**2006 IEEE International
Symposium on Signal Processing
and Information Technology**

Improving Bone-Conducted Speech Quality via Neural Network

Tetsuya Shimamura, Jun'ichiro Mamiya and Toshiki Tamiya

Department of Information and Computer Sciences
Saitama University
255 Shimo-Okubo, Sakura-Ku, Saitama 338-8570, Japan
Phone : +81-48-858-3496
Fax : +81-48-858-3716
Email : shima@sie.ics.saitama-u.ac.jp

Abstract

The quality of bone-conducted speech is low, but bone-conducted speech itself is not affected by noise. In this paper, we take into account such properties of bone-conducted speech, and derive a reconstruction filter to improve the quality of bone-conducted speech. The reconstruction filter is designed by learning a neural network on the basis of the bone-conducted speech and normal speech obtained from a speaker. Experimental results show that the reconstructed speech signal has better quality than the bone-conducted speech signal.
Keywords − bone conduction, speech quality, reconstruction filter, neural network

1. INTRODUCTION

The transmission of voice on bones is called bone conduction. When the voice waveforms are transmitted from the voice source (vocal cord) through the vocal tract wall and skull, they do not confront directly with noise. This is the reason why the bone-conducted speech signal is utilized to accomplish speech communications in a very noisy environment [1]. Recently, in [2], it was reported that bone-conducted speech could be effectively used for speech recognition even in a negative decibel signal-to-noise ratio environment. However, it is known that the quality of bone-conducted speech is comparatively lower than that of normal speech being transmitted through air. This may be caused by the fact that the frequency components more than 1kHz deteriorate in bone-conducted speech [3][4]. A straightforward method to improve the quality of bone-conducted speech is to emphasize the high frequency components. However, this has been not accepted in current communication systems. One of the reasons of this fact may be that the phenomenon of bone conduction is speaker dependent. As a speaker-dependent technique, the use of an air- and bone-conduction integrated microphone is suggested in [6]. In [7], it is reported that the quality of bone-conducted speech can be improved by utilizing both long-term spectra of the normal and bone-conducted speech signals.

In this paper, we set out to design a reconstruction filter for the speaker and improve the quality of the bone-conducted speech signal obtained from the speaker by means of filtering. The filter design method is derived from the concept of neural networks [5]. Considering the speaker dependency of bone-conducted speech, both the normal and bone-conducted speech signals are utilized to design the reconstruction filter.

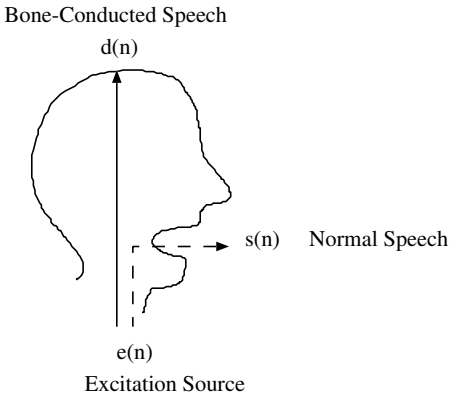

Figure 1: *Speech production and bone conduction*

The organization of this paper is as follows. Section 2 describes the principle of reconstruction filtering. Section 3 presents the proposed method. In Section 4, we discuss the results of experiments. Finally, we conclude this paper in Section 5.

2. PRINCIPLE OF RECONSTRUCTION

Figure 1 shows that a bone-conducted speech signal, $d(n)$, is measured at the head or ear of a speaker. When $d(n)$ is obtained, the corresponding normal speech signal, $s(n)$, is also obtained. Both $d(n)$ and $s(n)$ are assumed to be excited by a common excitation source, $e(n)$.

Figure 2 shows a system representation of Fig.1. The speech signal $s(n)$ should be modeled as the output of a vocal tract filter $V(z)$ (including the effect of lip radiation). The bone-conducted speech signal $d(n)$ may be obtained through a filter $B(z)$ which has the characteristics of bone-conduction.

If the above model is valid, $s(n)$ will be obtained from $d(n)$ as shown in Fig.3, where the combined system is denoted as $H(z)$. The purpose of this paper is to reconstruct the normal speech signal from the bone-conducted speech signal. Thus, the system shown in Fig.3 is a direct solution to this purpose. However, it is basically impossible to obtain the transfer function of the reconstruction filter, $H(z)$. Hence, we estimate $H(z)$ as $\hat{H}(z)$ and use it to reconstruct the original speech signal as shown in Fig.4.

