Journal of Signal Processing, Vol. 10, No. 6, pp. 407–417, November 2006

PAPER

A Study on Restoration of Bone-Conducted Speech with MTF-Based and LP-Based Models

Thang Tat Vu, Kenji Kimura, Masashi Unoki and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {vu-thang, kkimura, unoki, akagi}@.jaist.ac.jp

Journal of Signal Processing

信号処理

Journal of Signal Processing, Vol. 10, No. 6, pp. 407–417, November 2006

PAPER

A Study on Restoration of Bone-Conducted Speech with MTF-Based and LP-Based Models

Thang Tat Vu, Kenji Kimura, Masashi Unoki and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {vu-thang, kkimura, unoki, akagi}@.jaist.ac.jp

Abstract **Bone-conducted speech in an extremely noisy environment seems to be more advantageous than normal noisy speech (i.e., noisy air-conducted speech) because of its stability against surrounding noise. The sound quality of bone-conducted speech, however, is very low and restoring bone-conducted speech is a challenging new topic in the speech signal-processing field. We describe two types of models for restoration: one based on the modulation transfer function (MTF) and the other based on linear prediction (LP). The MTF-based model is expected to yield a restored signal with higher intelligibility while the LP-based model is expected to yield one that is not only more intelligible to human hearing systems but also enables automatic speech recognition (ASR) systems to achieve better performance. To evaluate the ability of these models to improve voice-quality, we compared them with the other previous two models using one subjective and three objective measurements. The mean opinion score (MOS) and log-spectrum distortion (LSD) were used to evaluate the improvements in intelligibility, which is useful for human hearing systems. The distances based on LP coefficients and mel-frequency cepstral coefficients (MFCCs) were used to evaluate the improvements in cepstral distances which are useful for ASR systems. The results proved that both the MTF-based and LP-based models are better than the other previous models for improving intelligibility. They particularly proved that the LP-based model produces the best results for both human hearing and ASR systems.**

Keywords: bone-conducted (BC) speech, air-conducted (AC) speech, modulation transfer function (MTF), linear prediction (LP), speech intelligibility

1. Introduction

The sound quality and intelligibility of speech are influenced by their transmission environments. It is very difficult for automatic speech recognition (ASR) systems as well as humans to accomplish speech communications in an extremely noisy environment. There are many different complex models and/or algorithms that are used as a solution to canceling or reducing interfering noises. These are only efficient at low and medium noise levels and are ineffective when the noise levels are too high.

Another possible solution is to use a special microphone to record the speech signal transmitted through the speaker's head and face. This recorded signal is referred to as "bone-conducted speech". Its stability against interfering noise from a noisy environment seems to make bone-conducted (BC) speech more advantageous than noisy air-conducted (AC) speech. Although BC speech is not affected by external noise as AC speech is, there is a drawback to using BC speech; the signal is attenuated in a complex manner when it is transmitted through bone-conduction. This causes the voice-quality of BC speech, which means both its intelligibility by human hearing systems and features that are robust in ASR systems, to be very poor. If the voice-quality of BC speech can be improved, the restored signal can be applied to speech applications in noisy environments with greater efficiency instead of using noisy AC speech. Such applications include human hearing aids and machine hearing systems. Since it is very difficult to blindly restore BC speech, this is a challenging new topic in the speech signal processing field.

The attenuation of the BC speech signal varies for different positions of measurement (BC microphone

positions), pronounced syllables, and speakers. This is because the characteristics of bone-conduction change for different measuring positions, and the distribution of frequency components varies with speakers who pronounce syllables differently. This attenuation is generally strong at high frequencies and it seems to be low-pass filtering with a cut-off frequency of about 1 kHz [1]. A straightforward method of restoring BC speech is to emphasize these attenuated frequency components by using high-pass filtering (inverse of the low-pass filtering previously described). However, it is difficult to adequately design one unique kind of high-pass filtering that is independent of pronounced syllables, speakers, and measuring positions. There are various methods of deriving inverse filtering such as the cross-spectrum method [2] and the long-term Fourier transform [3, 4], but these yield restored signals with artifacts such as musical noise and echoes so there are only slight improvements in voice-quality.

We investigated the relationships between BC and clean AC speech using an AC/BC speech database as an essential step toward constructing complete models to restore BC speech and find significant characteristics (for inverse filtering) that would be useful in restoring BC speech. We propose two models of restoration from these results. The first is based on the modulation transfer function (MTF) [5] and the second is based on linear prediction (LP) [6]. All current models (including the ones we propose) obtain their parameters by using various information on AC speech; we regarded the consideration of blindly determined model parameters to be the next step we will undertake in future work.

The MTF-based model originates from the idea that the temporal envelope contains most of the important information related to speech intelligibility, and this intelligibility can be improved by using power envelope inverse filtering such as that with the speech dereverberation method [7, 8]. The LP-based model originates from the idea that the information corresponding to the source (glottal) characteristics as the LP residue is the same for both AC and BC speech signals. Therefore, adaptive inverse filtering will be primarily derived from the LP coefficients of AC and BC speech related to filter information (vocal tract).

Both models manipulate source and filter information; the temporal envelopes and carriers are used in the MTF-based model and the LP residue and LP coefficients are used in the LP-based model. The difference is in the processing domain, the MTF-based model restores BC speech in the time domain with each sub-band (channel) and the LP-based model restores BC speech in the frequency domain with each frame. We investigated both models, which are expected to yield restored signals that are not only more intelligible to human hearing systems but which also enable ASR systems to achieve better recognition.

Table 1 List of equipment

Measurement site	Soundproof room
Measurement positions	5 positions
Number of speakers	10 people
Recorder	SONY, TCD-D10 ProII
Sampling frequency	48 kHz
Sample size	16 bits
Mic. A for AC speech	SONY, C536P
Mic. amp. A for AC speech	SONY, AC148F
Mic. B for BC speech	Temco, HG-17
Mic. C for BC speech	Handmade
Mic. amp. B for BC speech	Handmade

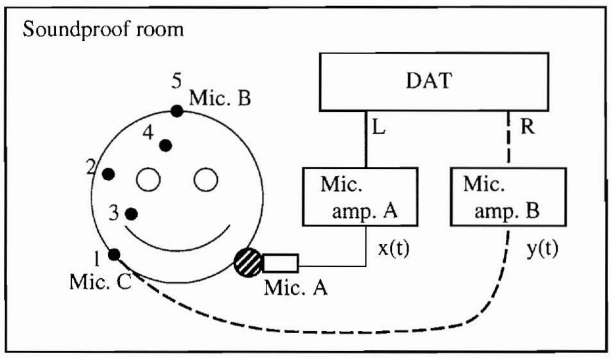

Fig. 1 Environment for recording AC/BC speech

The rest of this paper is organized as follows. The next section describes the AC/BC speech database that was used in our study. Section 3 explains our assumptions and approaches to BC speech restoration. We then present our MTF-based and LP-based models in sections 4 and 5. These models are evaluated and discussed in section 6. Section 7 concludes with a summary and mentions future work.

