2006 IEEE International
Symposium on Signal Processing
and Information Technology

ON EQUALIZATION OF BONE CONDUCTED SPEECH FOR IMPROVED SPEECH QUALITY

Kazuhiro Kondo, Tomoe Fujita† , and Kiyoshi Nakagawa

Yamagata University
Department of Electrical Engineering
4-3-16 Jonan, Yonezawa, Yamagata 992-8510, Japan
{kkondo,nakagawa}@yz.yamagata-u.ac.jp

Abstract-*We propose an equalizer that attempts to improve the perceived speech quality of bone-conducted speech input with ear-insert microphones, which can provide clean speech input in noisy environments. We first show that the transfer characteristics of bone-conducted speech are both speaker and microphone dependent, and propose an equalizer which is trained using simultaneously recorded airborne and bone-conducted speech. The short-term FFT amplitude ratio of airborne and bone-conducted speech is used. The amplitudes are averaged and smoothed extensively before the ratio is calculated. The trained equalizer is applied to bone-conducted speech in the frequency domain. We show that the proposed equalizer provides notable quality improvement on the bone-conducted speech input, both subjectively and objectively. We also show that the application of spectrum subtraction also helps decrease some constant level background noise found in these types of microphones.*

Keywords-*bone conducted speech, equalization, Fourier transform magnitude, spectrum subtraction*

1. INTRODUCTION

Mobile communication devices including cellular phones have become quite ubiquitous. We now are likely to input speech in environments where the surrounding noise is prohibitive, *e.g.* train stations, crowded streets, etc. Speech input under these conditions will become too noisy, and effective speech communication will become difficult. There are microphones that can cancel noise, *i.e.* noise cancelling microphones. These have shown some success in reducing noise, but require multiple input positioned at specific orientations. An alternative approach would be to input speech that conducts through the head (the skull). Since this path does not go through external aerial paths, it is virtually free of external noise. This is known as bone-conducted speech, and can be picked up typically at the ear canal using ear-insert microphones [1]. The conducting paths of an ear-insert microphone and a conventional aerial microphone are illustrated in Fig. 1. Ear-insert microphones are becoming increasingly attractive since they can potentially provide both input and output in one compact form factor, without a boom microphone placed in front of the mouth.

However, bone-conducting speech is known to be muffled and suffer lower quality and intelligibility than speech obtained through conventional aerial microphones. Fig. 2 shows spectrograms of speech recorded with a conventional aerial microphone (upper plot) and an ear-insert microphone. The recordings were of a female speaker reading a Japanese sentence. These samples were recorded simultaneously, as will be described in the next section. As can be seen, most of the higher frequency region that can be seen in the airborne recording is gone in the bone-conducted recording. There is also a very strong component close to DC in the latter, although this is somewhat difficult to see. These two seems to be the main contributing factor in the relatively low speech quality of the bone-conducted speech.

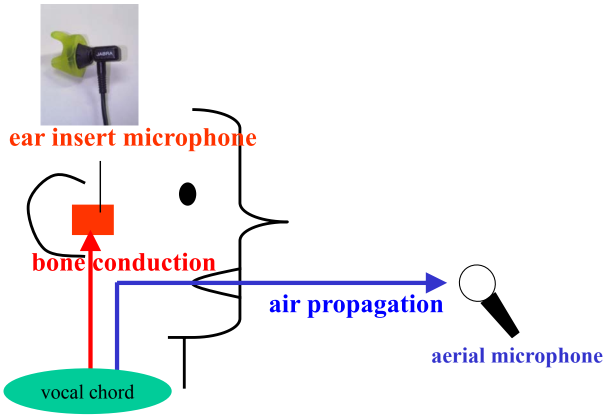


Fig. 1. Conducting paths of a bone-conducting ear-insert microphone and an ordinary microphone.

There have been attempts to equalize bone-conducted speech to enhance the high frequency ranges for improved quality. Tamiya et al. have attempted to use simple long-term FFT amplitude ratio of the airborne and bone-conducted paths [2]. They attempt to estimate airborne speech from the bone-conducted speech by applying this ratio in the frequency domain. The improvements they have shown seems promising but still limited and marginal. Kimura et al. have attempted to equalize bone-conducted speech using modulation filterbanks [3]. Although still in its preliminary phase, they have shown some improvements on a single speaker with closed training. Liu et al. have attempted to use a multiple-input microphone to obtain noise-free input for speech recognition systems [4, 5]. They

† Currently with Alpha Systems Inc.

