Reconstruction Filter Design for Bone-Conducted Speech

T.Tamiya, T.Shimamura

Department of Information and Computer Sciences
Saitama University
{tamiya,shima}@sie.ics.saitama-u.ac.jp

Abstract
Bone-conducted speech is of low intelligibility, but its quality is not affected by noise. In this paper, we take into account such properties of bone-conducted speech, and address a digital filter to reconstruct the quality of the bone-conducted speech signal obtained from a speaker. The reconstruction filter design method is derived based on a model assumption of pronunciation. Experimental results show that the reconstructed speech signal has better quality than the bone-conducted speech signal.

1. Introduction
To accomplish speech communications in a very noisy environment, a technique of picking up the vibrations of bone can be utilized [1] [2]. 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 is robust against noise [5]. However, although Kitamori et al.[6] studied the intelligibility of the bone-conducted speech signal picked up from the nose of a speaker, up to now the properties of bone-conducted speech are not well known. One known property of the speech signal measured through a bone conduction microphone is that in a noiseless environment, the quality of the bone-conducted speech signal is worse than that of the corresponding normal speech signal, which is the speech signal transmitted through air. This may be caused by the fact that the frequency components more than 1kHz deteriorates in bone-conducted speech [3]. A straightforward method to reconstruct the quality of bone-conducted speech is to emphasize the high frequency components. However, this technique has been not accepted in the current communication systems. One of the reasons of this may be that the phenomenon of bone conduction is speaker-dependent.

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 a model assumption in which the normal speech and bone-conducted speech are related mathematically. Considering the speaker dependency for bone-conducted speech, both the normal and bone-conducted speech signals are utilized to design the reconstruction filter.

2. Principle of Reconstruction
Figure 1 shows that a bone-conducted speech signal, \( d(n) \), is measured at the head 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, then \( s(n) \) and \( d(n) \) are related with the block diagram shown in Fig.3. We assume the combined system of \( B(z) \) with \( 1/V(z) \) as \( T(z) \). Furthermore, in the case where \( d(n) \) is inputted, the relation between \( d(n) \) and \( s(n) \) are shown in Fig.4, 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.4 is a direct solution to this subject. However, it is basically impossible to obtain the transfer function of the reconstruction filter, $H(z)$. Hence, we estimate $H(z)$.

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

![Diagram](image1.png)

Figure 4: A model transforms bone-conducted speech into normal speech

![Diagram](image2.png)

Figure 5: Reconstruction filter

![Diagram](image3.png)

Two transfer functions, $H(z)$ and $T(z)$, are inversely related as

$$H(z) = \frac{1}{T(z)}.$$  

(1)

In Fig.3, the relation

$$D(f) = T(f)S(f)$$  

(2)

is valid where $D(f)$ and $S(f)$ are the Fourier transforms of $d(n)$ and $s(n)$. $T(f)$ corresponds to the frequency response of $T(z)$. From (2), $T(f)$ is represented by

$$T(f) = \frac{D(f)}{S(f)}.$$  

(3)

Also from the relation in (1), the frequency response of the reconstruction filter $H(z)$, $H(f)$, is represented by

$$H(f) = \frac{S(f)}{D(f)}.$$  

(4)

Direct use of (4) is input-dependent. Common excitation source has to be assumed between $d(n)$ and $s(n)$. To avoid this situation, we utilize long-term spectra of $s(n)$ and $d(n)$ and find a reconstruction filter for the speaker. Our proposal is to use

$$\hat{H}(f) = \frac{|\hat{S}(f)|}{|\hat{D}(f)|}.$$  

(5)

where $|\hat{S}(f)|$ and $|\hat{D}(f)|$ are long-term spectra of $s(n)$ and $d(n)$. Once this filter is designed, it is always used for the speaker to transform any bone-conducted speech signals into the corresponding normal speech ones.

$\hat{H}(f)$ has only real values and the filter $\hat{H}(z)$ results in a zero-phase filter. When such a filter is used, the phase information of the input signal is not kept. However, it is well known that speech is not sensitive to phase changes. The filtering shown in Fig.5 utilizes this characteristic of speech.

3. Proposed Method

Based on the principle described in Section 2, we propose the following procedure.

1. Obtain the long-term spectra of the bone-conducted and normal speech signals, $|\hat{D}(f)|$ and $|\hat{S}(f)|$, for the speaker.
2. Design the reconstruction filter $\hat{H}(z)$ based on (5).
3. Filter the bone-conducted speech signal $d(n)$ by the reconstruction filter $\hat{H}(z)$ and obtain the reconstructed speech signal $\hat{s}(n)$.
4. Apply a speech enhancement method to the reconstructed speech signal $\hat{s}(n)$.