2. AC/BC Speech Database

A database is indispensable for analyzing the relationships and differences between BC speech and clean AC speech signals before any models are used to restore BC speech. We constructed a large-scale database containing pairs of BC and clean AC speech signals recorded simultaneously using a DAT system (two channels).

Figure 1 and Table 1 show the environment and equipment we used to construct this database. The BC speech was collected at five different positions on the head and face, i.e., the (1) mandibular angle, (2) temple, (3) philtrum, (4) forehead, and (5) calvaria. Thus, one position was associated with one pair of clean AC speech and BC speech. Microphone B was only used at position 5 and microphone C was used at the other positions. Ten speakers (eight males and two females) participated in the recording

of the pronounced speech of 100 Japanese words and 45 Japanese syllables.

The selected words were chosen from the NTT database by their degree of familiarity [9]. The database had two parts. The first was (i) a Japanese word dataset of 100 Japanese words selected from Japanese word lists compiled by NTT-AT (2003). With 10 speakers, 100 words, and 5 measurement positions, there were 5000 pairs of wave files. The second was (ii) a Japanese syllable dataset of 45 Japanese syllables. With 10 speakers, 45 syllables, and 5 measurement positions, there were 2250 pairs of wave files.

3. Restoration Approaches

3.1 Assumptions

In the work reported here, clean AC speech was used instead of noisy AC speech and this will be referred to as "AC speech" after this. AC speech was recorded/observed simultaneously with BC speech, as in the AC/BC speech database. We assumed that there were existing relationships between AC and BC speech that would be significant to restore BC speech. Therefore, our strategy was to construct restoration models through the following steps: (1) to investigate the relationship between clean AC and BC speech signals, (2) to find significant characteristics and restoration models to restore BC speech (toward clean AC speech), and (3) to consider how to blindly determine model parameters only from observed BC speech and apply these to realistic communication. This paper focused on the first two essential steps. We considered different approaches to designing inverse filtering for the restoration models, which will be discussed in the next section.

3.2 Approaches to restoring BC speech

There are three types of transfer functions that can be used as different approaches to restoring BC speech in Fig. 2. In general, these should be investigated as transmission characteristics from AC speech to BC speech before designing the inverse filtering to restore observed BC speech.

As we can see from Fig. 2(a), one straightforward approach is to design the inverse transfer function from $y(t)$ to $x(t)$. There are also the cross-spectral and long-term FFT methods [2]-[4]. These are used to construct the inverse transfer function from the BC spectrum to the AC spectrum using cross-spectrum or FFT methods. The LP-based model that we propose can be regarded as this kind of approach to restoration because it is used to design the inverse transfer function that can easily restore BC speech. However, here, its inverse filtering was designed by using the

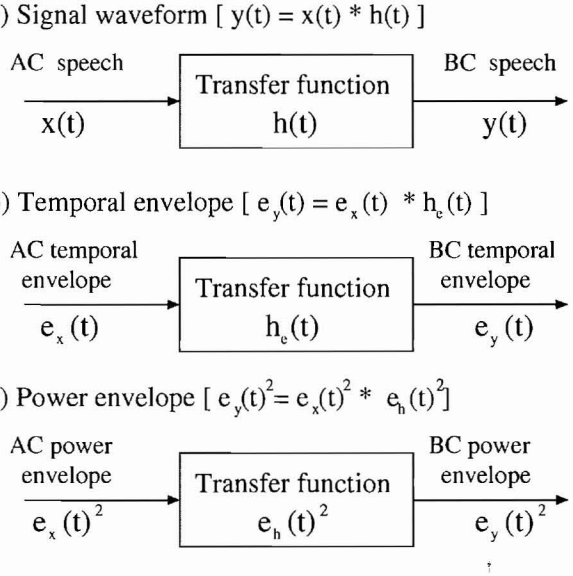

Fig. 2 Definitions of transfer functions: (a) signal waveform, (b) temporal envelope, and (c) power envelope

LP coefficients related to the spectral characteristics of signals.

The signal in each sub-band (channel) can be restored independently where the temporal envelope and carrier are the two components to be restored, when representing a signal in the filterbank. Moreover, the temporal envelope contains most of the important information related to speech intelligibility rather than the carrier difference [10, 11]. Thus, the primary goal is to restore the temporal envelope in each channel, as can be seen in Fig. 2(b). The cross-spectrum and long-term FFT methods, for example, can be applied to temporal-envelope inverse filtering (inverted $h_e(t)$) for BC speech.

When focusing on the representation of the power envelopes of signals (Fig. 2(c)) rather than those of the temporal envelopes (Fig. 2(b)), we can clearly express the characteristics of the power envelopes as follows [7, 8]:

$$e_y(t)^2 = e_h(t)^2 * e_x(t)^2 \quad (1)$$

where "$*$" denotes convolution, and $e_x(t)^2$, $e_y(t)^2$, and $e_h(t)^2$ correspond to the power envelopes of signals $x(t)$, $y(t)$, and $h(t)$. In this relation, the carriers were assumed to be mutually independent respective white noise random variable functions (see [7] for details). Thus, restoring the power envelope in each channel is the primary goal. As mentioned above, the temporal envelope contains most of the important information related to speech intelligibility. The modulation transfer function of $e_h(t)^2$ is strongly related to speech intelligibility and this can be applied to deriving power envelope inverse filtering, so that speech intelligibility

can be improved by using this kind of inverse filtering used in MTF-based speech dereverberation [7, 8].

Although the MTF-based and LP-based models seem to be different approaches, both, in fact, involve the same concept to restore the observed BC signal, i.e., the same representation of source and filter information. The MTF-based model tries to restore the temporal power envelope in each channel, while the LP-based one tries to restore the spectral envelope. Thus, their only difference lies in the processing domain, one processing in the time domain, and the other processing in the frequency domain. In the next two sections, we present the two models based on the MTF and the LP concepts in detail.

