A wearable bone-conducted speech enhancement system for strong background noises

Boyan Huang*, Yihan Gong, Jinwei Sun
Department of Automatic Testing and Control,
Harbin Institute of Technology,
Harbin 150001, China
*Email: byhuang@hit.edu.cn

Yi Shen
Department of Control Science and Engineering,
Harbin Institute of Technology,
Harbin 150001 China

***Abstract*—Wearable electronic systems have been and will continue to be utilized in both civil and military uses. Strong noise environments derived from large vehicles (e.g. ships, aircrafts, or military tanks) seriously affect the quality of speech communications especially for wearable systems without hermetic packaging. Bone conduction technology through the acquisition of skull vibration is able to obtain voice information, which can effectively avoid the interference of acoustic noise on the speech. In this paper, a high-performance wearable bone-conducted speech enhancement system is developed to reduce the distortion of the environment noises. Both bone-conducted and air-conducted voices are used to train the equalization function of bone-conducted speech to air-conducted speech based on the deep neural network, and spectrum coefficients of linear predictive coding is taken as feature information for conversion model.**

Keywords—Bone-conducted speech; speech enhancement; deep neural network;

I. Introduction

Wearable electronic systems has been and will continue to be utilized in both civil and military uses. For wearable devices without hermetic sealing or packaging, strong noise environments derived from large vehicles (e.g. ships, aircrafts, or military tanks) seriously affect the quality of speech communications. To alleviate the noise effects, air-conducted (AC) speech is easily disturbed by all kinds of noise in the process of communication, while it is difficult to isolate a speech signal from the background noise [1]. Although, there are a vast number of techniques developed for AC speech enhancement, all of them may lose their capabilities when the background noise is very harsh [2].

The bone-conducted (BC) speech signal collects the vibration of the skull through the highly sensitive vibration sensor and then converting into an audio signal, instead of collecting audio source not from the sound transducer [3]. As thus, BC speech is only relevant to human body, not to the ambient background noise [4]. Useful signals can also be clearly transmitted even under intensively noise environments [5]. However, BC speech has severe transmission attenuation of the high-frequency components, and the nonlinearity of the transmission path from the vibration pickup to the microphone will affect the system performance in a significant way [6]. For the above reasons, an effective nonlinear method should be applied to train and recover the high-frequency part of the speech.

In recent years, with the development of Deep Learning (DL), some improved neural network-based spectrum extension methods have been developed [7]. Deep neural networks (DNN), convolution neural networks, and deep confidence networks have achieved good results in terms of speech enhancement; they have also made striking developments in image classification. Deep learning is mainly applied to spectral feature conversion to achieve voice conversion and speech recognition.

In this paper, a successfully trained DNN model are applied in the BC speech conversion and high-frequency components recovery. A wearable system is developed for collecting skull vibrations from different positions such as the human nose wing, forehead, ear bones, throat, lips and cheeks etc. NI multi-channel data acquisition card is used to collect and preserve the simultaneous voice of the BC speech measured from different positions. In order to prevent taking a very longtime to gather a large number of data, buffer overflowing and losing data, acquisition and preservation program in the paper has taken a producer-consumer model, namely while collecting and preserving.

II. Spectrum Expansion of Bone-conducted Speech

The schematic diagram of bone-conducted speech spectrum expansion based on deep neural networks (DNN) is shown in Fig.2. The DNN training process consists of an unsupervised pre-training section based on the deep belief network (DBN) and a supervised fine-tuning section based on BP neural network. Tansig function is selected as the activation function for experiment, and the activation function of the output layer is a linear function. In order to avoid falling into the local minimum in the training of DNN, firstly DBN model for a bone-conducted speech to an air-conducted speech conversion is pre-trained. DBN model is a multi-layer probability generation model, and using the restricted Boltzmann machine (RBM) to achieve layer-by-layer training [7].