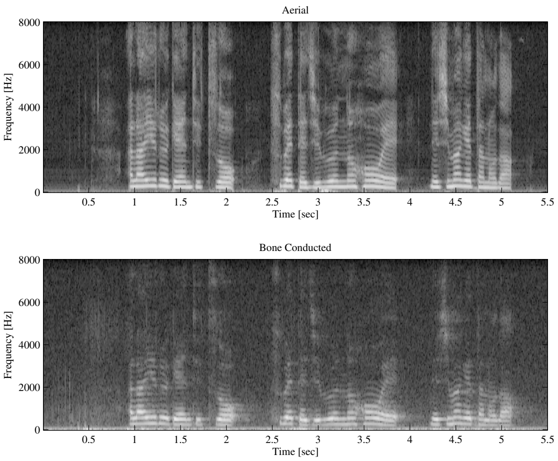


Fig. 2. Spectrograms of female speaker f02 recorded simultaneously using a conventional aerial microphone (upper plot) and an ear-insert microphone (lower plot).

attempt to use a headset with both a bone-conductive and close-talk microphone integrated into one. Simultaneous inputs from the microphones are combined into a single clean speech estimate. Their results seem promising, but require special multiple input hardware.

In this paper, we first show that the transfer function of the bone-conduction path is speaker and microphone dependent, and the transfer function should be individualized for effective equalization. Then, we propose a speaker-dependent short-term FFT based equalization with extensive smoothing and training. We show that the proposed scheme improves the subjective quality especially when the degradation due to bone-conduction is large. We then show that spectrum subtraction can decrease the low level background noise found in ear-insert microphones.

In the next section, we describe the speech database in which speech was recorded simultaneously with both a conventional microphone and a ear-insert microphone. Section 3 describes the proposed short-term equalization scheme, followed by experimental results. Section 5 describes the application of Wiener filters and its evaluation results. Finally, conclusions are given.

2. TRANSFER CHARACTERISTICS OF THE BONE-CONDUCTED PATH

We first collected a speech database where simultaneous recordings of airborne speech with a conventional microphone and bone-conducted speech with an ear-insert microphone were made. Recordings were made for three male speakers and three female speakers at 16 kHz sampling, 16 bits per sample. Each speaker read 50 phonetically balanced Japanese sentences defined by ATR [6]. The airborne speech was recorded with the Sennheiser headset microphone HMD410-6. The bone-conducted speech was recorded simultaneously with the JABRA earphone by JABRA Corporation. For comparison, we also made simultaneous recording with the same headset microphone and another ear-insert microphone, the PHS earphone microphone by Nihon Auto Giken Kogyo Co. Ltd., which picks up the bone-conducted speech at the cheek bone just below the ear lobe. Recordings were made for one male and one female speaker both reading the same 50 ATR sentences.

Fig. 2 showed spectrograms of female speech in which there was a large difference in the high frequency regions of the airborne and bone-conducted speech. However, Fig. 3 shows spectrograms of a male speaker where the difference between the two speech is quite small. It is almost impossible to tell the subjective quality difference as well. These two speakers are extreme cases, but we found that the transfer characteristics of the bone-conducting path differ significantly from speaker to speaker. The transfer characteristics also differ between the two bone-conducted speech microphones, the JABRA earphone and the PHS earphone. We also made a small number of recordings on a later date with the same speaker and microphone, but the transfer characteristics do not seem to differ significantly from day to day. Thus, it seems that the equalization needs to be speaker-dependent as well as microphone-dependent. In the next section, we propose an equalization scheme which takes into account these dependencies.

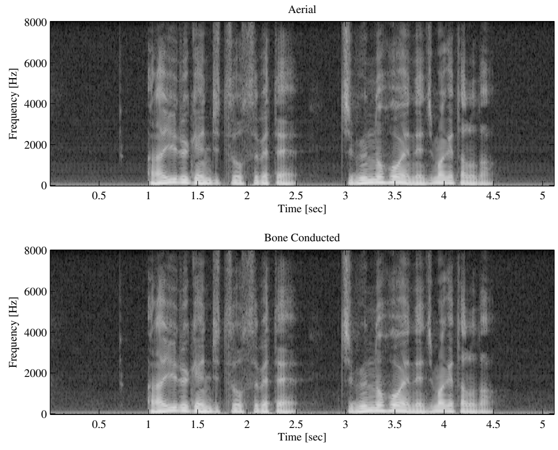


Fig. 3. Spectrograms of male speaker m01 recorded simultaneously using a conventional aerial microphone (upper plot) and an ear-insert microphone (lower plot).