4. MTF-Based Model

4.1 Model concept

The MTF concept was proposed by Houtgast *et al.* to measure the room acoustics to assess what effect the enclosure had on speech intelligibility [10, 12, 13]. The model originated from the idea that the temporal envelope contained most of the important information related to speech intelligibility, and this intelligibility could be improved if the power envelope of BC speech were restored, as processing did with the method of speech dereverberation [7, 8]. Thus, the primary goal of the MTF-based model was to restore the power envelopes of BC speech by using power envelope inverse filtering related to the MTF concept, in the filterbank [5]. According to this concept, the input signal is divided into sub-band signals, and the sub-band signal in each channel is then manipulated independently. Here, the power envelope and carrier are represented as the two components to be restored in each channel.

4.2 Definitions

Let $x(t)$ and $y(t)$ be AC speech and associated BC speech. With the N-channel band-pass filterbank, we assumed that the signals could be represented as

$$x(t) = \sum_{n=1}^{N} x_n(t) = \sum_{n=1}^{N} e_{x_n}(t) \cdot c_{x_n}(t) \quad (2)$$

$$y(t) = \sum_{n=1}^{N} y_n(t) = \sum_{n=1}^{N} e_{y_n}(t) \cdot c_{y_n}(t) \quad (3)$$

Here, $x_n(t)$ and $y_n(t)$ are the sub-band signal components, $e_{x_n}(t)$ and $e_{y_n}(t)$ are the temporal envelopes, and $c_{x_n}(t)$ and $c_{y_n}(t)$ are the carriers in the n-th channel of the filterbank.

We used the Hilbert transform to decompose the signal in each channel into an envelope and a carrier. This method was based on the calculation of the instantaneous amplitude of the signal, using low-pass filtering as post-processing to remove the higher frequencies components in the envelopes.

$$e_{y_n}(t) = \mathrm{LPF}\left[|y_n(t) + j \cdot \mathrm{Hilbert}(y_n(t))|\right] \quad (4)$$

$$c_{y_n}(t) = \frac{y_n(t)}{e_{y_n}(t)} \quad (5)$$

In these equations, Hilbert$(\cdot)$ is the Hilbert transform. LPF$[\cdot]$ denotes low-pass filtering with a 20 Hz cut-off frequency to remove the high-pass envelope [7, 8]. This cut-off value (20 Hz) was chosen because the important modulation region, for speech perception [14] and speech recognition [15, 16], ranges from 1 to 16 Hz. $e_{x_n}(t)$ and $c_{x_n}(t)$ can also be calculated from $x(t)$ using the same method (Eqs. (4) and (5)).

4.3 Analysis

We analyzed the relationships between all pairs of speech signals (BC and AC speech) to design the MTF-based inverse transfer function, using the following:

(1) Correlation

$$\mathrm{Corr}(e_x(t)^2, e_y(t)^2) = \frac{\int_0^T \Delta e_x(t) \Delta e_y(t) dt}{\sqrt{\left\{\int_0^T \Delta e_x(t)^2 dt\right\}\left\{\int_0^T \Delta e_y(t)^2 dt\right\}}} \quad (6)$$

$$\Delta e_x(t) = e_x(t)^2 - \overline{e_x(t)^2}$$

$$\Delta e_y(t) = e_y(t)^2 - \overline{e_y(t)^2}$$

(2) SNR (dB)

$$\mathrm{SNR}(e_x(t)^2, e_y(t)^2) = 20 \log_{10} \frac{\int_0^T e_x(t)^2 dt}{\int_0^T (e_x(t)^2 - e_y(t)^2) dt} \quad (7)$$

(3) Complex modulation transfer function MTF

$$M(\omega) = \left| \frac{\int_0^\infty e_h(t)^2 \exp(-j\omega t) dt}{\int_0^\infty e_h(t)^2 dt} \right| \quad (8)$$

(4) Transfer functions via long-term FFT

$$h(t) = F^{-1}[H(\omega)] = F^{-1}\left[\frac{F[y(t)]}{F[x(t)]}\right] \quad (9)$$

$$e_h(t)^2 = F^{-1}[E_h(\omega)] = F^{-1}\left[\frac{F[e_y(t)^2]}{F[e_x(t)^2]}\right] \quad (10)$$

where $F[\cdot]$ is the long-term Fourier transform and $F^{-1}[\cdot]$ is the inverse of the long-term Fourier transform.

The signal was resampled at a sampling frequency of 16 kHz with 16 bits per sample. Then, to analyze the signal within 8 kHz, we used a constant-band N-channel filterbank, with 200 channels and a constant bandwidth of 40 Hz. Figure 3 shows the analyzed

