LP-Based Method of Blind Restoration to Improve Intelligibility of Bone-Conducted Speech

Thang Tat Vu, Masashi Unoki, and Masato Akagi
5 October 2007
IS-RR-2007-011

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, JAPAN
vu-thang@jaist.ac.jp, unoki@jaist.ac.jp, akagi@jaist.ac.jp

ISSN 0918-7553

PAPER

LP-BASED METHOD OF BLIND RESTORATION TO IMPROVE INTELLIGIBILITY OF BONE-CONDUCTED SPEECH

Thang TAT VU[†a)], **Massashi UNOKI**[†b)], ***and*** **Masato AKAGI**[†c)],

SUMMARY Bone-conducted (BC) speech can be used instead of air-conducted (AC) speech in an extremely noisy environment. However, its intelligibility is degraded when transmitted through bone-conduction. Therefore, voice quality and the intelligibility of BC speech need to be blindly improved in actual communication through speech and this is a challenging new topic in the field of speech signal processing. We proposed a linear prediction (LP) based model to restore BC speech to improve voice quality in a previous study. While other methods such as Long-term Fourier transform need to use numerous AC speech parameters to restore BC speech, the model we proposed demonstrated the expressed ability of blindly restoring BC speech by predicting AC-LP coefficients from BC-LP coefficients. We improved the previous model by (1) extending long-term processing to frame-basis processing, (2) using line spectral frequency (LSF) coefficients on an LP representation, and (3) using a recurrent neural network for predicting parameters. We evaluated the improved model in comparison with others to find out whether it could adequately improve voice quality and the intelligibility of BC speech, using objective measures (i.e., LSD, MCD, and LCD) and carrying out a subjective measure — a Japanese-word intelligibility test (JWIT). The experimental results proved significant improvements to our newly proposed models (LSF and LSF-SRN). The LSF model demonstrated it had significant capabilities for improving BC speech, i.e., both voice quality and intelligibility of speech. Our proposed model, LSF-SRN, demonstrated an expressed capability for improving the intelligibility of BC speech even when using blind restoration.

key words: *Speech intelligibility, Bone-conducted speech, Simple recurrent network, Blind restoration.*

1. Introduction

It is very difficult for automatic speech recognition (ASR) systems or people to communicate through speech in extremely noisy environments. This is because of the poor sound quality and intelligibility of speech due to the influence the transmission environments have on speech features.

There have been many different complex models and/or algorithms that have been used to cancel or reduce the effects of interfering noise [1]. These approaches have only been efficient at low- and medium-noise levels and have been ineffective when these have been too high.

Another possible solution has been to use a special microphone to record the speech signals transmitted through the speaker's head and face [2], [3]. This recorded signal is referred to as "bone-conducted (BC) speech". Its stability against interfering noise from noisy environments makes BC speech more advantageous than noisy air-conducted (AC) speech.

Although BC speech is not affected by external noise while AC speech is, there is a drawback to using BC speech in that the signal is complexly attenuated when it is transmitted through bone conduction. BC speech is generally attenuated stronger at higher frequencies and the attenuation seems to be low-pass filtering with a cut-off frequency of about 1 kHz [2], [3]. The characteristics of BC vary for different pick-up points (BC microphones) and the distribution of frequency components varies with syllables and speakers who pronounce syllables differently [2], [3]. This causes the attenuation changed on different pick-up points, speakers, and pronounced syllables

The attenuation causes the voice quality to be degraded in BC speech, which means both the intelligibility of speech in human-hearing systems and the robustness for ASR systems. If the voice quality of BC speech can be improved, the restored signals can be used in speech applications in noisy environments with greater efficiency instead of using noisy AC speech. There are several studies on BC speech for applications such as human-hearing aids and machine-hearing systems but the results are still limited. A Gaussian Mixture Model (GMM)-based voice conversion model was applied to restore body-transmitted speech, which is like BC speech [4]. However, due to the difficulty of dealing with F0 features that might cause synthesis problems, this approach has only been applied to unvoiced speech such as whispered speech. BC speech has been used as an additional source to noisy AC speech and helps to reduce external noise [5], [6] in other approaches such as when air-and-bone conductive microphones are used.

The purpose of our approach is to restore BC speech to enable restored speech to be applied directly such as to human-hearing systems and to the front-end of ASR systems. Since it is very difficult to blindly re-

[†]The authors are from the School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan
a) E-mail: vu-thang@jaist.ac.jp
b) E-mail: unoki@jaist.ac.jp
c) E-mail: akagi@jaist.ac.jp

store the signals of BC speech and improve voice quality and the intelligibility of speech without using any other information, this topic is extremely challenging in the field of speech-signal processing.

The 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 form of high-pass filtering that is independent of speakers, pronounced syllables, or pick-up points. Although there are various methods of deriving inverse filtering such as the cross-spectrum [7], short-term Fourier transform [8] and the long-term Fourier transform [9], [10], these yield restored signals with artifacts such as musical noise and echoes, so that there are only slight improvements in voice quality [7]–[10].

We proposed MTF-based and LP-based models in our previous papers [12]–[16], which overcame the drawbacks of previous methods and yielded a restored signals that had better intelligibility. Both models, in fact, was based on the same concept for restoring the observed BC signal as in their representations of source and filter information. The MTF-based model tries to restore the temporal power envelope in each channel, while the LP-based model tries to restore the spectrum envelope. Thus, the difference here is just the processing domain, the first processing in the time domain, and the other processing in the frequency domain.