3. EQUALIZATION OF THE BONE-CONDUCTED PATH CHARACTERISTICS

From the analysis of the database collected as described in the previous section, we decided to use an equalizer based on speaker and microphone dependent short-term FFT magnitude ratio. In other words, assuming that $|H_{ab}(f_k)|$ is the magnitude FFT spectrum of the airborne speech, and $|H_{bc}(f_k)|$ is the magnitude FFT spectrum of the same bone-conducted speech, the trained magnitude equalization characteristics can be defined as

$$|H_{eq}(f_k)| = \frac{|H_{ab}(f_k)|}{|H_{bc}(f_k)|} \tag{1}$$

where f_k is the frequency bin number. Since we are using short-term FFT, the input speech utterance is segmented into a number of frames. Thus, we come up with a number of $|H_{eq}|$ estimates, which are averaged to obtain one mean estimate. The first and last three frames were left out of the averaging since the boundary frames are normally silent. The averaged estimate in some frequency bins seemed to include impulse noise components due to transient characteristics in some frames. Therefore, we used a simple moving averaging to smooth these transients.

$$\overline{|H_{eq}(f_k)|} = \sum_{i=k-j}^{k+j} \frac{|H_{eq}(f_i)|}{2j+1} \qquad (2)$$

The use of averaging to smooth the ratio estimates may in some cases degrade the equalization in transient areas. However, we came to the conclusion through informal testing that the benefit of having a "smoothed out" equalization characteristics in the dominant stable portions seems to outweigh having matching characteristics in transient areas.

Using the obtained equalization characteristics, the equalized bone-conducted speech, which should resemble its airborne counterpart, can be given as

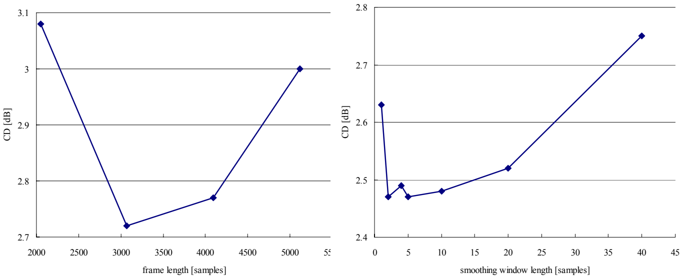
$$|\hat{H}_{ab}(f_k)| = |H_{bc}(f_k)| \cdot \overline{|H_{eq}(f_k)|} \qquad (3)$$

The phase is left as is.

The quality of the equalized speech was found to be dependent on both the FFT frame length and the moving average smoothing length. Fig. 4 shows the LPC Cepstrum Distance (LPC CD) [7] with different frame and window length. The speech was from female speaker f01. Both the bone-conducted and airborne speech for a single utterance was used to calculate $\overline{|H_{eq}(f_k)|}$, and this was used to equalize a separate bone-conducted speech utterance. The LPC CD calculation was done between this equalized speech and the simultaneously recorded airborne speech. For (a), we kept the moving average length to 5 samples (2 before and 2 after). For (b), the FFT frame length was kept constant at 4096 samples.

As can be seen, there is an apparent optimum value for both the FFT frame length and the moving average smoothing window length. In the following experiments, we will use 3072 samples for the former, and 10 samples for the latter.

We also attempted to use multiple utterances to obtain a better estimate of the average magnitude values to calculate $\overline{|H_{eq}(f_k)|}$. For all utterances, the first and last three frames were excluded since most are silent frames. The results are shown in the next section.



(a) FFT frame length (b) Smoothing window length

Fig. 4. Effect of frame and window length on equalized speech.

4. EXPERIMENTS

We evaluated both subjective and objective performance of the proposed equalization scheme. We used speech in the database described in Section 2. Three male (speakers m01, m02, m03) and three female speakers (f01, f02, f03) were evaluated. The JABRA earphone was used for all speakers. For speakers m01 and f01, we also evaluated the performance using the PHS earphone. In all cases, we trained the equalization filter using 1, 4 and 48 utterances. The training was speaker and microphone dependent, *i.e.*, the training was done using utterances spoken by the same speaker and microphone, then the equalization was applied to a different utterance spoken by the same speaker with the same microphone. For comparison, we also included a closed training condition, *i.e.*, equalization was trained on one utterance, and equalization is applied on the same utterance. This is obviously not possible in real time, but should give an upper bound on the performance. Table 1 gives the LPC CDs for the JABRA earphone, and Table 2 gives the LPC CDs using the PHS earphone.