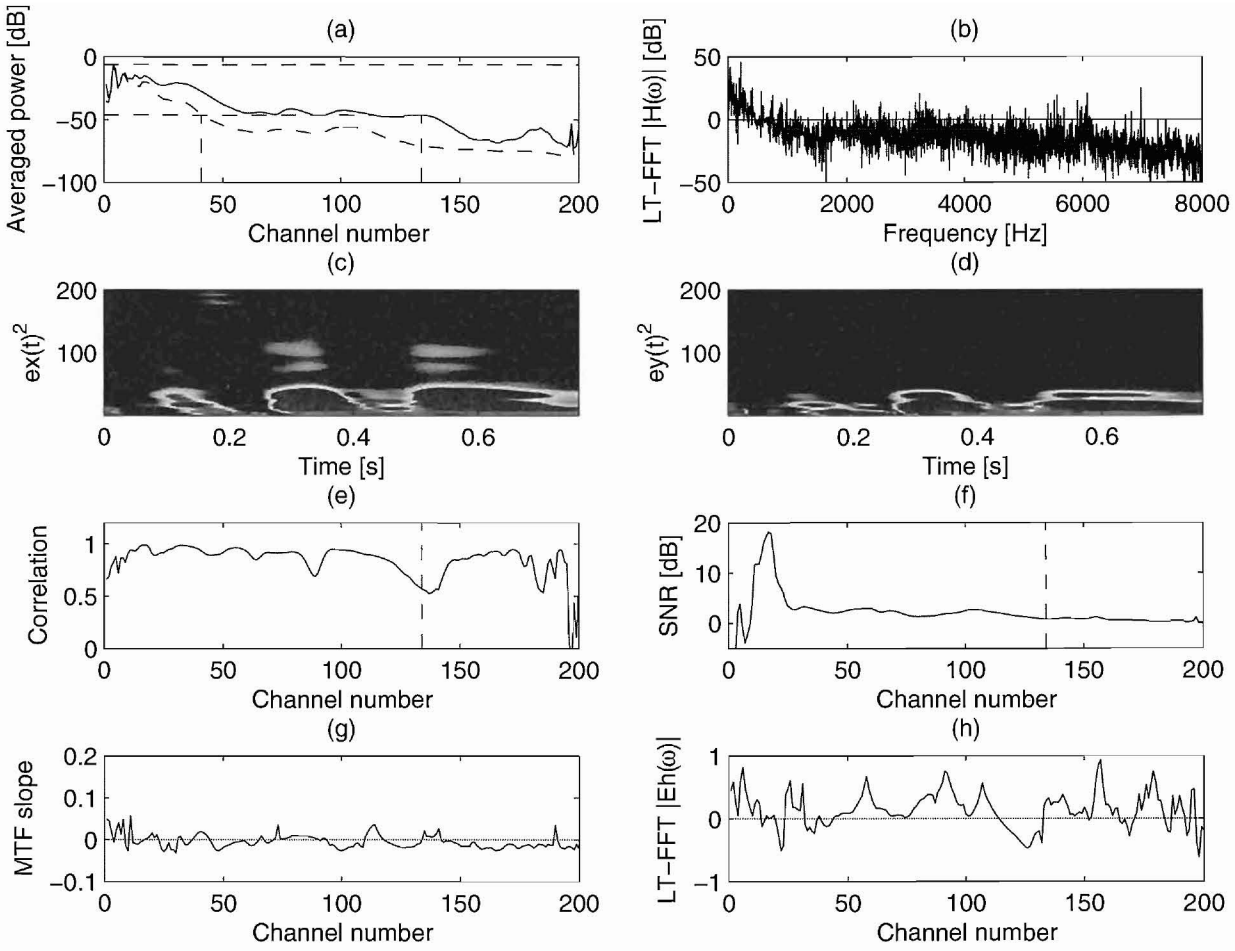

Fig. 3 Analysis results: (a) averaged $e_x(t)^2$ (solid line) and $e_y(t)^2$ (dashed line) in channels, (b) $|H(\omega)|$ via long-term FFT, (c) $e_x(t)^2$, (d) $e_y(t)^2$, (e) power envelope correlation $\mathrm{Corr}(e_x(t)^2, e_y(t)^2)$, (f) SNR $(e_x(t)^2, e_y(t)^2)$, (g) slope of the MTF $M(\omega)$, and (h) slope of $|E_h(\omega)|$ via long-term FFT in each channel

results for a pair of AC/BC male speech /asahan/, recorded at position 5 via Microphone B.

Figure 3(a) shows the averaged powers of AC (solid line) and BC speech (dashed line). The averaged powers of AC speech were reduced to less than 40 dB in the first 130 channels, while this was only in the first 40 channels with BC speech. Figure 3(b) showes the transfer function, $H(\omega)$, as a low-pass filter. Therefore, generally speaking, these figures indicated low-pass characteristics as their transfer functions. Figures 3(g) and 3(h) show the magnitude curves of MTF $M(\omega)$ and $|E_h(\omega)|$ in 200 channels that we analyzed. The signs of their values correspond to the characteristics of the transfer function in each channel. A positive value implies a high-pass filter and a negative value implies a low-pass filter. From Figs. 3(g) and 3(h), we know that the significant characteristics to restore BC speech in the power envelope can be interpreted as low-pass or hig-hpass filtering. In Figs. 3(e) and 3(f), the correlation between the power envelopes of AC and BC speech, $\mathrm{Corr}(e_x(t)^2, e_y(t)^2)$, is high within about 100 channels while the gain reduction, $\mathrm{SNR}(e_x(t)^2, e_y(t)^2)$, is small in most channels, and variants of these 100 channels. Therefore, the shapes of the power envelopes seem to be almost the same (Fig. 3(e)) and the difference is only in magnitude (Fig. 3(f)). Therefore, the relative reduction in the BC power envelopes with AC speech can be approximately interpreted as a reduction in constant gain within 100 channels (respective to 4 kHz).

These results were used to design the MTF-based inverse transfer function to construct the restoration model as descussed in the following.

4.4 Restoration method

From the above results, we predicted that the MTF-based transfer function with constant gain would be useful for restoring BC speech. We there-

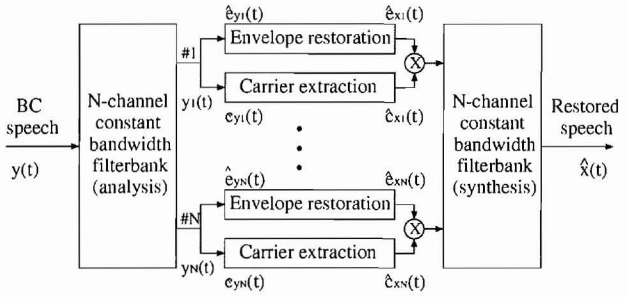

Fig. 4 Restoration of BC speech by MTF model

fore propose a method of restoration based on MTF with the power envelope gain compensation outlined in Fig. 4 [5]. The BC speech signal is decomposed into temporal envelopes and carriers in N-channels as Eq. (3), using Eqs. (4) and (5). Here, the power envelope could be obtained as the square of the temporal envelope. After processing in the BC power envelopes and the BC carriers, the restored sub-band signals in the channels are reconstructed into a restored speech signal, $\hat{x}(t)$, using synthesis processing.

Because the carriers are not important for speech intelligibility [11] during the carrier process, we used the BC carriers as the AC carriers as in the MTF-based model in Unoki *et al.* [5], i.e., $\hat{c}_{x_n} = c_{y_n}$.