The MTF-based model compensated for the reduced values of temporal power envelopes in the channels of the filterbank model. It overcame the drawbacks with previous methods and yielded a restored signal with enhanced voice quality [12]. Although its aim was to restore BC for human-hearing, no consideration was given to improving ASR systems. The LP-based model originates from the idea that the LP residue information corresponding to the source (glottal) characteristics is the same for both BC and AC speech signals. Therefore, adaptive inverse filtering was primarily derived from the LP coefficients, which are related to the filter (vocal tract) characteristics. This model showed its ability to yield restored signals that were not only more intelligible to human-hearing systems in experiments but it also enabled ASR systems to achieve better recognition [12]–[15].

Information on AC speech is needed to construct the inverse filtering in restoration models [7]–[14] and this is a serious drawback in practice when we have no information on AC speech. Inverse filtering using cross-spectrum and Fourier transform methods [7], [9], [10] depends on the AC spectrum. The gain values for the power envelope in the MTF-based model [12] and the LP coefficients of AC speech in the LP-based one [12]–[14] are essential to construct inverse filtering. Averaged gains or averaged LP coefficients can be chosen for these methods using averaged filtering [8]. However, the model that is achieved will be difficult to adapt to BC speech signal.

A recurrent neural network was shown to be effective in designing an inverse filter from BC to AC speech in one study [11] to adapt an inverse filter with a BC speech signal. We proposed an LP-based model with the ability of blind restoration in our previous study by predicting various parameters [15], [16] from the fact that model only depends on a few unknown parameters, i.e., the LP coefficients of AC speech (AC-LP coefficients). Machine learning methods were applied to predicting AC-LP coefficients from BC-LP coefficients. The results from this study revealed that the existing relationship between the LP coefficients of AC and BC speech signals is helpful for blindly restore BC speech [15].

Although reasonable results were obtained to improve the voice quality and intelligibility of BC speech, the LP-based model [15] suffered from some serious limitations. The LP coefficients were not stable or suitable to enable prediction with statistical models due to the different roles played by LP coefficients and their relatively large dynamic ranges. Even small prediction errors of AC-LP coefficients could easily cause problems with filter instability [16], [20]. Also, inverse filtering was only determined to remain unchanged for an entire BC speech signal as in long-term processing.

We improved the model of LP-based blind restoration by (1) extending long-term processing to frame-based processing, (2) using LSF coefficients on LP representations, and (3) predicting LSF parameters on a frame-by-frame basis via a recurrent neural network. Since LSF coefficients play the same role in the presentation of the spectrum envelope and their values are limited within a range $(0,\ \pi)$, these coefficients could help alleviate the limitations with LP coefficients in prediction using statistical methods. The processes of restoration on a frame-by-frame basis could also be adapted to inverse filtering in real time. Since a speech signal contains a series of speech frames, the restoration of neighboring frames should be related; a simple recurrent network (Elman network) was applied to predict BC-LSF coefficients to complete the blind restoration system.

This paper is organized as follows: Section 2 describes the AC/BC speech database that we constructed. Section 3 describes the framework and the algorithm for LP-based BC speech restoration. Section 4 explains how we implemented the proposed model as a blind restoration model. Section 5 discusses how we evaluated the models using objective and subjective evaluations. Section 6 draws conclusions and gives perspectives regarding further work.

2. AC/BC speech database

We assumed that there were existing relationships be-

Table 1 List of equipments.

Measurement site	Soundproof room
Number of pick-up points	5
Number of speakers	10
Recorder	MARANZ, PMD671
Coding method	PCM
Sampling frequency	48 kHz
Sample size	16 bits
Number of channels	2 (Left:AC, Right:BC)
Mic. A for AC speech	SONY, C536P
Mic. power supply A	SONY, AC148F
Mic. B for BC speech	TEMCO, HG-17
Mic. C for BC speech	TEMCO, SK-1
Mic. amp. B & C	Handmade

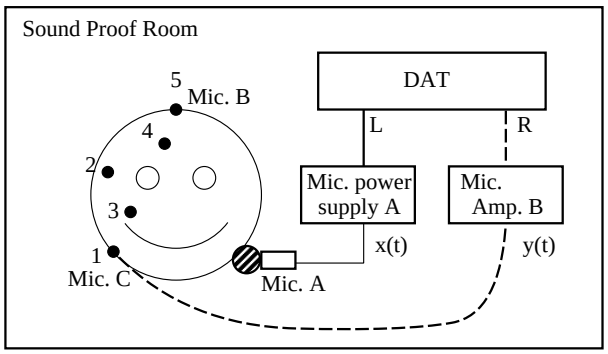


Fig. 1 Environment for recording AC/BC speech.

tween AC and BC speech that were significant in restoring BC speech. Therefore, a database was essential for analyzing the relationships and differences between BC speech and clean AC speech signals before any models were 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 (2 channels).

Figure 1 and Table 1 show the environment and equipment used to construct this database. The BC speech was collected at five different pick-up points on the head and face: the (1) mandibular angle, (2) temple, (3) philtrum, (4) forehead, and (5) calvaria. These points were chosen among other several pick-up points since their pick-up signals was clearer and better quality than that of the others [19]. One pick-up point was associated with one pair of clean AC and BC speech. Microphone B was only used at point 5 and microphone C was used at the other pick-up point. Ten speakers (five males and five females) participated in the recording of speech pronouncing 100 Japanese words and all 101 Japanese syllables.