The vector dimension of the line spectral pairs (LSP) parameter corresponds to the number of explicit nodes V , and the hidden layer node H is made to optimize the model. The energy functions of RBM depend on the assumptions to which the nodes are subjected. As speech signals generally follow the

978-1-5386-2972-7/17/$31.00 ©2017 IEEE

Gaussian distribution, the explicit layer is assumed to be Gaussian distribution, while the hidden layer still retains the assumed Bernoulli distribution, which is the Gaussian-Bernoulli RBM model.

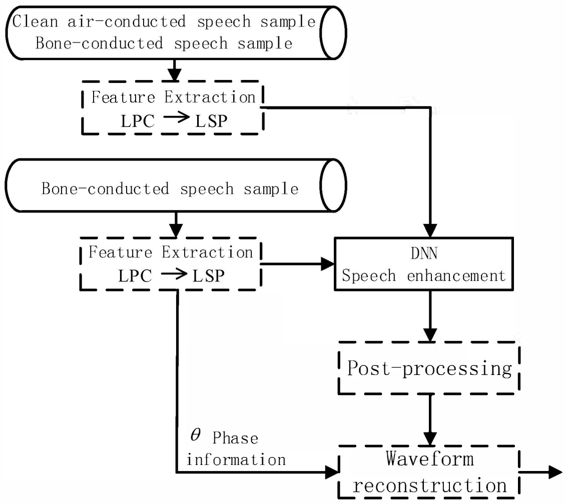

Fig. 1. Conversion model of bone-conducted speech based on DNN

12-dimensional LSP coefficients are selected as the feature parameters of the input bone-conducted speech. In order to keep each explicit node a unit variance, the parameters are Gaussian. As the nodes of explicit and hidden layers of RBM are conditional independent, the conditional probability distribution of GBRBM is as follows:

$$p\left(v_i=1\middle|h\right)=N\left(\sigma_i\sum_j h_jW_{ij}+a_i,\sigma_i^2\right) \qquad (1)$$

$$p\left(h_i=1\middle|v\right)=f\left(\sum_j \frac{v_i}{\sigma_i}W_{ij}+b_i\right) \qquad (2)$$

where $f(x)=\frac{1-\exp(-x)}{1+\exp(-x)}$ is activation function, $N(\ )$denotes the Gaussian distribution function.

The back propagation algorithm based on the MMSE objective function is used to train the DNN. In the fine-tuning section, all parameters are monitored and trained, then compared to the pre-trained part of the initialized parameters for the first few hidden layers. The following error equation is improved by small random gradient descent algorithm.

$$E_r=\frac{1}{N}\sum_{n=1}^{N}\left\|\hat{A}_n(Y_{n-\tau}^{n+\tau},W,b)-A_n\right\|_2^2 \qquad (3)$$

where E_r is the mean square error, A_n and $\hat{A}_n(Y_{n-\tau}^{n+\tau},W,b)$indicate linear predication cepstrum coefficient and its estimation respectively, N is the value of minimum batch, $Y_{n-\tau}^{n+\tau}$ is logarithmic spectral feature vector. The weight and bias parameter $(W,b)$are updated by learning factor λ, and calculated by the following equation,

$$\Delta(W_{n+1}^l,b_{n+1}^l)=-\lambda\frac{\partial E_r}{\partial(W_n^l,b_n^l)}-\kappa\lambda(W_n^l,b_n^l)+\omega\Delta(W_n^l,b_n^l),1\le l\le L+1 \qquad (4)$$

where L is the total number of hidden layers, κ is the attenuation coefficient, ω is the momentum term.