A gain value is normally used as compensation for the BC power envelope during the power envelope process; this value is obtained from the average differences between BC and AC power envelopes in each channel. However, when the correlation is lower than 0.8 and the BC power envelope is low (≥ -20 dB), inverse filtering, $E_h^{-1}(z)$, as in some methods of MTF-based dereverberation [7, 8] is used to restore the BC power envelope as follows:

$$E_h^{-1}(z) = \frac{1}{a^2}\left\{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}\right\} \qquad (11)$$

where f_s is the sampling frequency (16 kHz), a is the gain factor, and T_R is the delay parameter. Parameters a and T_R can be estimated using the following algorithm:

$$\hat{T}_R = \max\left(0 \leq T_R \overset{\arg\min}{\leq} T_{R,\max}\left\{\int_0^T \min(\hat{e}_{x,T_R}(t)^2, 0)dt\right\}\right) \qquad (12)$$

$$\hat{a} = \sqrt{\frac{1}{T}\int_0^T \frac{e_x(t)^2}{e_y(t)^2}dt \Big/ \int_0^T \exp\left(\frac{-13.8t}{T_R}\right)dt} \qquad (13)$$

where $T_{R,\max}$ is the upper limited region of T_R, $\hat{e}_{x,T_R}(t)^2$ is the set of candidates for the power envelope restored as the function of T_R, and T is the signal duration of $y(t)$. These algorithms for parameters $\hat{T}_R$ and $\hat{a}$ originated in Unoki *et al.* [7, 8] and were modified for AC/BC speech. The algorithm for $\hat{a}$ was changed with a gain factor of $\sqrt{\frac{1}{T}\int_0^T \frac{e_x(t)^2}{e_y(t)^2}dt}$ that

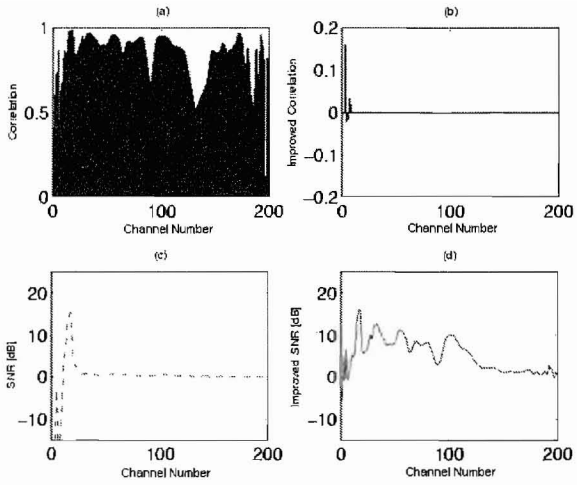

Fig. 5 Signal improvement in N-channel filterbank ($N = 200$, 40 Hz constant bandwidth): (a) correlation $\mathrm{Corr}(e_x(t)^2, e_y(t)^2)$, (b) $\mathrm{SNR}(e_x(t)^2, e_y(t)^2)$ (dB), (c) improved correlation, and (d) improved SNR (dB)

reflected the relation between AC and BC power envelopes. Here, the BC power envelope was restored using its gain control when parameter $\hat{T}_R$ was 0 (≥ -40 dB).

Figure 5 compares a restored signal and the BC speech signal. There are improvements in the correlation and SNR of the power envelope in each channel ($N = 200$). The left panel (a) in Fig. 5 shows the correlation between AC/BC power envelopes and (c) shows their SNR. Using the MTF-based model with power envelope gain compensation, the correlation was improved in lower frequency regions as shown in Fig. 5(b), and the SNR was also improved as seen in Fig. 5(d). These results prove the advantages of the MTF-based model.

5. LP-Based Model

5.1 Model concept

Linear prediction (LP) is one of the most powerful techniques for analyzing speech. It provides extremely accurate estimates of speech parameters such as fundamental frequency, formants, spectra, and vocal tract area functions, and it is relatively efficient in computation [6]. The LP-based model can be interpreted as a source-filter model; the LP residue is related to the source, which corresponds to the characteristics of the glottis and the LP coefficients are related to the filter, which corresponds to the characteristics of the vocal tract. The LP-based model originates from the idea that the information corresponding to the source (glottal) characteristics can be the

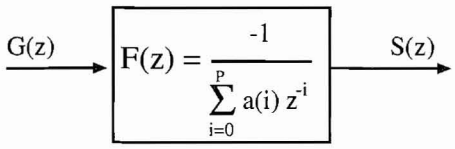

Fig. 6 LP as source-filter model

same for both BC and AC speech. Therefore, inverse filtering is primarily derived from the LP coefficients. Moreover, since the LP coefficients correspond to the vocal tract, inverse filtering based on LP coefficients could be adapted for the different characteristics of speakers and those of the syllables they pronounce.

5.2 Definition

When the LP order is sufficiently high, the all-pole model provides a good representation for almost all speech sounds, $s(n)$, as in Rabiner [6]

$$s(n) = \sum_{i=1}^{P} s(i)a(n-i) + g(n) \tag{14}$$

where P is the LP order, $a(i)$ is the i-th LP coefficient, and $g(n)$ is the LP residue of speech signal $s(n)$. We can rewrite Eq. (14) in the z-domain as

$$-G(z) = S(z)\sum_{i=0}^{P} a(i)z^{-i} \tag{15}$$

where $a(0) = -1$, $G(z)$ and $S(z)$ are the z transforms of $g(n)$ and $s(n)$. As we can see in Fig. 6, LP representation can be interpreted as a source-filter model. The LP residue, $G(z)$, is related to the source, which corresponds to the characteristics of the glottis. The LP coefficients are related to the filter, $F(z)$, which corresponds to the characteristics of the vocal tract.

Let $x(t)$ and $y(t)$ be the AC and its associated BC speech. The signals $x(n)$ and $y(n)$ are discrete signals of $x(t)$ and $y(t)$ with a sampling frequency of 16 kHz. Thus, the two signals, $x(n)$ and $y(n)$, are represented by the LP model in the z-domain as:

$$-G_x(z) = X(z)\sum_{i=0}^{P} a_x(i)z^{-i} \tag{16}$$

$$-G_y(z) = Y(z)\sum_{i=0}^{Q} a_y(i)z^{-i} \tag{17}$$

where $a_x(0) = -1$, $a_y(0) = -1$, $X(z)$ and $Y(z)$ are z transforms of $x(n)$ and $y(n)$, P and Q are LP orders, $a_x(i)$ and $a_y(i)$ are i-th LP coefficients, and $G_x(z)$ and $G_y(z)$ are z transforms of LP residues $g_x(n)$ and $g_y(n)$, respectively.