The database was divided into two parts. The first was (i) a Japanese word dataset of 100 Japanese words selected from Japanese word lists by NTT-AT (2003) [17]. With 10 speakers, 100 words, and 5 pick-up points, there were 5000 pairs of wave files. The second part was a (ii) Japanese syllable dataset of all 101 Japanese syllables. With 10 speakers, 101 syllables, and 5 pick-up points, there were 5050 pairs of wave files. The selected words in the Japanese word dataset were chosen by the degree of familiarity with the NTT database [17]. Word familiarity in the NTT database was ranked from 1 (low familiarity) to 7 (high familiarity) for all 80,000 entry words and subtitles in the Shinmeikai Japanese Dictionary (Fourth Ed.). The population of words was then divided into four groups in different familiarity ranges [18]. There were R1(1.0,2.5) – low familiarity range, R2(2.5,4.0) – medium low familiarity range, R3(4.0,5.5) – medium high familiarity range, and R4(5.5,7.0) – high familiarity range. Since there is complementary relationship between the familiarity of a word and its intelligibility [17], [18], it is often difficult even for a Japanese to recognize what a low familiarity word is in clean condition. We selected 100 words with 25 words chosen for each familiarity range from the Japanese word dataset.

3. LP-based BC Speech Restoration

3.1 Signal Restoration Diagram Based on LP

LP is one of the most powerful techniques for analyzing speech. The all-pole model provides a good representation of almost all speech sounds when the order of LP is sufficiently high [20]. Let $x(t)$ and $y(t)$ be 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 [12]–[16] as

$$-G_x(z) = X(z)\sum_{i=0}^{P} a_x(i)z^{-i}, \qquad a_x(0) = -1, \quad (1)$$

$$-G_y(z) = Y(z)\sum_{i=0}^{Q} a_y(i)z^{-i}, \qquad a_y(0) = -1, \quad (2)$$

where $X(z)$ and $Y(z)$ are the z-transforms of $x(n)$ and $y(n)$, P and Q are LP orders, and $a_x(i)$ and $a_y(i)$ are i-th LP coefficients. Here, $G_x(z)$ is the z-transforms of the LP residues of $g_x(n)$ and $G_y(z)$ is that of $g_y(n)$.

Since the LP residues, $g_x(n)$ and $g_y(n)$, are related to the source information (glottal information) of $x(n)$ and $y(n)$, this kind of information may remain unchanged in both AC and BC speech signals. Figure 2 has a typical example of the relation between AC and BC speech signals. The AC and BC vowel /i/ signals, which have been recorded simultaneously, are in Figs. 2(a) and 2(b). Figure 2(c) shows that the correlation between $g_x(n)$ and $g_y(n)$ is very high. Each correlation value here is associated with an AC/BC speech pair at 4−ms frames. Figure 2(d) shows that the ratio of LP residues in the frequency domain is almost constant. These facts suggest that the AC and BC residues are

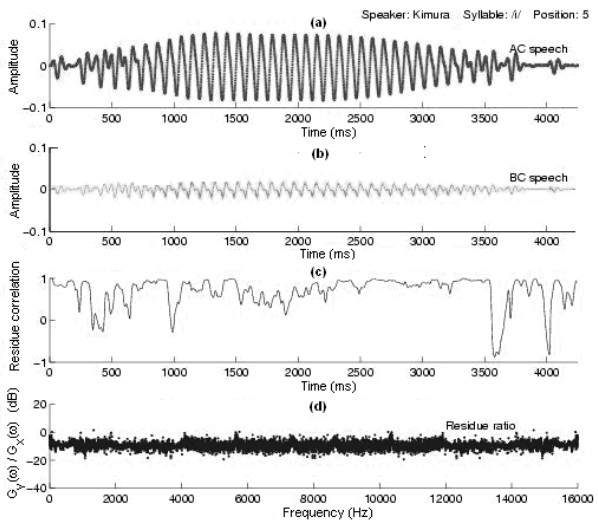


Fig. 2 Ratio of AC-BC residues: (a)AC speech, (b) BC speech, (c)residue correlation Corr($g_x(n)$, $g_y(n)$), and (d)residue ratio $G_y(z)/G_x(z)$.

almost the same except for magnitude. We can represent this approximately as a constant factor, k, as

$$G_x(z)/G_y(z) = k. \tag{3}$$

Let us assume that the mathematical description of transfer function $h(n)$ from $x(n)$ to $y(n)$ is an M-order FIR filter. In the z-domain, it is represented as

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{i=0}^{M} h(i) z^{-i}. \tag{4}$$

Figure 3 outlines a typical conversion from AC to BC speech with transfer function $H(z)$. The inverse filter $H^{-1}(z)$ can be found as the inverse of $H(z)$ and used to restore BC to AC speech in a straightforward way. All equations in the figure have been derived from Eqs. (1)-(3). We can obtain the equation for $H^{-1}(z)$ simply from these as

$$H^{-1}(z) = \frac{1}{H(z)} = k \cdot \frac{\sum_{i=0}^{Q} a_y(i) z^{-i}}{\sum_{i=0}^{P} a_x(i) z^{-i}}. \tag{5}$$

We should obtain restored speech from observed BC speech with inverse filtering $H^{-1}(z)$ which can be decomposed into two parts. In the first, the constant value, k, 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, the relation between the LP coefficients of AC and BC speech signals in the LP-based model is essential to restore BC speech [12]–[14]. Moreover, these LP coefficients in the LP-based model have to be predicted from

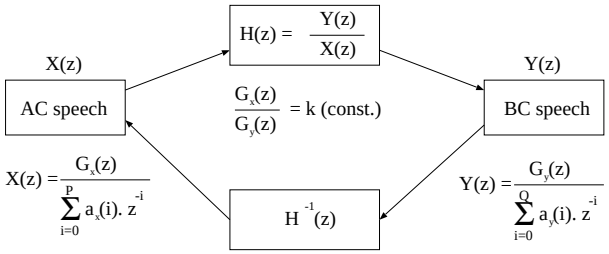


Fig. 3 Transfer function of LP-based model.

observed BC speech. Although LP coefficients could be predicted with the previous LP-based model with some good result [15], they were inappropriate parameters for statistical models of prediction because they played different roles and had a relatively wide dynamic range. LSF coefficients are thus used as more appropriate parameters in this paper.