As can be seen, use of some form of equalization lowers the LPC CDs compared to speech with no equalization (denoted "none"). Closed training gives minimum distances in the majority of cases, but not significantly; the extensive averaging and smoothing seems to increase the distances in this case. With open training, in most cases, one utterance is enough to give decent reduction in the LPC CDs.

With speaker m01 using the JABRA earphone, an exceptional case where the speech quality of the airborne and bone-conducted speech was virtually indistinguishable, the distances are significantly lower than other speakers to start with. Thus equalizers have no significant effect in this case. Interestingly, with the PHS earphone, the same speaker m01 shows significantly larger LPC CDs compared to the airborne speech, showing similar characteristics as the other speakers.

Table 1. LPC CDs of equalized speech using the JABRA earphone

speaker	equalization				
	none	closed training	open training		
			1 utt.	4 utt.	48 utt.
m01	1.09	0.62	0.67	1.23	1.18
m02	3.68	2.91	2.84	2.54	2.83
m03	4.47	3.10	2.42	2.69	2.72
f01	3.79	2.73	2.73	3.09	3.31
f02	4.02	3.36	3.33	3.30	-
f03	3.53	2.25	2.53	2.29	-

We also conducted Mean Opinion Score (MOS) testing for the same speech utterances in order to rate the subjective quality improvements. Ten listeners rated each utterance on a five point scale, 5 for "very good", 4 "good", 3 "fair", 2 "bad" and 1 "very bad". Table 3 lists the results for the JABRA earphone, while Table 4 lists the results for the PHS earphone.

In most cases, use of an equalizer provides improvements over bone-conducted speech with no equalization ("none"). Interestingly,

Table 2. LPC CDs of equalized speech using the PHS earphone

speaker	equalization				
	none	closed training	open training		
			1 utt.	4 utt.	48 utt.
m01	4.47	3.72	3.87	3.32	3.26
f01	4.33	3.22	3.54	3.09	3.07

the open trained equalizer in many cases outperforms the equalizer with closed training. Again, one or four utterances seem to be enough to train a decent equalizer.

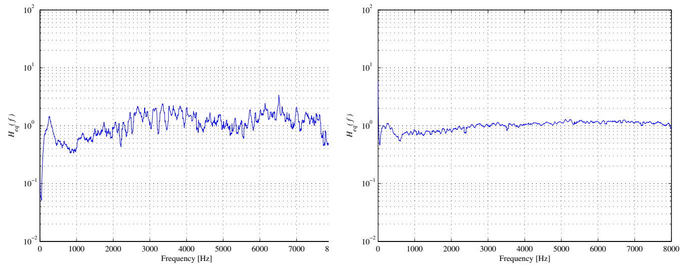
Fig. 5 shows the trained equalizer amplitude ratio for speaker f02 and m01 using the JABRA earphone. For speaker f02, the lower frequency regions are attenuated while the higher frequencies have a gain larger than 1.0. There is also a notable attenuation close to DC which apparently tries to reduce the spectrum components close to DC, as visible in Fig. 2. For speaker m01, since the airborne and bone-conducted speech is virtually identical, the trained characteristics are flat across all bandwidth.

Table 3. MOS results of equalized speech using the JABRA earphone

speaker	air-borne	bone conducted, equalization				
		none	closed training	open training		
				1 utt.	4 utt.	48 utt.
m01	3.5	3.7	3.9	3.7	3.5	3.3
m02	4.0	2.3	2.2	2.6	2.4	2.5
m03	3.9	2.6	2.9	3.0	2.8	3.0
f01	3.6	2.3	2.4	2.6	2.5	2.7
f02	3.8	2.5	2.5	3.3	2.5	-
f03	3.2	2.3	2.9	2.7	2.8	-

Table 4. MOS results of equalized speech using the PHS earphone

speaker	air-borne	bone conducted, equalization				
		none	closed training	open training		
				1 utt.	4 utt.	48 utt.
m01	4.0	2.4	2.8	2.5	2.6	2.1
f01	4.0	2.3	3.1	2.8	2.6	2.6



(a) speaker f02

(b) speaker m01

Fig. 5. Trained equalizer amplitude ratio for the JABRA earphone.