In the learning process, DNN is often used in the study of the mapping function; the relationship between the bone-conducted speech and the air-conducted speech does not make any assumptions. It will automatically learn this complex relationship for sufficient training sample conditions to make the conversion

III. Experiments

Speech signals are collected from 4 men and 4 women in a quiet room, each one records 9 fragments Chinese speech 3 minutes in total for DNN training, and 9 fragments Chinese speech 10 seconds in total for DNN testing. The air-conducted speech is assumed to be clear voice in the quiet room. The bone-conducted and air-conducted speech are recorded synchronously, and the experimental frame length is set to 25ms, the frame shift is set to 10ms, the window function is Hanning window. A 256-point Fourier transform is used for time-frequency transformation. The parameters of the 12-order LSP are extracted and the mean square error of the trained model is 0.0047. Characteristic parameters of extended frequency speech can be transformed from the tested speech through the trained model, after that, the LSP is converted to linear predictive coding (LCP). The proposed speech can be synthesized by the prediction coefficients of the test speech and LCP parameters of the extended frequency speech. The DNN takes 6 hidden layers and the 30 nodes, the experimental results are as follows:

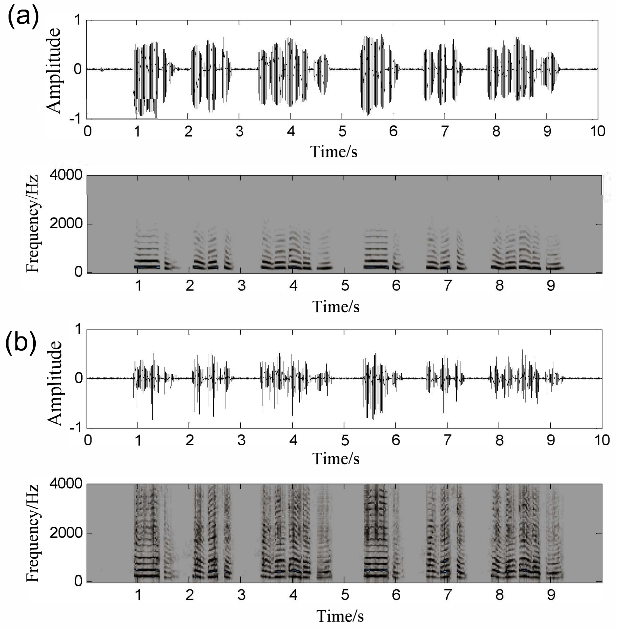

Fig. 2. Bone-conducted speech of throat, (a) original and (b) improved results.

It can be seen from the above experimental results as follows.

- By the training of DNN model, part of the high frequency information has been recovered, and the intelligibility is also greatly improved.

TABLE I. SNR AND SNRSEG(DB) OF BONE-CONDUCTED SPEECH

Speech type	Nose wing	cheek	forehead	Ear bone	throat
Original BC speech	-5.15/-4.02	-5.07/-3.94	0.75/-1.86	-3.51/-2.76	5.78/-4.54
Expanded BC speech	-4.84/-3.79	-3.75/-2.86	-0.64/-4.15	-1.06/-2.08	-1.00/-1.32

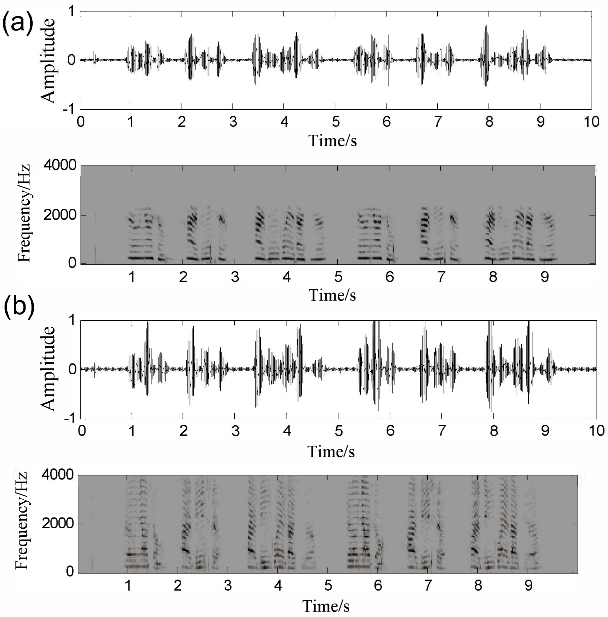

Fig.3. Bone-conducted speech of forehead, (a) original and (b) improved results.