3.2 LSF Representation

While LP coefficients are known to be inappropriate for statistical models because of their different roles and their relatively large dynamic range, LSFs would be a better choice. LSFs are an alternative representation of LP coefficients with the same spectral information, such as reflection coefficients and log area ratios, but reduce the above problems with LP coefficients. LSF parameters have both a well-behaved dynamic range and filter stability, and can be used to encode LP spectral information more efficiently than any other parameters [20]. The almost equivalent roles of LSF coefficients, on the other hand, would be suitable for statistical models.

Let $A(z)$ be a general LP filter on an LP representation. The LSF coefficients, ϕ and θ, can be derived from a symmetric polynomial and an anti-symmetric polynomial, $U(z)$ and $V(z)$, as the phase of conjugated zeros.

$$A(z) = \sum_{i=0}^{P} a(i) z^{-i}, \qquad a(0) = -1 \tag{6}$$

$$U(z) = A(z) + z^{-(P+1)} A(z^{-1}), \tag{7}$$

$$V(z) = A(z) - z^{-(P+1)} A(z^{-1}), \tag{8}$$

where $U(z)$ has a root of $z = -1$ and $V(z)$ has a root of $z = 1$. $U(z)$ and $V(z)$ have conjugated zeros that can be expressed as $e^{\pm j\phi}$ and $e^{\pm j\theta}$. Phases ϕ_i and θ_i of the conjugated zeros of $U(z)$ and $V(z)$ are interlaced with each other in the interval $(0, \pi)$. There are LSF coefficients in this case:

$$0 < \phi_1 < \theta_1 < \phi_2 < \theta_2 < \cdots < \pi \tag{9}$$

Because of the interlacing properties of these frequencies, LSF coefficients exclusively determine $U(z)$ and $V(z)$, then $A(z)$. We can equivalently turn the coefficients of $A(z)$ into the phases ϕ_i and θ_i of the zeros of $U(z)$ and $V(z)$.

3.3 LSF Transfer Function

Substituting Eqs. (6)-(8) into Eq. (5), we can obtain the equation for the inverse filtering as

$$H^{-1}(z) = k\frac{U_y(z) + V_y(z)}{U_x(z) + V_x(z)}. \quad (10)$$

Here, $U_y(z)$ and $V_y(z)$ are $(Q + 1)$ order symmetric and anti-symmetric polynomials for BC speech that are determined from conjugated zeros or LSF coefficients. $U_x(z)$ and $V_x(z)$ are also similar $(P + 1)$ order polynomials for AC speech. Here, the inverse filtering depends on the LSF coefficients of speech signals, instead of the LP coefficients in Eq. (5).

We usually chose $k = 1$, set the LP-orders to $P = Q = 20$, and the order of transfer function $h(n)$ to $M = 20$ in our experiments. These values for P, Q and M were chosen after considering the sampling frequency, 16 kHz, of the signals. We need to automatically predict AC-LSF coefficients from BC-LSF coefficients to blindly restore BC speech. The relation between the LSF coefficients of AC and BC speech signals is essential to restoring BC speech. Let us consider the problem of predicting AC-LSF in the next section.

4. Blind BC restoration model

All restoration models [7]–[14] needed some information of AC speech to construct inverse filtering to restore BC speech signals. We proposed an LP-based blind restoration model in a previous study with the ability of blind restoration [15] as shown in the block diagram in Fig. 4(a). However, this model has serious limitations due to the mentioned inadequacies of LP coefficients. As this model estimates the inverse filtering based on long-frame of BC speech signal, the inverse filtering is can not be adapted to real-time changes.

We introduced LP representation and LSF coefficients in the previous section. This section proposes an LP-based model using LSF coefficients to estimate inverse filtering. Figure 4(b) is a block diagram of the proposed model. We need to predict the LSF coefficients of AC speech for each respective frame to obtain a blind restoration model as restoration was processed on a frame basis. Since there is overlap between every two neighbors in the series of frames, their restoration should be related. Therefore, a recurrent neural network (RNN) may be a good choice for automatically predicting AC-LSF coefficients on a series of frames. Using RNN, we could predict AC-LSF coefficients from the BC-LSF coefficients of current and previous frames. We propose the application of an Elman network to the problem of prediction.

4.1 Prediction of AC-LSF coefficients

`Problem:` Let $\mathbf{V}_Y$ be the observed vector of BC-LSF coefficients $\mathbf{V}_Y(l_y(1), l_y(2), ..., l_y(Q))$, and let $\mathbf{V}_X$ be the associated vector of AC-LSF coefficients $\mathbf{V}_X(l_x(1), l_x(2), ..., l_x(P))$. We need to approximately predict the best match series of output vector $\mathbf{V}_X$ from the series of input vector $\mathbf{V}_Y$. Since the characteristics of LSF coefficients are those in Eq. (9), the LSF coefficients in vectors $\mathbf{V}_X$ and $\mathbf{V}_Y$ have to be satisfied as

$$0 < l_x(1) < l_x(2) < ... < l_x(P) < \pi, \quad (11)$$
$$0 < l_y(1) < l_y(2) < ... < l_y(Q) < \pi. \quad (12)$$

LSF differentials have positive values in the range of $(0, \pi)$.

$$\Delta(i) = \begin{cases} l(1) & \text{if } i = 1, \\ l(i) - l(i-1) & \text{if } i > 1. \end{cases} \quad (13)$$

Using LSF differentials can help simplify the requirements as in Eqs. (11) and (12) for the prediction problem instead of directly using LSF coefficients. The dynamic range of LSF differentials also varies less widely. Let $\Delta\mathbf{V}_Y$ be the observed vector of BC-LSF differences $\Delta\mathbf{V}_Y(\Delta_y(1), \Delta_y(2), ..., \Delta_y(Q))$, and let $\Delta\mathbf{V}_X$ be the predicted vector of AC-LSF differential $\Delta\mathbf{V}_X(\Delta_x(1), \Delta_x(2), ..., \Delta_x(P))$. We need an Ω model that enables the best match series of output vector $\Delta\mathbf{V}_X$ to be approximately predicted from a series of input vector $\Delta\mathbf{V}_Y$ as: $\overline{\Delta}\mathbf{V}_X \leftarrow \Omega(\Delta V_Y)$.

