{d\times T}$ which is composed of multiple feature frames $\{\cdots, y_{t-1}, y_t, y_{t+1}, \cdots\}$ be the magnitude spectrogram of its corresponding clean AC speech, where d is the dimension of feature, that is the number of frequency bins in each frame, and T is the length of the spectrogram, t denotes the current timestep.

Given a training set of $D = \{(X_n, Y_n)\}_{n=1}^{N}$ including N parallel pairs of BC and AC speech spectrograms, the problem of BC speech enhancement can be formalized as finding a mapping relationship $f_\theta : X \rightarrow Y$ that maps the magnitude spectrograms of BC speech to that of clean AC speech, where f_θ is a DNN-based mapping model parameterized by θ. Usually, logarithmic X and Y are used to reduce the dynamic range, and global mean-variance normalization is applied to make the DNN training amenable.

Traditionally, MSE is used to measure the error between the estimated and the target spectrograms, then we find the best model parameter θ by minimizing the following loss function:

$$J_{MSE}(\theta) = \frac{1}{2N}\sum_{n=1}^{N}\|f_\theta(X_n) - Y_n\|_2^2 \tag{1}$$

In the enhancement stage, waveform reconstruction is completed by appending the original phase information of BC speech to the spectrogram reconstructed by the learnt model.

B. SSIM Loss Function

Different from MSE, SSIM comprehensively takes **illumination**, **contrast** and **structure** of the local image patch into consideration, and is proven to be more appealing to human visual system than pixel-based assessment [14]. The magnitude spectrogram of speech signal includes most of the speech information and are often used as an image format. When viewed as an image, the spectral magnitude can be seen as the luminance, and the structural information is contained in the harmonics.

Suppose two local spectrogram patches centered at the spectro-temporal point x and y, the three components in SSIM metric are computed individually as following:

$$L(x,y) = \frac{2\mu_x\mu_y + C_1}{{\mu_x}^2 + {\mu_y}^2 + C_1} \tag{2}$$

$$C(x,y) = \frac{2\delta_x\delta_y + C_2}{{\delta_x}^2 + {\delta_y}^2 + C_2} \tag{3}$$

$$S(x,y) = \frac{\delta_{xy} + C_3}{\delta_x\delta_y + C_3} \tag{4}$$

where square patch is usually chosen, μ_x, μ_y, δ_x and δ_y denote mean spectral magnitude and standard deviation of spectral magnitude in the patch, δ_{xy} is the covariance coefficient. C_1, C_2 and C_3 are constants introduced to avoid instability when the denominators in (2)-(4) are very close to zero. Specifically, they are computed as:

$$C_1 = (K_1L)^2, C_2 = (K_2L)^2, C_3 = (K_3L)^2 \tag{5}$$

where in image processing, L is the dynamic range of the pixel values (255 for 8-bit gray scale images). K_1 and K_2 are predefined small constants. The three components are then combined together to yield an overall similarity measure:

$$SSIM(x,y) = L(x,y)^\alpha C(x,y)^\beta S(x,y)^\gamma \tag{6}$$

To simplify the expression, α, β and γ are all set to 1 and $C_3 = C_2/2$, and the final formulation of SSIM is:

$$SSIM(x,y) = \frac{2\mu_x\mu_y + C1}{{\mu_x}^2 + {\mu_y}^2 + C_1} \cdot \frac{2\delta_{xy} + C_2}{{\delta_x}^2 + {\delta_y}^2 + C_2} \tag{7}$$

From (7) we can clearly note that SSIM is very different from conventional point-based error metric, because it includes the local statistics such as μ_x, δ_x and δ_{xy}.

In practice, in order to resist the undesirable "blocking" artifacts, a Gaussian filter with standard deviation δ_G is used to replace the uniform window to compute the means and standard deviations. As the filter moves point-by-point over the entire spectrogram to compute SSIM scores, a map of SSIM scores based on per-point can be formed. In image quality assessment, the mean SSIM score of the entire images is typically used to evaluate the quality. The larger the score is, the more similar the two images are.