In Step 4, the iterative spectral subtraction method addressed in [7] is deployed. In addition to the speech enhancement process, we low-pass filters the output with a cut-off frequency of 3.5kHz so that the quality of the reconstructed speech signal is further improved.

4. Experiments

In order to investigate the performance of the proposed reconstruction method, we conducted a listening test. In a sound-isolated room, we measured simultaneously the normal speech and bone-conducted speech pronounced by a person. We gathered 2 male (Speakers A and B) and 2 female (Speakers C and D) speech data. The sampling frequency is commonly 11.025kHz.

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 which is a bone conduction microphone. The HG-17 is a headgear type, and the bone-conducted speech signal is picked up at the top of the head of the person.

The recorded speech signal consists of two sentences which have balanced phonemes, three short sentences used for daily life and successive five vowels. They are shown below (from 1to 6).

1. kare wa izen kara kagakugijyutu no sinpo to ningen no yuuki ga, harukana utyuu eno tabi wo kanou ni sita no dento kangaete imashita.
2. hitobito no byoubue to nyoraizou ni taisuru kyoumi wa happykunen no nengetsu no yotte syoujita hyoumen no bimyou na sikisaihenka ni aru.
3. kuroi doresu no hito wa dare desuka?
4. anata yori se ga takai no desuka?
5. daisyou samazama na tehburu ga arimasu.
6. “/a/i/u/e/o/"
What kinds of normal speech signals should be used for designing the reconstruction filter is an issue to be considered. This is because in a noisy environment it is impossible to utilize the normal speech signal recorded simultaneously with the bone-conducted speech signal. The above Sentence 1 and Sentence 2 were connected and one long sentence was created. This sentence was used as one of two normal speech signals to design the reconstruction filter. The other was Sentence 6, successive vowels, which are normal, too. To compare the accuracy of the filter design, additionally we used the normal speech sentence of the same phrase as the bone-conducted speech sentence. They are summarized as

a) original bone-conducted speech sentence
b) normal speech sentence of the same phrase as the bone-conducted speech sentence
c) “/a//i//u//e//o/”
d) combined long sentence

In the experiments, as the benchmark, the bone-conducted speech itself is also considered for assessment.

In the proposed method, the long-term spectra were obtained based on the calculation method described in [4]. The frame length was set to 512. For filtering by the reconstruction filter, a Fast Fourier Transform (FFT) based convolution was used with 2048 frequency points. In Step 4 of the proposed method, that is actually in the iterative spectral subtraction method, the spectral subtraction method was repeated five times based on the results in [7].

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 reconstructed speech signals, and listeners judged which reconstructed speech signal has higher intelligibility. The rate of selection was averaged for the twenty listeners and the score was obtained where the averaged selection rate was changed into Thurstone Scale. At first, the listeners heard the original normal speech sound. Next they heard the two picked reconstructed speech. And then they judged. We did not inform a priori the listeners what kinds of speech signals were used.

Tables 1 to 4 are the selection rate averaged with 5 sentences (from 1 to 5) of each speaker. Figure 6 shows the listening score averaged with the 5 sentences (in Thurstone Scale the quality of speech is better as the value becomes larger.) In the tables and figure, “bone”, “same”, “vowel” and “balance” mean the use of a, b, c, and d as the normal speech signals for designing the reconstruction filter ( only a is of bone-conducted speech ). From Tables 1 to 4 and Figure 6, it is clearly observed that the 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 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.

From Tables 1 to 4 and Figure 6, it is also observed that using the long sentence which has balanced phonemes produces better quality of the reconstructed speech signal than using successive vowels. As expected, using the speech sentence of the same phrase as the bone-conducted speech sentence produces the best performance.

Figure 7 shows the spectrograms of male (Speaker B’s) Sentence 5. And Figure 4 shows the spectrograms of female (Speaker C’s) Sentence 4. Commonly the 5 spectrograms in
each figure are, from the top, the normal speech signal, the bone-conducted speech signal, the reconstructed speech signal using the same phrase as the bone-conducted speech signal, the reconstructed speech signal using successive vowels and that using the long sentence. From these figures, we can confirm that high frequency components of the bone-conducted speech signal are restored by the proposed method. This would be the reason why the reconstructed speech signal has a better quality than the bone-conducted speech signal.

Figure 7: Spectrograms for Speaker B

Figure 8: Spectrograms for Speaker C

6. Concluding Remarks

In this paper, it has been shown that the long-term spectrum of speech could be effectively used in designing the reconstruction filter of bone-conducted speech. From the experimental results, it has been observed that the quality of the reconstructed speech is improved by utilizing a long sentence which has balanced phonemes. Future work will be focused on optimizing the parameters used in the proposed method.

7. References