5. INTRODUCTION OF SPECTRUM SUBTRACTION

As seen in the previous section, the equalizer does improve the quality for some speakers. However, the constant low frequency noise is still present and noticeable. Since this noise is constant, the application of classical spectrum subtraction [8] can potentially reduce this noise. However, since the proposed equalizer is essentially a high frequency emphasizer, musical noise which may be introduced by spectrum subtraction will be emphasized in the high frequency regions as well, which could be prohibitively annoying. Thus, we decided to use Wiener filters [9] since these filters are known to introduce less musical noise than explicit subtraction as described in [8]. We also modify its definition to introduce a control factor to limit the noise subtraction to a "modest" level.

Wiener filters can be defined as

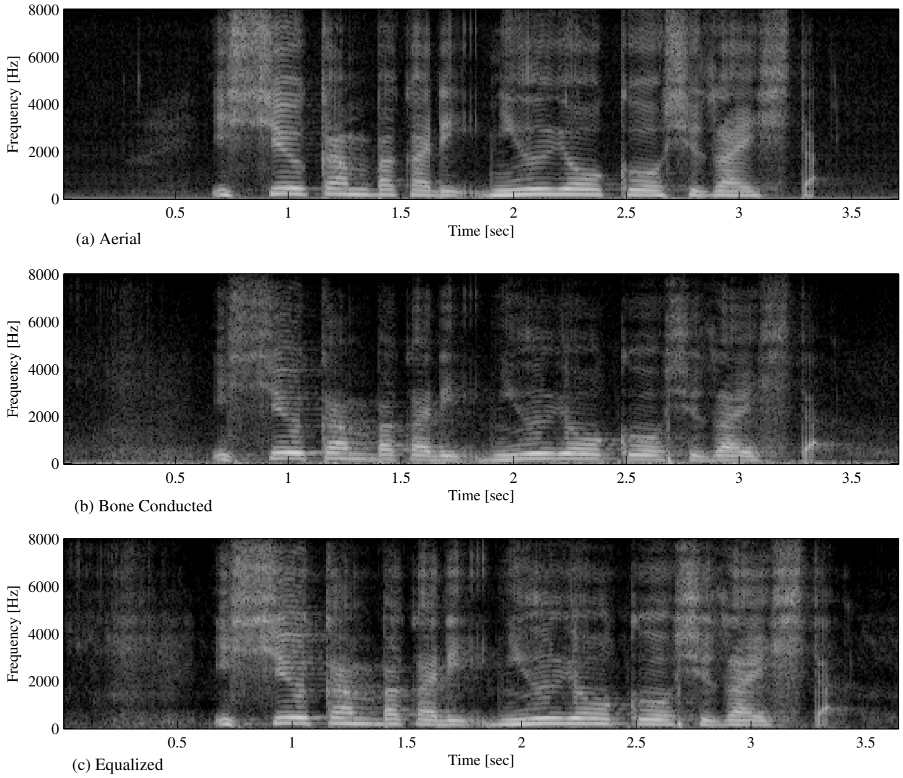
$$\hat{S}(\omega) = H(\omega)X(\omega) \tag{4}$$

where $\hat{S}(\omega)$ is the filtered signal, $X(\omega)$ is the noisy signal, and $H(\omega)$ is the Wiener filter. We modified $H(\omega)$ to be

$$H(\omega) = \frac{P_x(\omega)}{P_x(\omega) - \alpha P_n(\omega)} \tag{5}$$

where $P_x(\omega)$ is the power spectrum of the noisy signal, $P_n(\omega)$ is the power spectrum of the estimated noise, and α is the control parameter. The control parameter α is empirically set to 0.5 to balance between noise reduction and musical noise introduction, as stated above. This filter is applied before any processing on both the bone-conducted speech and aerial speech, both during training and equalization.

Tables 5 and 6 are the MOS results for equalization with spectrum subtraction with the JABRA and the PHS earphone respectively. Ten subjects rated the speech, as in previous experiments. The other experimental conditions are the same as described in section 4, except the number of utterances used in the training, which was 1, 12, and 48. As can be seen, for some speakers, the MOS improves modestly, while for others, the MOS stays the same. As expected, the background noise does seem to noticeably decrease with no apparent effect on the speech in some speakers, thereby improving the overall subjective impression. However in some speakers, the high frequency components of noticeable musical noise is emphasized, decreasing the subjective quality. Fig. 6 shows the spectrograms of simultaneously recorded airborne speech (uppermost plot), the bone-conducted speech (middle plot), and the equalized speech using Wiener filters (lowest plot) for female speaker f03. The higher frequency regions in the bone-conducted speech which are somewhat attenuated (thinner), is regained to some extent using the proposed equalizer.