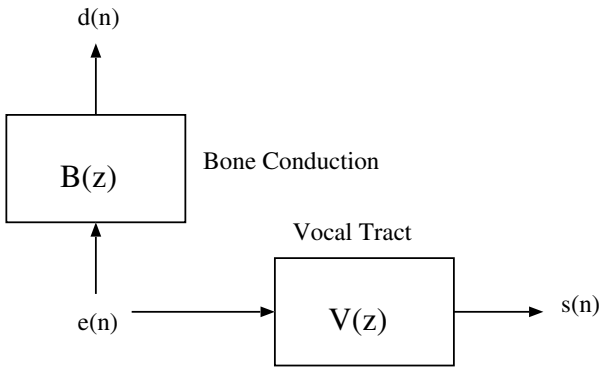

Figure 2: *System representation of Fig.1*

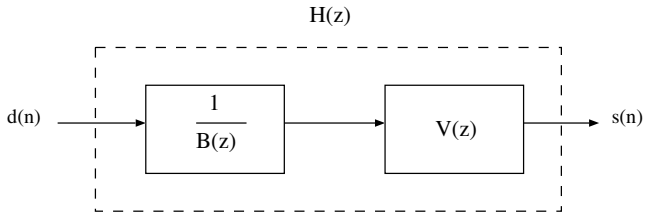

Figure 3: *A model that transforms bone-conducted speech into normal speech*

3. FILTER DESIGN BY LEARNING

Based on the principle described in Section 2, we propose the following procedure to design the reconstruction filter.

1. Obtain the bone-conducted and normal speech signals, $d(n)$ and $s(n)$, for a speaker.
2. Make a multi-layer perceptron (MLP) filter [5] learn by using $d(n)$ and $s(n)$ as shown in Fig.5.
3. Continue Step 2 until $|e(n)|$ for each n results in a specified small real value or until the iteration number of learning for each n is beyond a specified maximum number.

As the structure of the MLP filter, we use three nodes for input layer, four nodes for hidden layer and single node for output layer. The bone-conducted speech signal $d(n)$ and its delayed versions, $d(n-1)$ and $d(n-2)$, are used for the input layer. All nodes are connected with a weight coefficient for each connection, without connections for the same layer. All weight coefficients are learned by the back-propagation algorithm [5]. The nonlinearity used in the back-propagation algorithm is a sigmoidal activation function given by

$$f(x) = \frac{1 - e^{-x}}{1 + e^{-x}} \tag{1}$$

where x is a variable.

For a preliminary experiment, we observed that only the use of the reconstruction filter designed by the MLP filter produces noise as by-products, which is one enhanced by the reconstruction filter. From this observation, we use a low-pass filter with a cut-off frequency of 3.5 kHz in addition to the reconstruction filter. The low-pass filter is cascaded with the reconstruction filter as a post-filter.

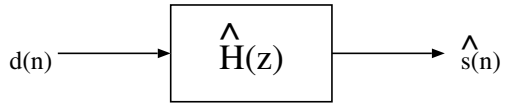

Figure 4: *Reconstruction filter*

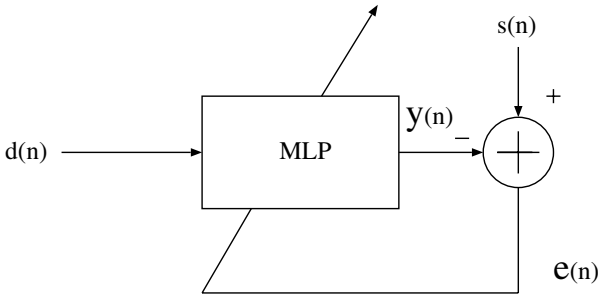

Figure 5: *MLP filter*

4. EXPERIMENTS

4.1. Selection of Bone-Conducted Speech Microphone

At first, we compared two bone-conducted speech microphones; one is Temco Japan HG-17 which is a headgear type and the other is Temco Japan EM-7B-05 which is an earphone type. In a sound-isolated room, we measured simultaneously the normal speech and bone-conducted speech pronounced by a person. To get the normal speech data, we used Panasonic RP-VK25 which is a standard microphone. On the other hand, to get the bone-conducted speech data, we used Temco Japan HG-17 or EM-7B-05. The sampling frequency is commonly 11.025kHz.