4.2 Elman - simple recurrent network

A RNN is a feed back network, i.e., a model with bi-directional data flow. While a feed-forward network such as a multilayer perception network (MLP) propagates data linearly from input to output, RNN also propagates data from later to earlier processing stages. Using RNN could help us model the relationship between output with current input and also with previous inputs.

The Elman network, which is also called as simple recurrent network (SRN), has one hidden layer as in Fig. 5, with connections back to a special copy layer. The copy layer is a memory of hidden units, delayed for one time step, and treated as partial inputs. Therefore, standard back-propagation learning techniques for feed-forward networks can be used for training the network [21]. Back-propagation algorithm was used for training then helps to modify network weights with respect to minimize error by the stochastic gradient descent error. A simple calculation of forward-propagation was also used to compute the outputs as the same as in a common MLP network [22], [23]. Here, the S-shaped hyperbolic tangent function was used in the hidden layer,

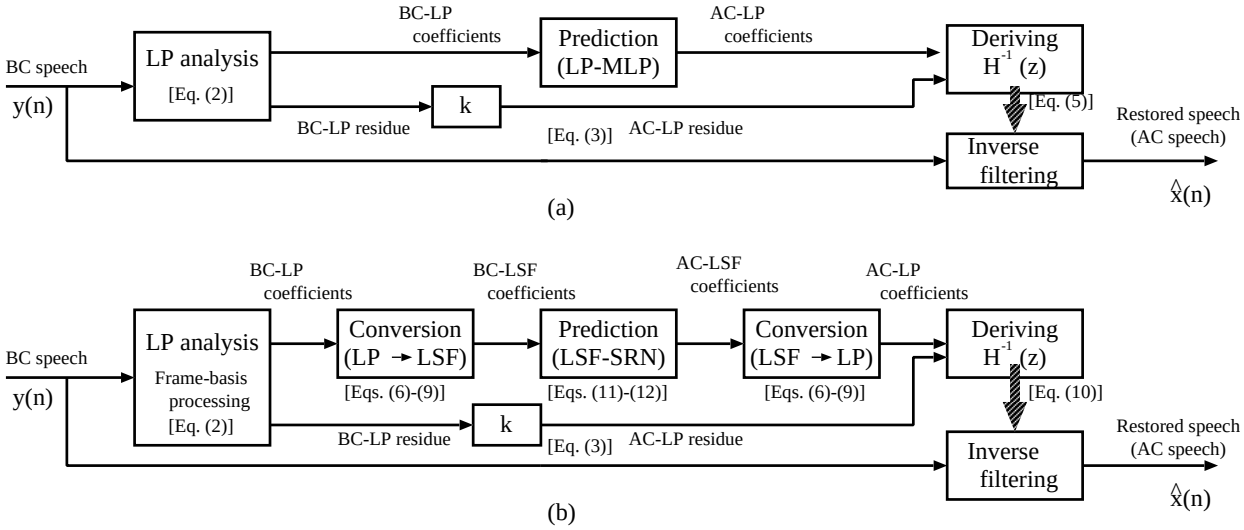


Fig. 4 Block diagrams for (a) previous LP-based blind model and (b) proposed model

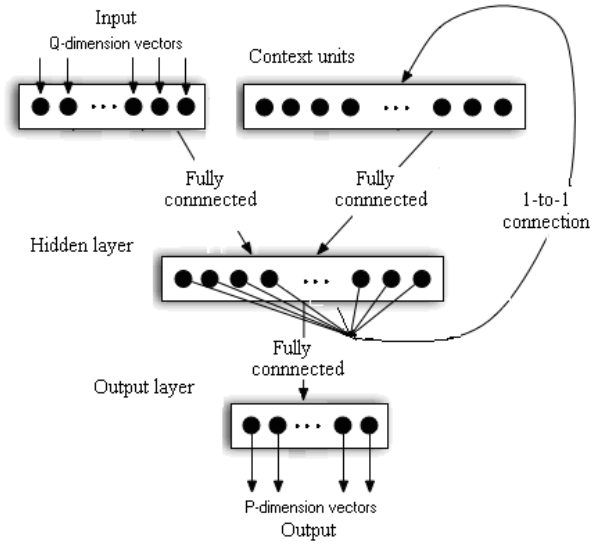


Fig. 5 Simple recurrent network structure.

but linear functions were used at the input and output layers.

Since we chose both LP-orders as $P = Q = 20$, this means that the input and output vectors of SRN had 20 dimensions. The SRN structure is outlined in Fig. 5. There were 20 nodes for the input layer, 20 for the output layer, and 20 for the hidden layer. We then had 1200 connection in this Elman topology. Our proposed model could restore BC speech for every speech frame in real time. We chose a frame length of 250 ms, and the overlap between two neighbors was 125 ms. These values were selected optimally for model's performance, to keep the frame-length sufficiently short, and also to reduce the number of LSF differential vectors (approximately 10.000 samples) to train a small LSF prediction model, as will be discussed in the next section.