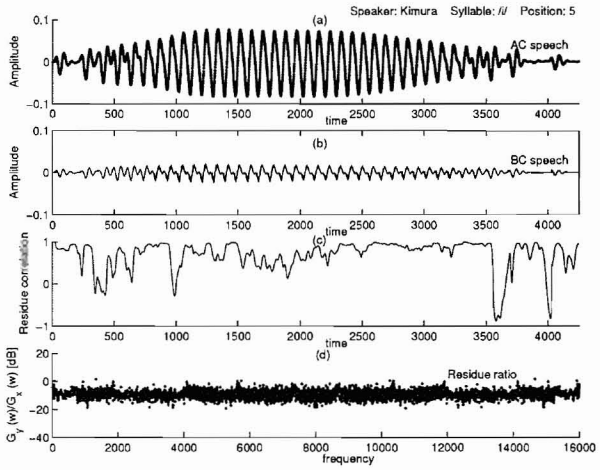

Fig. 7 Ratio of AC/BC residues: (a) AC speech, (b) BC speech, (c) residue correlation $\mathrm{Corr}(g_x(n), g_y(n))$, and (d)residue ratio $G_y(z)/G_x(z)$

Assuming that the mathematical description of $h(t)$ is an M-order FIR filter, the transfer function from $x(t)$ to $y(t)$ in the z-domain, $H(\omega)$, is represented as

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{i=0}^{M} h(i)z^{-i} \tag{18}$$

From Eqs. (16)-(18), we have

$$\sum_{i=0}^{M} h(i)z^{-i} = \frac{\sum_{i=0}^{P} a_x(i)z^{-i}}{\sum_{i=0}^{Q} a_y(i)z^{-i}} \cdot \frac{G_y(z)}{G_x(z)} \tag{19}$$

5.3 Analysis

Since the LP residues $g_x(t)$ and $g_y(t)$ are related to the source information (or glottal information) of $x(t)$ and $y(t)$, this kind of information may remain unchanged in both the AC and BC speech signals. To verify this supposition, we analyzed every pair of AC and BC speech signals in a Japanese syllable dataset. The technique for LP analysis as autocorrelation using the Levinson-Durbin recursion algorithm. There is a typical example of the results of analysis in Fig. 7. The AC and BC vowel /i/ signals are in Figs. 7(a) and 7(b), respectively. Figure 7(c) indicates that the correlation between $g_x(n)$ and $g_y(n)$ is very high. Each correlation value here is associated with a pair of 4-millisecond AC/BC speech frames. Figure 7(d) shows that the ratio of the LP residues in the frequency domain, $G_x(z)/G_y(z)$, is almost constant. Although this ratio is not stable at any frequency, we can approxi-

mately represent it as

$$\frac{G_y(z)}{G_x(z)} = k \tag{20}$$

where k is a constant factor.

5.4 Restoration method

From Eqs. (19) and (20), we can rewrite this ratio as

$$\left(\sum_{i=0}^{M} h(i)z^{-i}\right) \cdot \left(\sum_{i=0}^{Q} a_y(i)z^{-i}\right) = k\sum_{i=0}^{P} a_x(i)z^{-i} \tag{21}$$

Deriving the zero-th variable of both sides in Eq. (21), we obtain

$$h(0) = k = \frac{G_y(z)}{G_x(z)} \tag{22}$$

Figure 8 outlines a typical conversion from AC speech to BC speech with transfer function $H(z)$. The inverse filter, $H^{-1}(z)$, can be found as the inverse function of $H(z)$ and used to straightforwardly restore BC speech to AC speech. All equations in the figure have been implied from Eqs. (16),(17), and (22). From these, we can obtain the equation for $H^{-1}(z)$ simply as

$$H^{-1}(z) = \frac{1}{h(0)} \cdot \frac{\sum_{i=0}^{Q} a_y(i)z^{-i}}{\sum_{i=0}^{P} a_x(i)z^{-i}} \tag{23}$$

We should obtain the restored speech from observed BC speech with the inverse transfer function, $H^{-1}(z)$, which depends on the LP coefficients and residue ratio of the AC and BC speech signals. Equation (23) gives us a different way to estimate this transfer function. Inverse transfer function $H^{-1}(z)$ is decomposed into two parts. In the first part, the constant value, $h(0)$, can be chosen manually and used to control the magnitude of restored speech. The second part primarily depends on the LP coefficients of signals. Therefore, in the LP-based model, the relationship between the LP coefficients of AC and BC speech signals is essential to restore BC speech. Here, we chose $h(0) = 1$ and set the LP orders at $P = Q = 20$.

6. Evaluation

This section discusses the feasibility of the models to restore BC speech signals. The main aim of our evaluation was to determine which model would be the most useful with regard to sound quality to achieve two different purposes: speech intelligibility for human hearing systems and robustness for ASR

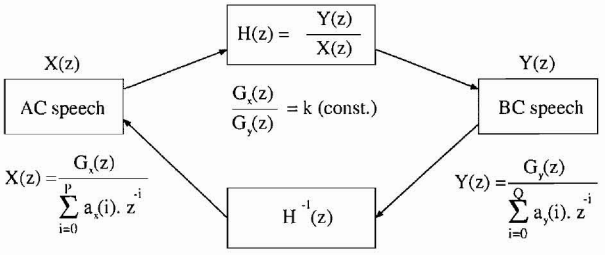

Fig. 8 Transfer function of LP-based model

systems. Using a number of different measurements, we evaluated two previous models and our two proposed models. The two previous were (1) the cross spectrum signal (CrossSig) and the (2) the long-term FFT (LTFSig) models. Ours were (3) the MTF-based (MTFGain) and (4) the LP-based (LPSig) models.

Log-spectrum distortion (LSD) and the mean opinion score (MOS) were used to evaluate the improvements in intelligibility, and the LP coefficient distance (LCD) and mel-frequency cepstral coefficient distance (MCD) were used to evaluate the improvements in the cepstral distance of restored speech signals. The Japanese vowel dataset from the AC/BC database was used to evaluate improvements with the objective measurements (LSD, LCD, and MCD), and ten words that were chosen randomly from the Japanese word dataset were used to evaluate improvements with the subjective measurements (MOS).

6.1 Objective evaluations

We used LSD, LCD, and MCD for the Japanese syllable dataset to objectively evaluate the four methods. These three measurements were computed as follows:

$$\text{LSD} = \sqrt{\frac{1}{W}\sum_{\omega}^{W}\left[20\log_{10}\left(\frac{|S(\omega)|}{|\hat{S}(\omega)|}\right)\right]^2} \tag{24}$$

$$\text{LCD} = \sqrt{\frac{1}{P}\sum_{i=1}^{P}\left(a_x(i) - a_y(i)\right)^2} \tag{25}$$

$$\text{MCD} = \sum_{i=0}^{12}\left(c_{x,i} - c_{y,i}\right)^2 \tag{26}$$

where W is the upper frequency (8 kHz in this case), and $S(\omega)$ and $\hat{S}(\omega)$ are the amplitude spectra obtained by 1024-point FFT calculation of 25-millisecond frames. The time these frames overlapped was 15 ms. Here, $a_x(i)$ and $a_y(i)$ are the i-th LP coefficients of signals with the LP order being set at $P = 20$, and $c_{x,i}$ and $c_{y,i}$ are the i-th MFCC of the signals.