In DNN training, our objective is to maximize the sum of SSIM scores across all spectro-temporal points in spectrograms, and the SSIM loss function can be defined as following:

$$J_{SSIM}(\theta) = -\frac{1}{N}\sum_{n=1}^{N} SSIM(f_\theta(X_n), Y_n) \tag{8}$$

Due to the difference between natural image and speech spectrogram image, some adjustments should be made for using SSIM loss function in BC speech enhancement task. **Firstly**, since SSIM metric requires the signal non-negative, we can only apply it to measure the similarity between non-negative spectra such as magnitude spectra or power spectra, so we first transform the log spectral magnitude back to raw magnitude to compute it. **Secondly**, $K_1 = 0.01$ and $K_2 = 0.03$ is set as the default value according to [14], but we change L to 7. We choose this value based on the dynamic range of magnitude spectra of AC speech. Though the maximum value of magnitude spectra is about 35, 95% values are under 7. If L is set too large, $C1$ and $C2$ will be oversized which will cause $SIMM(x, y)$ to be constant 1. **Thirdly**, It has been found that smaller values of δ_G produce better results at edges [15]. In fact, δ_G decides the size of Gaussian filter. In image processing task, $\delta_G = 1.5$ which covers 11×11 region is often used. By observing that one harmonic texture covers about 2-4 frequency bins in our spectrograms, we choose $\delta_G = 0.5$ which corresponds to 3×3 region. The result of the experiments we carried out later verifies the effectiveness of our choice.

III. Experiments

A. Datasets and Evaluation Metrics

The speech data is collected in an anechoic chamber. We select 1000 Chinese Mandarin sentences as the corpus, and each of them lasts for about 3-5 seconds. BC and AC speech are recorded synchronously at 32kHz sampling rate. Five male and five female speakers are required to read 200 different sentences selected randomly in the corpus, among which, 160 sentences are used for training and the rest are for testing. No overlap between the two sets.

Perceptual Evaluation of Speech Quality (PESQ) and Log-spectral Distance (LSD) are employed to evaluate the speech quality objectively. PESQ score ranges from -0.5 to 4.5. It measures the overall speech quality and has high correlation with subjective evaluation scores. Higher scores mean better speech quality. LSD is used to measure the spectral distortion between the referenced speech and enhanced speech. Small LSD means less speech distortion.

B. Experimental setup and Results

Currently, we consider enhancing the BC speech of 8kHz sampling rate, because in telecommunications, 8kHz is still the mainstream sampling rate. In addition, the effective spectral component of BC speech is about 2kHz, it is difficult to recover the lost components to very high bands. Thus, the acoustic signals are first down-sampled to 8kHz in our experiments, then windowed by 32ms with 8ms frame shift. 129-dimensional log spectral magnitude are extracted using a 256 point STFT. An individual enhancement model is trained for each speaker.

1) The effect of δ_G in SSIM loss: It has been found that the standard deviation δ_G can affect the quality of generated image, especially the edges of images [15]. In this experiment, we explore the effect of different δ_G values on the mapped spectra.

We conduct the experiment on the data of one male and one female speakers, and deploy BLSTM-RNN as our feature mapping model which is composed of 3 hidden layers of 512 memory cells and a fully connected linear layer. SSIM loss function is used to BLSTM-RNN and we denote this method as **BLSTM-SSIM**. Adam optimizer [17] is used to train the model with an initial learning rate of 0.002. The results are shown as Table I.

TABLE I: Objective Results of BLSTM-SSIM with different δ_G values

		male		female	
Sigma	Filter Size	PESQ	LSD	PESQ	LSD
0.01	1×1	3.039	0.960	3.382	0.869
0.50	3×3	**3.056**	**0.949**	**3.399**	**0.856**
1.00	7×7	3.042	0.965	3.377	0.872
1.50	11×11	3.035	0.970	3.369	0.875
2.50	15×15	3.024	0.981	3.355	0.883
5.00	29×29	2.982	1.002	3.323	0.913

In Table I, it can be noted that when $\delta_G = 0.5$, the best PESQ and LSD scores can be acquired for both speakers. We infer it that one harmonic texture covers about 2 to 4 frequency bins in the spectrogram image, so if the size of filter is chosen around 2 to 4, the mapped spectra tends to preserve better structural details. Therefore, we can conclude that the optimal δ_G value should be chosen according to the characteristics of the harmonic, and if the resolution of the spectra which depends on the factors such as STFT points and sampling rate of speech is changed, the δ_G value should be reconsidered.