5. Evaluation

We discuss the feasibility of models restoring BC speech signals in this section. The main aim of our evaluation was to investigate whether the proposed model could adequately restore BC speech to attain better voice quality and speech intelligibility in human-hearing systems and/or ASR systems and whether this could work well blindly. Moreover, this evaluation was done to find what a significant model for prediction should be.

We evaluated the previous long-term Fourier transform model [4], [5] and the two LP-based models (the first was non-blind model and the second was blind model). In these three models, there were two non-blind models: (1) a long-term Fourier transform (LTF) and (2) an LP-based model using LSF coefficients and frame-based processing (LSF). The blind restoration model was (3) LP-based blind restoration - applying SRN to LSF (LSF-SRN).

LSD (log-spectrum distortion) and JWIT were used to evaluate the improvement in intelligibility, and LCD (LP coefficient distance) and MCD (Mel-frequency cepstral coefficient (MFCC) distance) were used to evaluate the improvements in the cepstral distance of restored speech signals. The Japanese syllable dataset from the AC/BC database was used for the evaluation with objective measures (LSD, LCD, and MCD). JWIT was a subjective measure which is used to estimate intelligibility based on the average recognition accuracy score of subjects. Eighty words (20 words for each familiarity range) were chosen randomly from Japanese word dataset and were used in JWIT as mentioned in Section 5.2.

Table 2 Average improvements of restored speech signals. (Improvement is distance of BC& AC minus that of restored & AC)

Objective Measures	Un-blind models		Blind model
	LTF	LSF	LSF-SRN
Improved LSD	0.75	**1.70**	*0.87*
Improve MCD	1.15	**2.99**	*1.13*
Improve LCD	0.29	**0.96**	*0.21*

Table 3 Stimulus lists for Japanese-word intelligibility test. Five groups: A, B, C, D, E. Four familiarity ranges: R1, R2, R3, R4 (familiarity value effects on speech intelligibility).

Word Index		BC	LTF	LSF	LSF-SRN	AC
R1 (1.0–2.5) *low familiarity*	1–4	A	B	C	D	E
	5–8	E	A	B	C	D
	9–12	D	E	A	B	C
	13–16	C	D	E	A	B
	17–20	B	C	D	E	A
R2 (2.5–4.5) *medium low familiarity*	21–24	A	B	C	D	E
	25–28	E	A	B	C	D
	29–32	D	E	A	B	C
	33–36	C	D	E	A	B
	37–40	B	C	D	E	A
R3 (4.5–5.5) *medium high familiarity*	41–44	A	B	C	D	E
	45–48	E	A	B	C	D
	49–52	D	E	A	B	C
	53–56	C	D	E	A	B
	57–60	B	C	D	E	A
R4 (5.5–7.0) *high familiarity*	61–64	A	B	C	D	E
	65–68	E	A	B	C	D
	69–72	D	E	A	B	C
	73–76	C	D	E	A	B
	77–80	B	C	D	E	A

5.1 Objective evaluations

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

$$\mathrm{LSD} = \sqrt{\frac{1}{W}\sum_{\omega}^{W}\left[20\log_{10}\left(\frac{|S(\omega)|}{|\hat{S}(\omega)|}\right)\right]^2}, \quad (14)$$

$$\mathrm{LCD} = \sqrt{\frac{1}{P}\sum_{i=1}^{P}\left(a_x(i)-a_y(i)\right)^2}, \quad (15)$$

$$\mathrm{MCD} = \sum_{i=0}^{12}\left(c_{x,i}-c_{y,i}\right)^2, \quad (16)$$

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

LSD, LCD and MCD carried out the differential between two speech signals (i.e., restored and AC speech signals). Improved LSD in Table 2 means LSD between original BC and AC speech minus LSD between restored speech and AC speech. This value shows the improvement of restored speech from BC toward AC speech. The calculations were as the same for Improved MCD and Improved LCD. Table 2 showed us average improvements of restored speech signals that are yielded by the restoration models. The LSF model generally produced the best results, due to the shorter distances between restored and AC speech. Thus, the LSF model should be a significant restoration model that improves both voice quality in human-hearing systems and spectral distance for the front-end of ASR systems.

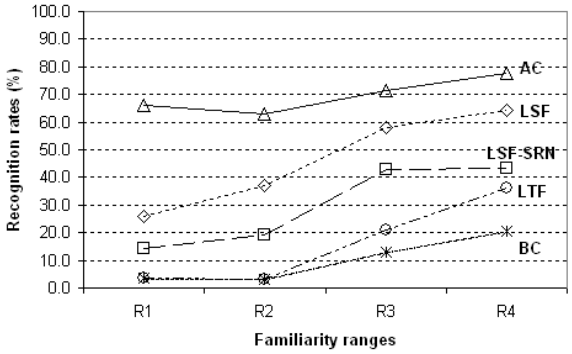


Fig. 6 Japanese-word intelligibility test results.

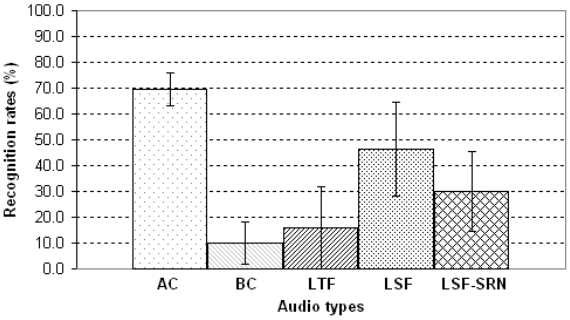


Fig. 7 Average results for Japanese-word intelligibility test.