After measuring the distances between the clean AC speech signal and the other signals, i.e., the ob-

Table 2 Average LSD improvements in restored speech

Method	Measurement position				
	1	2	3	4	5
CrossSig	1.14	5.00	2.14	3.53	5.54
LTFSig	**3.18**	4.64	**4.58**	3.95	6.27
MTFGain	1.94	4.04	2.60	1.68	6.55
LPSig	3.02	**5.80**	4.29	**5.00**	**7.71**

Table 3 Average LCD improvements in restored speech

Method	Measurement position				
	1	2	3	4	5
CrossSig	0.60	1.09	0.56	0.89	0.57
LTFSig	0.62	1.24	**1.15**	**1.07**	0.78
MTFGain	0.61	0.18	0.08	0.25	0.31
LPSig	**0.73**	**1.26**	0.91	0.92	**0.82**

Table 4 Average MCD improvements in restored speech

Method	Measurement position				
	1	2	3	4	5
CrossSig	6.13	7.00	7.74	5.42	1.71
LTFSig	4.91	10.13	13.84	7.98	7.56
MTFGain	5.72	4.22	7.22	5.53	4.91
LPSig	**9.69**	**11.66**	**13.95**	**10.09**	**8.33**

Table 5 Results of MOS test

BC speech	Cross -Sig	LTF -Sig	MTF -Gain	LP -Sig	AC speech
2.44	1.72	2.45	2.68	2.91	4.33

served BC speech and the restored speech signals, we evaluated the improvements in the restored speech in comparison with BC speech. Tables 2, 3, and 4 list the average improvements in the three objective measurements. The LP-based model is generally the best for all the measurements. Although the MTF-based model does not have advantages in the objective measurements, it does yield better results than the previous models with the LSD measurement at measurement in Position 5. Figure 9 shows typical LSD curves in this position.

6.2 Subjective evaluation

We carried out MOS (mean opinion score) tests using the four methods for the subjective evaluation. We conducted these with five subjects who had normal hearing. The MOS tests were used to measure the sound quality of restored speech using the four methods in five evaluations graded as perceived by the subjects. The levels rated were: bad (1), poor (2), fair (3), good (4), and excellent (5). The speech signals in these tests were ten words that had been

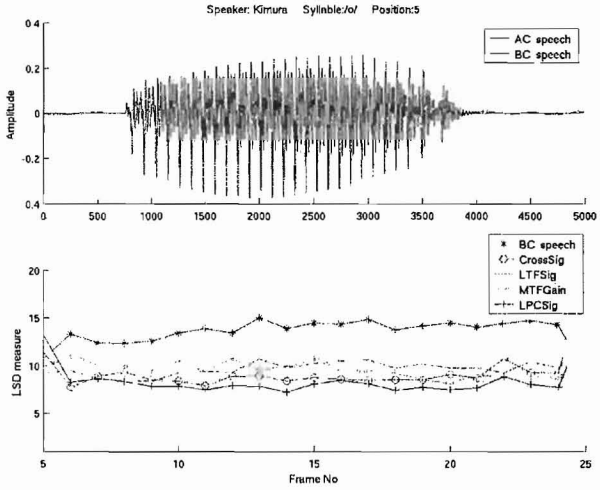

Fig. 9 LSD between AC and restored speech signals

randomly chosen from the Japanese word dataset.

Table 5 lists the mean scores for the subjective ratings. The LP-based model is the best for restoring BC speech to AC speech. This is followed by the MTF-based model. These subjective results also prove that the previous models were not as good. The improvements with the LTFSig model are almost zero, and even minus with the CrossSig model.

6.3 Discussion

As previously mentioned, we used LSD and MOS to evaluate improvements in speech intelligibility, which is useful in human hearing systems. LCD and MCD were used to evaluate the cepstral distances, which are significant for ASR systems. With LCD and MCD measurements, the MTF-based model did not seem to have as good ASR robust-features. Although the LSD measurements demonstrated that the MTF-based model significantly improved intelligibility, the subjective MOS measurements also revealed its advantages in comparison with the two previous models. The objective measurements were better for evaluating the models for ASR systems, and the listening tests were better for evaluating speech intelligibility. Overall, the MTF-based model was best for improving intelligibility.

All the objective measurements (LSD, LCD, and MCD) and the subjective measurements (MOS) revealed that the proposed models, i.e., MTF-based and LP-based, were better at improving voice-quality than the other previous methods. In particular, the LP-based model was the best for both human hearing and ASR systems.

The proposed models still currently need to use some information from AC speech to restore the observed BC speech. The gain values of the power en-

velope in the MTF-based model, and the LP coefficients of AC speech in the LP-based model, are essential in constructing inverse filtering. This also means that there are only a few parameters (gain, LP coefficients) that affect the ability of the proposed models in restoration. These parameters should depend on the characteristics of pronounced sounds such as vowels and consonants from each specific position of measurement. By investigating the variances in these parameters along with BC speech sounds, we should be able to find practical algorithms to determine them automatically without AC speech.

7. Conclusion

We constructed a large-scale AC/BC speech database (5 measurement positions, 10 speakers, and 145 stimuli for both AC and BC speech) to investigate the significant characteristics in the relationship between AC and BC speech signals. By analyzing all AC/BC datasets in this database, we found that the gain of the power envelope is approximately constant in about 100 channels, corresponding to 4 kHz. We also found a constant ratio of LP residues of AC and BC speech signals with the LP method. These characteristics seem to be significant in restoring BC speech. We then proposed two models according to these characteristics, i.e., MTF-based and LP-based models. Both models worked with the same concept, but there were differences in their processing domains. The MTF-based model decomposed a signal into subband signals and then separately manipulated temporal envelopes and carriers in each channel, while the LP-based model manipulated the LP residue and LP coefficients in each frame. These differences can result in different restoration goals in improving speech intelligibility and features that are robust in ASR.

We evaluated both these models and demonstrated their advantages by comparing them with two other methods (CrossSig and LTFSig). As a result, we found that both the MTF-based and LP-based models were better than the other two methods for improving voice-quality. The MTF-based model, in particular, efficiently restore speech intelligibility which is useful for human hearing systems. The LP-based model efficiently improved voice-quality. We therefore verified both the proposed methods based on our concept could adequately restore BC speech to improve not only its intelligibility but also the performance of ASR systems.