(a) Aerial

(b) Bone Conducted

(c) Equalized

Fig. 6. Spectrograms of female speaker f03 recorded simultaneously using a conventional aerial microphone (upper plot) and an ear-insert microphone (middle plot), and its equalized speech (lower plot).

Table 5. MOS results of equalized speech with spectrum subtraction using the JABRA earphone

speaker	air-	bone conducted, equalization				
		none	closed	open training		
	borne		training	1 utt.	12 utt.	48 utt.
m01	3.3	3.5	4.0	4.0	4.0	4.2
m02	3.8	2.8	2.6	2.6	2.7	2.9
m03	4.0	2.7	2.7	2.7	2.9	3.0
f01	3.3	2.0	2.4	2.0	2.4	2.1
f02	3.7	2.1	2.0	2.2	2.3	2.4
f03	3.9	2.4	2.9	2.7	3.1	3.4

Table 6. MOS results of equalized speech with spectrum subtraction using the PHS earphone

speaker	air-	bone conducted, equalization				
		none	closed	open training		
	borne		training	1 utt.	4 utt.	48 utt.
m01	3.7	2.0	2.3	2.8	2.4	2.2
f01	3.7	2.2	2.3	2.3	2.8	2.5

6. DISCUSSIONS AND CONCLUSIONS

In this paper, we proposed an equalizer which attempts to improve the intelligibility of bone-conducted speech input with an ear-insert microphone. We first showed that the transfer characteristics of bone-conducted speech are both speaker and ear-insert microphone dependent. We then proposed an equalizer which is designed using simultaneously recorded airborne and bone-conducted speech. The equalizer is based on the short-term FFT amplitude ratio of airborne and bone-conducted speech. The FFT amplitudes are averaged and smoothed before the ratio is calculated. This equalizer is applied to bone-conducted speech in the frequency domain. We showed that the proposed equalizer provides a notable improvement of the bone conducted speech quality both objectively (using LPC Cepstrum Distance) and subjectively (using Mean Opinion Scores). We also showed that spectrum subtraction using Wiener filters can decrease the constant low frequency noise introduced in ear-insert microphones.

In practice, the proposed method does require an initial training phase. However, it should not be unreasonable to ask the user to use the ear-insert microphone, and any PC microphone they may have in hand, and have them read the prompted sentence in a clean environment before initial use. As we have seen, the training should not require many sentences to be read for decent equalizer training. The trained equalizer can then be used for the same user again by having them pick from an enlisted user id.

Although the proposed equalizer was applied in the frequency domain, it is relatively easy to implement the designed filter in the time domain, which should reduce the complexity, as well as the processing delay.

We would also like to further improve the quality of the equalized speech. Since the characteristics of the equalizer is fixed to high frequency emphasis, the resident high frequency noise in the silent and low energy speech regions may be overly emphasized, which can lower the perceived quality. Thus, the introduction of adaptive equalization according to either energy (silence detection) or phoneme (low energy consonant detection) may help. Also, limiting the effective frequency region of the spectrum subtraction to lower regions might help.

ACKNOWLEDGEMENT

Part of this work was supported by the city of Yonezawa. The authors would like to thank their support. They also thank the members of the Nakagawa Laboratory for their support and participation in the discussions and experiments.

REFERENCES

[1] R. Black, "Ear-insert microphone," *J. Acoust. Soc. Am.*, vol. 29, no. 2, pp. 260–264, Feb. 1957.

[2] T. Tamiya and T. Shimamura, "Reconstruction filter design for bone-conducted speech," in *Proc. ICSLP 2004*, Aug. 2004, vol. II, pp. 1085–1088.

[3] K. Kimura, M. Unoki, and M. Akagi, "A study on a bone-conducted speech restoration method with the modulation filterbank," in *Proc. NCSP05*, March 2005, pp. 411–414.

[4] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, "Direct filtering for air- and bone-conductive microphones," in *Proc. IEEE MMSP*, 2004.

[5] Z. Liu, A. Subramanya, Z. Zhang, J. Droppo, and A. Acero, "Leakage model and teeth clack removal for air- and bone-conductive integrated microphones," in *Proc. ICASSP*, March 2005, vol. I, pp. 1093–1096.

[6] Japan Information Processing Development Corporation, "ASJ continuous speech corpus for research," 1991.

[7] A. H. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 24, no. 5, pp. 380–391, Oct. 1976.

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 27, no. 2, pp. 113–120, April 1979.

[9] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, chapter 5, John Wiley & Sons, West Sussex, England, 1996.