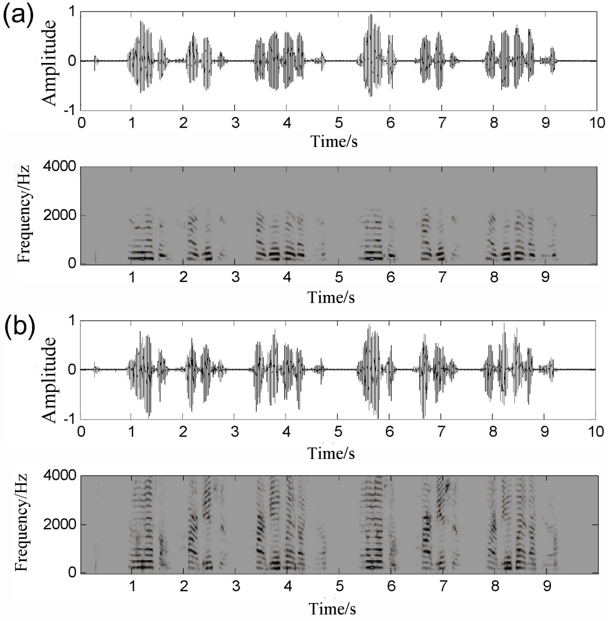

Fig.4. Bone-conducted speech of Nose wing, (a) original and (b) improved results.

- By comparing the speech from different points, it is found that the nose wing is the best position, followed by the ear bone, cheek, throat and forehead.
- The energy and distortion of the speech feature information detected from the above positions are different. The value of the information obtained from each measuring points has corresponding advantages and disadvantages in reflecting the characteristics of the speech signal.

IV. CONCLUSIONS

In this paper, a wearable bone-conducted speech enhancement system was developed, vibrations from different positions of the skull were collected and compared. AC speech was synchronously collected to compensate the high-frequency components of BC signal by spectrum expansion based on DNN. Numerous experimental results showed that part of the high frequency information has been recovered, and the quality and intelligibility of the speech were much improved.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61601140).

REFERENCES

[1] S. Gannot, D. Burshtein, and E. Weinstein, “Iterative and sequential Kalman filter-based speech enhancement algorithms,” IEEE Trans.Speech, Audio Process., vol. 6, no. 4, pp. 373-385, Jul. 1998.

[2] X. Zou, P. Jancovic, J. Liu, and M. Kokuer, “Speech signal enhancement based on MAP algorithm in the ICA space,” IEEE Trans. Signal Process.,vol. 56, no. 5, pp. 1812-1820, May 2008.

[3] N. Yousefian and P. C. Loizou, “A dual-microphone speech enhancement algorithm based on the coherence function,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 2, pp. 599-609, Feb. 2012.

[4] S. Suge, D. Koizumi, and M. Fukumi, “Speaker verification method using bone-conduction and air-conduction speech,” Intelligent Signal Processing and Communication Systems, pp. 449-552, 2009.

[5] H. S. Shin, H. G. Kang, and T. Fingscheidt. “Survey of speech enhancement supported by a bone conduction microphone,” Speech Communication, pp. 1-4, 2012

[6] T. Shimamura and T. Tamiya, “Learning for bone-conducted speech via adaptive and neural filters,” Proc. Intl. Conf. Signals and Electronic Systems, Sep. 2006.

[7] Y. Xu, J. Du,and L. Dai. A Regression Approach to Speech Enhancement Based on Deep Neural Networks.IEEE Transactions on Audio,Speech, and Language Processing. 2015, 23(1):7-20.