5.2 Subjective evaluation

We carried out JWIT with forty subjects who had normal hearing for the subjective evaluation. The speech signals of eighty words were played in random order in the tests. The subjects had not heard these words previously and had not been trained before the experiment. They were asked to listen to each word only once and write down what they heard in Hiragana to avoid training effects in determining words with lower familiarity.

We used five types of audio (AC speech, BC speech and three types of restored signals using the three models). We intended to evaluate the intelligibility of these signals in four different familiarity ranges (R1, R2, R3, and R4 as mentioned in Section 2). Since subjects only

Table 4 Japanese-word speech intelligibility test (correction (%)).

Familiarity	BC	LTF	LSF	LSF-SRN	AC
R1 (1.0–2.5)	3.5	3.5	**26.0**	*14.5*	66.0
R2 (2.5–4.5)	3.0	3.0	**37.0**	*19.0*	63.0
R3 (4.5–5.5)	13.0	21.0	**58.0**	*43.0*	71.5
R4 (5.5–7.5)	20.5	36.0	**64.5**	*43.5*	77.5
Avg. (1.0–7.5)	10.0	15.9	**46.4**	*30.0*	69.5

heard each word once, we divided them into five listening groups, i.e., A, B, C, D, and E to listen to 400 stimuli (20 words for each couple in the four familiarity ranges and five audio types). Table 3 shows how the 400 stimuli for the five listening groups were arranged. There were eight subjects in each listening group. Intelligibility could generally be evaluated using the average recognition accuracy scored by all subjects.

Table 4 lists the recognition accuracy scores for the five different audio types in the four ranges, and also the averages. The LSF model is also the best for subjective evaluation followed by LSF-SRN. The subjective evaluation again confirms the restoration of intelligibility of BC speech with the models based on LP.

5.3 Discussion

The non-blind LP-based model, i.e., the LSF model, significantly restored BC speech signals according to the results of evaluation listed in Tables 2 and 4, both in terms of intelligibility in human-hearing systems (with LSD and JWIT) and the spectral distance for the front end of ASR systems. From Fig. 7, the LSF model improved the average recognition accuracy scores of BC speech by more than 36.4% The LSF-SRN model improved the average recognition accuracy scores of BC speech by more than 20.0%.

Figure 6 gives more detail on the improved intelligibility of models in different familiarity ranges. We generally found that it was more difficult to restore BC speech signals in low-familiarity ranges. The LTF model demonstrated no improvements in low familiarity ranges (R1 and R2). The results improved quickly with greater familiarity. The LSF model even improved the average recognition accuracy by about 45% in high familiarity ranges (R3 and R4). The LSF-SRN model improved BC speech in these familiarity ranges up to almost the same average recognition accuracy scores of about 43%. LSF-SRN, even as a blind model, generally improved the intelligibility of BC speech signals. This means that SRN was adequately trained to predict AC-LSF coefficients to enable the LSF-SRN model to restore the intelligibility of BC speech signals as almost the same as that using non-blind AC-LSF coefficients.

As previously mentioned, LSD and JWIT were used to evaluate improvements in the intelligibility of speech, which is necessary for human-hearing systems. LCD and MCD were used to evaluate cepstral distances, which are important in ASR systems. We found that although LSF-SRN significantly improved the intelligibility of BC speech signals from LCD and MCD measures (Table 2), it only demonstrated the same improvement in spectral distance as the LTF method. However, LSF demonstrated sufficient ability to improve the spectral distance. Thus, these results suggest better results for the spectral distance of LSF-SRN can be attained by more SRN training. The blind restoration LSR-SRN model should therefore be able to successfully restore BC speech signals, i.e., not only intelligibility in human-hearing systems but also the spectral distances for the front-end of ASR systems.

6. Conclusions

We improved the LP-based model by (1) extending long-term processing to frame-basis processing, (2) using LSF features, and (3) using a simple recurrent network to predict parameters. Using LSF features helped alleviate the limitations with LP coefficients in prediction using statistical methods. Then, the frame-basis processing with ability of SRN helped LP-based model easily adapt predict inverse filtering to BC speech in real time.

The improved model (LSF) was very efficient in restoring BC speech, both in terms of intelligibility for human-hearing systems and for the spectrum features of ASR systems. Our evaluations demonstrated LSF outperformed the previous LTF model. SRN was then applied to predict the AC-LSF differentials, which are needed to achieve LSF inverse filtering. The results of evaluation confirmed that we could blindly predict these LSF differentials and then use them for a significant restoration model (LSF-SRN). The improved blind restoration model, LSF-SRN, can currently significantly improve the intelligibility of BC speech. Its ability to improve the spectral distance for front-end of ASR systems was the same as that of the previous non-blind LTF model.

We intend to examine this model over a larger AC/BC dataset by considering different pick-up points and also discover what effect it has on restoring different syllables and speakers. The next challenge is to improve the spectral distances since the blind restoration model we propose is still limited in this regard. We intend to achieve an LP-based blind restoration model to vastly improve BC speech signals. Significant improvements in both intelligibility and spectral distances remain problems that needs to be solved to construct a blind restoration model that will be as good as the LSF model. Moreover, the LSF model can be even further improved such as by considering the changing of k since this factor currently is assumed as constant. Besides, BC speech maybe effected by other effects in particular environments. These problems should be considered as the next stage in our future studies.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (No. 17650048) and a scheme for the "21st Century COE Program" in Special Coordination Funds for the Promotion of Science and Technology made available by the Ministry of Education, Culture, Sports, Science, and Technology in Japan. This was also partially supported by a Grant Program by the YAZAKI Memorial Foundation for Science and Technology and SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1] Benesty, J., Makino, S., and Chen, J., Speech enhancement, Springer, Berlin, 2005

[2] Kitamori, S. and Takizawa, M. "An Analysis of Bone Conducted Speech Signal by Articulation Tests," *IEICE Trans.* vol. J72-A, no. 11, pp. 1764–1771, Nov. 1989.