Figures 6 and 7 show a comparison of spectra on vowels /i/ and /u/, respectively. In Figs. 6 and 7, "ear VBC" and "head VBC" mean bone-conducted speech signals obtained by EM-7B-05 (earphone type) and HG-17(headgear type), respectively. It is commonly observed in Figs. 6 and 7 that bone-conducted speech levels are lower than normal speech counterparts in the region of frequencies more than 1.5kHz. This may be one of the reasons why the quality of bone-conducted speech is degraded. However, the bone-conducted speech obtained by EM-7B-05 has a more similar spectral envelope to that of the normal speech. In particular, the first formant is emphasized more clearly in the bone-conducted speech obtained by EM-7B-05. We also observed that for vowels /a/ /e/ /o/, the formant characteristics both the bone-conducted and normal speech signals have are not distinguished clearly. From these observations, we selected the use of EM-7B-05 as the bone-conducted speech microphone for the following experiment.

4.2. Listening Test

In order to investigate the performance of the proposed method, we conducted a listening test.

In a sound-isolated room, we measured again simultaneously the normal speech and bone-conducted speech pronounced by a person. We gathered two male (Male 1 and Male 2) and two female (Female 1 and Female 2) speech data. The sampling frequency is commonly 11.025kHz.

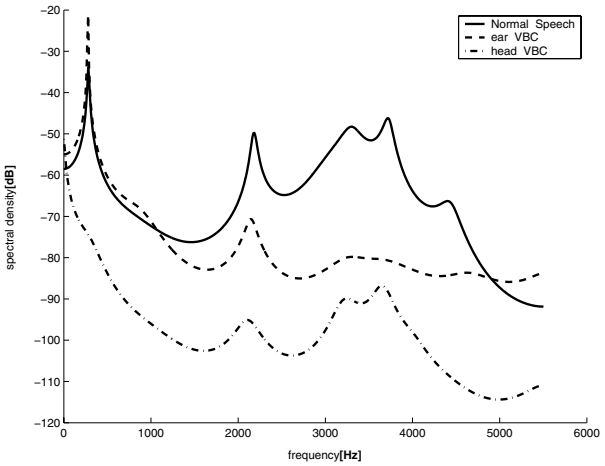

Figure 6: *Spectra on vowel /i/*

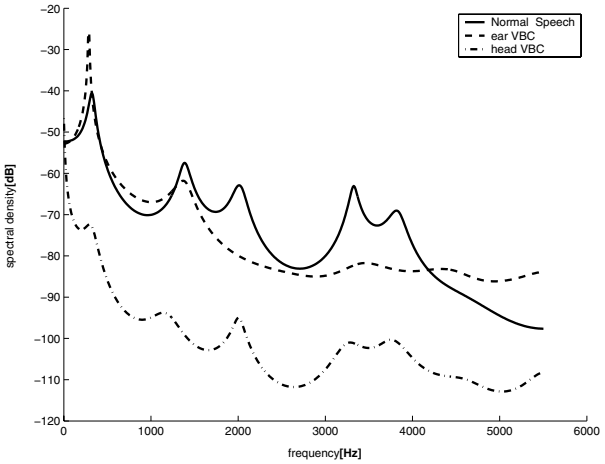

Figure 7: *Spectra on vowel /u/*

To get the normal speech data, we used again Panasonic RP-VK25 which is a standard microphone. On the other hand, to get the bone-conducted speech data, we used Temco Japan EM-7B-05.

The recorded speech signal consists of two sentences which have balanced phonemes as shown below.

1. suki na tabe mono wa nan desu ka ?
2. anata no teda suke ga hitsu you desu.

The assessment was made by twenty listeners. The pair comparison method was used for the listening test where the reconstructed speech signals were judged. In the pair comparison method, one-pair was picked among the original bone-conducted speech signal and reconstructed speech signals, and listeners judged which speech signal has higher intelligibility. The rate of selection was averaged for the twenty listeners and final score was obtained where the averaged selection rate was changed into Thurstone's Scale. We did not inform the listeners what kinds of speech signals were used a priori.

Tables 1 to 4 is the selection rate averaged for male speech. "A" is the original bone-conducted speech. "B" is the bone-conducted speech processed by MLP filter. "C" is the bone-conducted speech processed by MLP filter and low-pass filter. Tables 5 to 8 give the selection rate averaged for female speech. Figures 8-11 show the listening scores for each case in Thurstone's Scale where the quality of speech is better as the value becomes larger. From Tables 1 to 8 and Figures 8-11, it is clearly observed that quality of the reconstructed speech signal is better than that of the bone-conducted speech signal, while the quality of the reconstructed speech signal is dependent on the speakers and sentences used. This suggests that from the bone-conducted speech signal, the proposed method produces a speech signal which is more similar with the original normal speech signal.