These results were obtained as the first steps toward investigating the possibility of restoring BC speech. We thus focused on analyzing the significant relationship between AC and BC speech signals, and the feasibility of models to restore BC speech signals. The proposed models still currently need to use AC speech to determine the coefficients of MTF-based and LP-based inverse filtering. The gain values of power envelope inverse filtering in the MTF-based model were determined by the ratio of the AC/BC envelopes. The inverse transfer function in the LP-based model was determined using the LP coefficients of AC/BC speech signals.

As the next step toward developing blind restoration of BC speech in future work, we intend to investigate the variances in the model parameters in association with BC speech signals before finding practical algorithms to automatically calibrate these parameters only from the characteristics of BC speech signals. A different development, i.e., a hybrid model based on the same concept (of both MTF-based and LP-based models) may be able to be proposed, which would restore temporal-spectral information in both the time and frequency domains. It would have superior performance for both human hearing and ASR systems.

Acknowledgments

This work was supported by a Grant-in-Aid for Science Research (No. 17650048) and a scheme for the "21st Century COE Program" in Special Coordination Funds for promoting Science and Technology made available by the Ministry of Education, Culture, Sports, Science, and Technology.

References

[1] M. Kumashita, T. Shimamura and J. Suzuki: Property of voice recorded by bone-conduction microphone, Proc. 1996 spring meeting on Acoust. Soc. Jpn, 2-Q-3, pp. 269-270, March 1996 (in Japanese).

[2] S. Ishimitsu, H. Kitakaze, Y. Tsuchibushi, H. Yanagawa and M. Fukushima: A noise-robust speech recognition systemmaking use of body-conducted signals. Acoust. Sci. & Tech., Vol. 25, pp. 166-169, 2004.

[3] T. Tomikura and T. Shimamura: A study on improving the quality of voice of bone conduction, Proc. 2003 spring meeting on Acoust. Soc. Jpn,2-Q-14, pp. 401-402, 2003 (in Japanese).

[4] T. Tamiya and T. Shimamura: Reconstruct filter design for bone-conducted speech, Proc. ICSLP2004, II, pp. 1085-1088, 2004.

[5] M. Unoki, K. Kimura and M. Akagi: A study on a bone-conducted speech restoration with the modulation transfer function, Trans. Tech. Comm. Psycho. Phyciol. Acoust. ASJ, Vol. 35, No. 3, pp. 191-196, H-2005-33, April 2005 (in Japanese).

[6] L. R. Rabiner: Digital Processing of Speech Signals, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[7] M. Unoki, M. Furukawa, K. Sakata and M. Akagi: An improved method based on the MTF concept for restoring the power envelope from a reverberant signal, Acoust. Sci. & Tech., Vol. 25, No. 4, pp. 232-242, 2004.

[8] M. Unoki, K. Sakata, M. Furukawa and M. Akagi: A speech dereverberation method based on the MTF concept in power envelope restoration, Acoust. Sci. &Tech., Vol. 25, No. 4, pp. 243-254, 2004.

[9] Database for speech intelligibility testing using Japanese word lists, NTT-AT, March 2003.

[10] T. Houtgast, H. J. M. Steenken and R. Plomp: Predicting speech intelligibility in rooms from the modulation transfer function, Part I General Room Acoustics, Acustica, Vol. 46, pp. 60-72, 1980.

[11] R. Drullman: Temporal envelope and fine structure cues for speech intelligibility, J. Acoust. Soc. Am, Vol. 97, pp. 585-592, 1995.

[12] T. Houtgast and H. J. M. Steenken: The modulation transfer function in room acoustics as a predictor of speech intelligibility, Acustica, Vol. 28, pp. 66-73, 1973.

[13] T. Houtgast and H. J. M. Steenken: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria, J. Acoust. Soc. Am., Vol. 77, pp. 1069-1077, 1985.

[14] T. Arai, M. Pavel, H. Hermansky and C. Avendano: Syllable intelligibility for temporally filtered LPC cepstral trajectories, J. Acoust. Soc. Am., Vol. 105, pp. 2783-2791, 1999.

[15] N. Kanedera, T. Arai, H. Hermansky and M. Pavel: On the importance of various modulation frequencies for speech recognition, Proc. EuroSpeech 97, pp. 1079-1082, 1997.

[16] N. Kanedera, T. Arai and T. Funada: Robust automatic speech recognition emphasizing important modulation spectrum, IEICE Trans. D-II, J84-D-II, pp. 1261-1269, 2001.

Thang Tat Vu received his B.E. and M.E. in electronic & telecommunication engineering from the Hanoi University of Technology (HUT), in 2002 and 2004. He was a member of the Speech Processing Group at the Institute of Information Technology (IOIT) of the Vietnamese Academy of Science and Technology (VAST) from 2002. He has been a Ph.D. candidate at the School of Information Science of the Japan Advanced Institute of Science and Technology (JAIST) since 2005.

Kenji Kimura received his B.E. in electrical engineering from Doshisha University in 2003, and his M.S. from the Japan Advanced Institute of Science and Technology (JAIST) in 2005. He has been with the Tokyo Electric Power Company since 2005.

Masashi Unoki received his M.S. and Ph.D. (Information Science) from the Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999. His main research interests are in auditorymotivated signal processing and the modeling of auditory systems. He was a JSPS research fellow from 1998 to 2001. He was associated with the ATR Human Information Processing Laboratories as a visiting researcher from 1999-2000, and he was a visiting research associate at the CNBH in the Department of Physiology at the University of Cambridge from 2000 to 2001. He has been on the faculty of the School of Information Science at JAIST since 2001 and is now an Associate Professor. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Acoustical Society of America (ASA). He is also a member of the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). Dr. Unoki received the Sato Prize from the ASJ in 1999 for an Outstanding Paper and the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation in 2005.

Masato Akagi received his B.E. from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. Eng. from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST) and is now a Professor. His research interests include speech perception, the modeling of speech perception mechanisms of human beings, and signal processing of speech. During 1998, he was associated with the Research Laboratories of Electronics, MIT as a visiting researcher, and in 1993, he studied at the Institute of Phonetics Science, Univ. of Amsterdam. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronic Engineering (IEEE), the Acoustical Society of America (ASA), and the International Speech Communication Association (ISCA). Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, and the Sato Prize for Outstanding Papers from the ASJ in 1998 and 2005.

(Received May 30, 2006; revised August 14, 2006)