Fig.1 shows the spectrograms of utterance enhanced by BLSTM-SSIM with different δ_G values. From the oval boxes in Fig.1, we can see that the edge of harmonic is kept more complete and clear at $\delta_G = 0.5$ than at other δ_G values.

2) SSIM loss vs MSE loss: In this experiment, SSIM loss function and MSE loss function are compared.

The data of five male and five female speakers are conducted in this experiment. The hyper-parameter δ_G is set to 0.5 in SSIM loss function according to the result of the previous experiment. Four methods are implemented for comparisons, which are **DNN-MSE**, **DNN-SSIM**, **BLSTM-MSE**, **BLSTM-SSIM** respectively. DNN and BLSTM denote the RBM-DNN and BLSTM-RNN neural networks, MSE and SSIM denote the MSE and SSIM loss functions to train the neural network.

The architecture of BLSTM-RNN is introduced in the previous experiment. In DNN-based approaches, the input of the feature is expanded to 11 frames (5 to the left, 5 to the right), so that the frame-wised contextual information is used to estimate the middle frame feature. The architecture of DNN

TABLE II: Objective Evaluation Results of Different Methods

Person	PESQ					LSD				
	BC	DNN -MSE	DNN -SSIM	BLSTM -MSE	BLSTM -SSIM	BC	DNN -MSE	DNN -SSIM	BLSTM -MSE	BLSTM -SSIM
male1	2.277	2.719	2.802	2.913	**3.056**	1.482	1.047	1.017	0.961	**0.949**
male2	1.963	2.257	2.322	2.739	**2.877**	1.480	0.991	0.963	0.899	**0.877**
male3	1.931	2.324	2.417	2.580	**2.726**	1.455	1.061	1.035	0.961	**0.936**
male4	2.281	2.762	2.861	3.085	**3.203**	1.172	0.981	0.952	0.857	**0.843**
male5	2.102	2.403	2.489	2.661	**2.840**	1.440	1.014	0.984	0.955	**0.930**
female1	2.508	3.139	3.181	3.322	**3.399**	1.369	0.912	0.889	0.865	**0.856**
female2	2.023	2.533	2.627	2.832	**2.994**	1.305	0.962	0.934	0.924	**0.904**
female3	2.078	2.469	2.561	2.716	**2.811**	1.427	1.133	1.107	1.070	**0.992**
female4	2.294	2.611	2.708	2.849	**2.941**	1.239	0.978	0.960	0.953	**0.927**
female5	1.847	2.214	2.312	2.394	**2.495**	1.389	1.047	1.002	0.995	**0.966**
Average	2.130	2.543	2.628	2.809	**2.934**	1.376	1.013	0.984	0.944	**0.918**

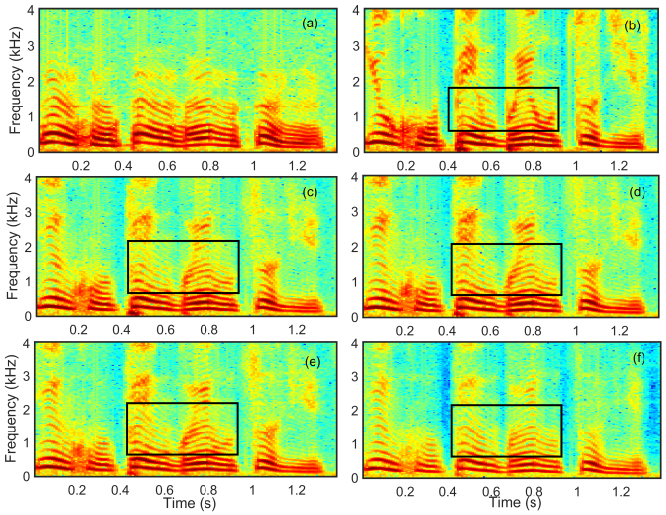

Fig. 1: Spectrograms of an utterance enhanced by BLSTM-SSIM with different δ_G values (a) BC speech (b) AC speech (c) Speech enhanced by BLSTM-SSIM with $\delta_G = 0.01$ (d) Speech enhanced by BLSTM-SSIM with $\delta_G = 0.5$ (e) Speech enhanced by BLSTM-SSIM with $\delta_G = 1.5$ (f) Speech enhanced by BLSTM-SSIM with $\delta_G = 2.5$

is set to 3 hidden layers of 1024 hidden units. The exponential linear function (ELU) [18] and linear function are chosen as the activation function for hidden layers and output layer respectively. SGD algorithm is used to train DNN [19] with initial learning rate of 0.001.