Figures 12-14 show the waveforms of normal speech, bone-conducted speech and bone-conducted speech processed by MLP filter for Male 2, Sentence 2. From these figures, we can confirm that the bone-conducted speech signal is improved by the proposed method, resulting in a similar waveform to that of the original normal speech.

Table 1: Selection rate for Male 1, Sentence 1

x\y	A	B	C	Average
A	$\cdots$	.10	.00	.05
B	.90	$\cdots$	.05	.475
C	1.00	.95	$\cdots$	.975

Table 2: Selection rate for Male 1, Sentence 2

x\y	A	B	C	Average
A	$\cdots$	.30	.10	.20
B	.70	$\cdots$	.00	.35
C	.90	1.00	$\cdots$	.95

Table 3: Selection rate for Male 2, Sentence 1

x\y	A	B	C	Average
A	$\cdots$	.00	.00	.00
B	1.00	$\cdots$	.00	.50
C	1.00	1.00	$\cdots$	1.00

Table 4: Selection rate for Male 2, Sentence 2

x\y	A	B	C	Average
A	$\cdots$	.15	.05	.10
B	.85	$\cdots$	.30	.575
C	.95	.70	$\cdots$	.825

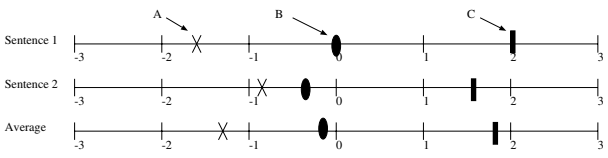

Figure 8: *Listening scores for Male 1*

Table 5: *Selection rate for Female 1, Sentence 1*

x\y	A	B	C	Average
A	· · ·	.25	.05	.15
B	.75	· · ·	.10	.425
C	.95	.90	· · ·	.925

Table 6: *Selection rate for Female 1, Sentence 2*

x\y	A	B	C	Average
A	· · ·	.20	.30	.25
B	.80	· · ·	.45	.625
C	.70	.55	· · ·	.625

Table 7: *Selection rate for Female 2, Sentence 1*

x\y	A	B	C	Average
A	· · ·	.00	.00	.00
B	1.00	· · ·	.40	.70
C	1.00	.60	· · ·	.80

Table 8: *Selection rate for Female 2, Sentence 2*

x\y	A	B	C	Average
A	· · ·	.00	.00	.00
B	1.00	· · ·	.45	.725
C	1.00	.55	· · ·	.775

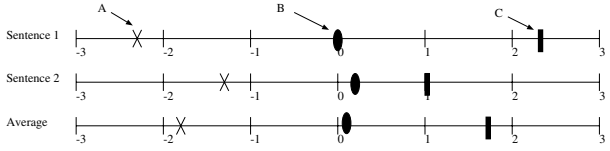

Figure 9: *Listening scores for Male 2*

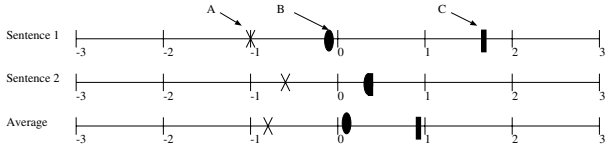

Figure 10: *Listening scores for Female 1*

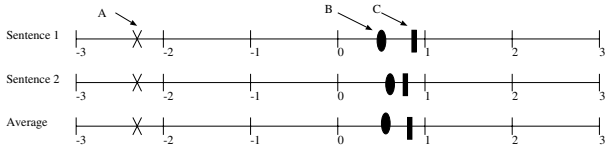

Figure 11: *Listening scores for Female 2*

5. CONCLUDING REMARKS

In this paper, it has been shown that the neural network could be effectively used in designing the reconstruction filter of bone-conducted speech. In experiments, it has been observed that the quality of bone-conducted speech is improved by utilizing an MLP filter. Future work will be focused on learning the MLP filter with all phonemes independently of the speaker.

6. REFERENCES

[1] H.M.Moser and H.J.Oyer, "Relative intensities of sounds at various anatomical locations of the head and neck during phonation of the vowels", Journal of Acoustical Society of America, vol.30, no.4, pp.275-277, 1958.

[2] S.Ishimitsu, H.Kitakaze, Y.Tsuchibushi, H.Yanagawa and M.Fukushima, "A noise-robust speech recognition system making use of body-conducted signals", Acoustical Science and Technology, vol.25, no.2, pp.166-169, 2004.