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

[4] Nakagiri, M., Toda, T., Kashioka, H., and Shikano, K. "Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion," *Proc. ICSLP2006*, pp. 2270–2273, Sept. 2006.

[5] Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., Acero, A., and Huang, X. "Air-and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement," *Proc. ASRU*, pp. 249–253, Dec. 2003.

[6] Subramanya, A., Zhang, Z., Liu, Z., Droppo, J., and Acero, A. "A Graphical Model for Multi-Sensory Speech Processing in Air-and-Bone Conductive Microphones," *Proc. Eurospeech2005*, pp. 2361–2364, Lisbon, Portugal, Sept. 2005.

[7] Ishimitsu, S., Kitakaze, H., Tsuchibushi, Y., Yanagawa, H., and Fukushima, M. "A noise-robust speech recognition system making use of body-conducted signals," *Acoust. Sci. & Tech.*, vol. 25, no. 2, pp. 166–169, 2004.

[8] Kondo, K. Fujita, T., and Nakagawa, K. "Equalization of Bone Conducted Speech for Improved Speech Quality," *Proc. 6th IEEE International Symposium on Signal Processing and Information Technology*, pp. 426–431, Vancouver, BC, Canada, Aug. 2006.

[9] Tomikura, T. and Shimamura, T. "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.

[10] Tamiya, T. and Shimamura, T. "Reconstruct Filter Design for Bone-Conducted Speech," *Proc. ICSLP2004*, vol. II, pp. 1085–1088, Oct. 2004.

[11] Shimamura, T., Mamiya, J. and Tamiya, T. "Improving Bone-Conducted Speech Quality via Neural Network," *Proc. 6th IEEE International Symposium on Signal Processing and Information Technology*, pp. 628–632, Vancouver, BC, Canada, Aug. 2006.

[12] Thang, V. T., Kimura, K., Unoki, M., and Akagi, M. "A study on restoration of bone-conducted speech with MTF-based and LP-based model," *J. Signal Processing*, vol. 10, no. 6, pp. 4070-417, Nov. 2006.

[13] Thang, V. T., Unoki, M., and Akagi, M., "A study on an LPC-based restoration model for improving the voice-quality of bone-conducted speech," 2006 RISP International Workshop on Nonlinear Circuits and Signal Processing (*NCSP'06*), pp. 110–113, Hawaii, USA, Mar. 2006.

[14] Thang, V. T., Unoki, M., and Akagi, M., "A Study on Restoration of Bone-conducted Speech with LPC-based Model," IEICE Technical Report, SP2005-174, pp. 67-72, March 2006.

[15] Thang, V. T., Unoki, M., and Akagi, M. "A study on an LP-based model for restoring bone-conducted speech," the International Conference on Communications and Electronics (*Proc. ICCE' 2006*), pp. 294–299, Hanoi, Vietnam, Oct. 2006.

[16] Thang, V. T., Unoki, M., and Akagi, M., "An LP-based blind restoration method for improving intelligibility of bone-conducted speech" IEICE Technical Report, SP2006-172, pp. 19–24, March 2007.

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

[18] Sakamoto, S., Iwaoka, N., Suzuki, Y., Amano, S., and Kondo, T. "Complementary relationship between familiarity and SNR in word intelligibility test," *Acoust. Sci. & Tech*, vol. 25, no. 4, pp. 290–992, 2004.

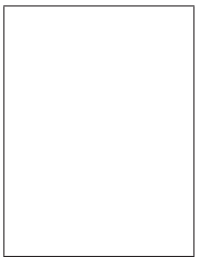
[19] Saitou, Y. , Niigaki, T. ,Nagano, Y., Fukushima, M., Ishimitsu, S., and Yanagawa, H. , "Change of the voice picked up by accelerometer of the face," Proc. 2002 autumn meeting on Acoust. Soc. Jpn, 3-P-22, pp. 623-624, 2002.

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

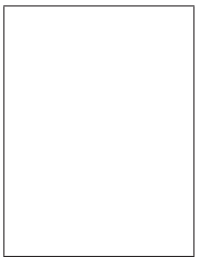
[21] Elman, J. L., "Finding structure in time," Cognitive Science, vol 14, pp. 179-211, 1990.

[22] Bishop, C. M., Neural networks for pattern recognition, Oxford University Press, Oxford, UK, 1995

[23] Nabney, I. T., NETLAB: Algorithms for Pattern Recognition, Springer-Verlag, London, 2002.

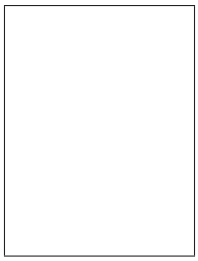


Thang tat Vu was born in Hanoi, Vietnam, in 1979. He received his B.E. and M.S. in electronic & telecommunication engineering from the Hanoi University of Technology (HUT) in 2002 and 2004. He was a member of the Speech Processing Group at the Institute of Information Technology (IOIT) of the Vietnamese 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. He is a member of the Research Institute of Signal Processing (RISP).



Masashi Unoki was born in Akita Prefecture, Japan, in 1969. He 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 auditory motivated 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 from 2000-2001 he was a visiting research associate at the CNBH of the Department of Physiology at the University of Cambridge. Since 2001 he has been on the faculty of the School of Information Science at JAIST 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. He is also a member of the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). Dr. Unoki received the Sato Prize for an Outstanding Paper from the ASJ in 1999 and the Yamashita Taro Prize for Young Researchers from the Yamashita Taro Research Foundation in 2005.



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