The objective results of different methods are shown in Table II, which are computed between the enhanced speech and referenced AC speech. "BC" represents the results of the original BC speech.

From Table II, we can see that the average PESQ score of BC speech is around 2.1, indicating that the quality of BC speech is not satisfying. From the PESQ and LSD scores with underlying lines, it can be noticed that whether the methods are based on DNN or BLSTM-RNN, using SSIM loss function to train neural network can achieve better results than MSE loss function. Therefore, we can conclude that SSIM loss function can performs better than MSE loss function in recovering the spectra of BC speech. We can also see that BLSTM-SSIM can greatly improve the PESQ scores of original BC speech by approximately 0.80 and reduce the average LSD score by approximately 0.46, which is equivalent to 37% and 33% improvement.

The spectrograms of an utterance enhanced by different methods are shown in Fig.2, which are best viewed by zooming in on the electronic copy.

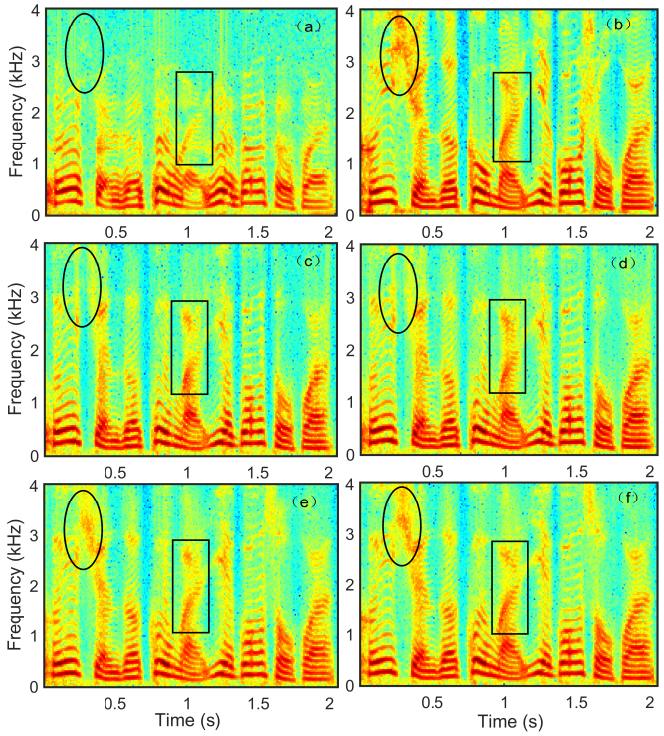

Fig. 2: Spectrograms of an utterance enhanced by different method (a) BC speech (b) AC speech (c) Speech enhanced by DNN-MSE (d) Speech enhanced by DNN-SSIM (e) Speech enhanced by BLSTM-MSE (f) Speech enhanced by BLSTM-SSIM

Comparing Fig.2 (a) with Fig.2 (b), we can see that the components above 2kHz and some plosives almost disappear completely in BC speech. From the rectangle boxes, it can be noted that SSIM loss function obtains better mapped spectra with clearer harmonic structure than MSE loss function on

both the DNN-based and BLSTM-based methods. BLSTM-RNN is able to recover the plosive components, which can be seen from the oval boxes.

3) Ablation study of the three components in SSIM: SSIM is composed of three kind of similarity measurement, including luminance, contrast and structure. In this experiment, we try to figure out how the three components affect the mapped spectra.