[3] M.Kumashita, T.Shimamura and J.Suzuki, "Property of voice recorded by bone-conduction microphone", Proc. Spring Meeting of the Acoustical Society of Japan, 2-Q-3, pp.269-270, 1994.

[4] T.Shimamura, "Noise reduction techniques for speech signals (Part 2)", Journal of Signal Processing, vol.9, no.3, pp.183-188, 2005.

[5] S.Haykin, "Neural Networks : A Comprehensive Foundation", Macmillan, 1994.

[6] Z.Liu, Z.Zhang, A.Acero, J.Droppo and X.D.Huang, "Direct filtering for air- and bone-conductive microphones", Proc. IEEE International Workshop on Multimedia Signal Processing, 2004.

[7] T.Shimamura and T.Tomikura, "Quality improvement of bone-conducted speech", Proc. European Conference on Circuit Theory and Design, 2005.

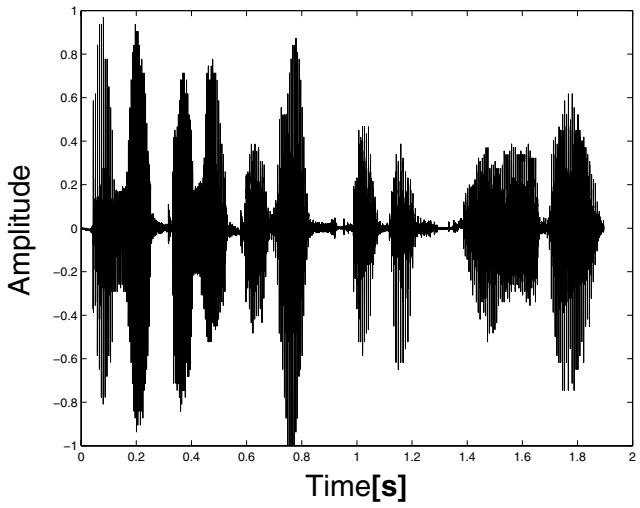

Figure 12: *Waveform of normal speech (Male 1, Sentence 2)*

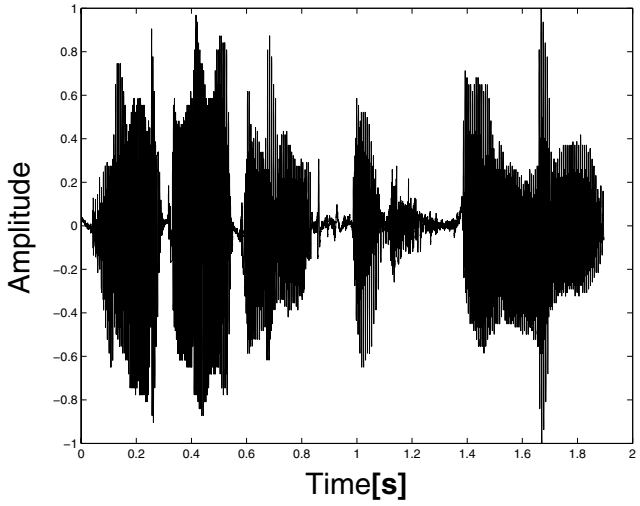

Figure 13: *Waveform of bone-conducted speech (Male 1, Sentence 2)*

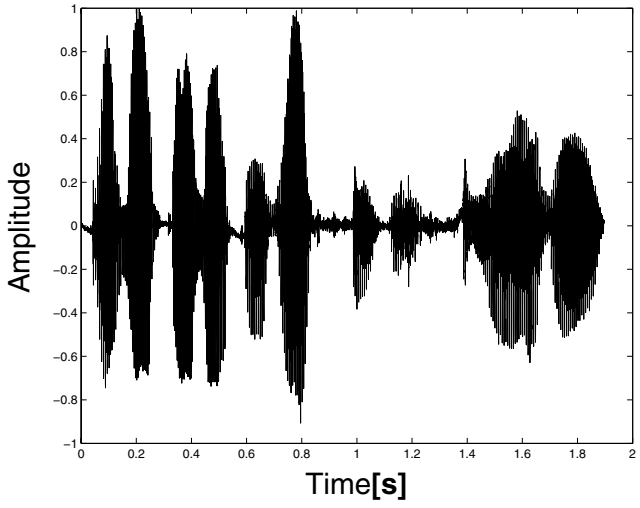

Figure 14: *Waveform of reconstructed speech(Male 1, Sentence 2)*