We take the $L(x, y)$, $C(x, y)$ and $S(x, y)$ defined in (2), (3) and (4) as the similarity measurement respectively, and the loss function is defined as following:

$$\mathcal{L}(X, Y) = -\frac{1}{M}\sum_{j=1}^{M} F(x_i, y_j), F = L, C, S \qquad (9)$$

We denote the corresponding loss function as **L-SSIM**, **C-SSIM** and **S-SSIM**. BLSTM-RNN is trained with the three loss functions respectively and the results are shown in Fig.3.

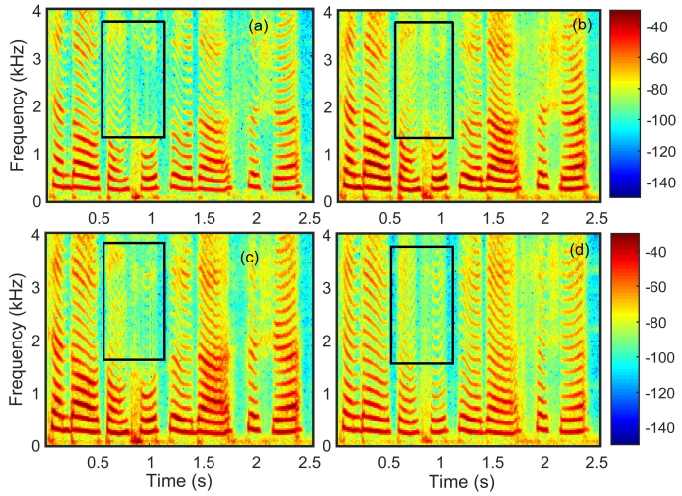

Fig. 3: Spectrograms of an utterance enhanced by different components in SSIM loss function (a) AC speech (b) Speech enhanced by L-SSIM (c) Speech enhanced by C-SSIM (d) Speech enhanced by S-SSIM

When compared with Fig.3 (b) with Fig.3 (c), we can note that L-SSIM and C-SSIM acquire very similar spectra, but the C-SSIM seems to get clearer background. We infer that the contrast which is used to distinguish the foreground and background in images has better ability to suppress the background noise than the illumination punishment. From Fig.3 (d) we can see that S-SSIM tends to obtain clear and complete harmonic structures. However, some unexpected structures are also generated and the energy of structures are very similar between the low-frequency band and high-frequency band. Since unexpected harmonics make speech muffled, we think S-SSIM needs to cooperate with energy constraint to take its advantages.

Generally, the three components in SSIM have their own advantages and disadvantages. Therefore, in order to get better results, it may be reasonable to adjust the portation of the three components by changing the values of α, β and γ in (6) according to different speech tasks and speech characteristics.

IV. Conclusion

In this paper, SSIM loss function is used to train the spectral mapping model in BC speech enhancement. The experimental results show that SSIM loss function can acquire better quality of enhanced speech than conventional MSE loss function. Besides, some hyper-parameters in SSIM loss function are adjusted due to the difference between natural images and magnitude spectrogram, and the optimal choice of them is suggested. Meanwhile, the effects of three components in SSIM loss function on the mapped spectra are analyzed individually. In the future, we would like to further modify SSIM loss function to suit the characteristics of speech better and explore the generality of it in other tasks such as speech de-noising and voice conversion.

Acknowledgment

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

References

[1] H. 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*. VDE, 2012, pp. 1–4.

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

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

[4] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, "Multi-sensory microphones for robust speech detection, enhancement, and recognition," 2004.

[5] A. Shahina and B. Yegnanarayana, "Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 087219, 2007.

[6] K. Vijayan and K. S. R. Murty, "Comparative study of spectral mapping techniques for enhancement of throat microphone speech," in *Communications (NCC), 2014 Twentieth National Conference on*. IEEE, 2014, pp. 1–5.

[7] M. T. Turan and E. Erzin, "Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7049–7053.

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

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

[10] Y. Xu, J. Du, and L.-R. Dai, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[11] C. Zheng, X. Zhang, M. Sun, J. Yang, and Y. Xing, "A novel throat microphone speech enhancement framework based on deep blstm recurrent neural networks," 2018.

[12] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[13] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.
[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[15] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.
[16] L. Marchegiani, X. Fafoutis, and S. Abbaspour, "Speech identification and comprehension in the urban soundscape," *Environments*, vol. 5, no. 5, p. 56, 2018.
[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[18] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.
[19] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.