2007 International Symposium on Nonlinear Theory and its Applications
NOLTA'07, Vancouver, Canada, September 16-19, 2007

A blind restoration model for bone-conducted speech based on a linear prediction scheme

Thang Tat Vu†, Masashi Unoki†, and Masato Akagi†

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

Abstract—Bone-conducted (BC) speech can be used instead of air-conducted (AC) speech for speech communication systems in extremely noisy environments. However, it has very poor sound quality and its intelligibility is degraded when transmitted through bone conduction. Therefore, blindly improving voice quality and the intelligibility of BC speech is a challenging topic. We propose a linear prediction (LP) scheme based blind-restoration model to improve voice quality and the intelligibility of BC speech. This LP-based method originates from the linear predictive concept, which regards speech signals as the representation of source and filter information. We evaluated the proposed model in comparison with other models to find out whether it could adequately improve voice quality and the intelligibility of BC speech, using objective measures (LSD, MCD, and LCD) and carrying out word intelligibility tests for Japanese words and modified rhyme tests for English words. The experimental results for objective and subjective evaluations proved the practicability of blind BC restoration.

1. Introduction

It is very difficult for automatic speech recognition (ASR) systems or humans to accomplish speech communications in extremely noisy environments. Many different complex models have been used to reduce interfering noise to solve this but these are ineffective when the noise levels are too high. Another possible solution is to use bone-conducted (BC) speech due to its stability against interfering noise [1]. Although not affected by external noise, the BC speech is attenuated in a complex way when transmitted through bone conduction. The voice quality and intelligibility of BC speech are degraded due to bone-conduction. We are therefore presented with a new challenge in the speech signal-processing field since it is very difficult to blindly restore those of BC speech.

The purpose of our approach was to restore BC speech so that it could be directly applied to human-hearing systems and the front end of ASR systems. This means that BC speech should be restored blindly without other information such as that on AC speech, and the restored processing should adapt incoming BC speech. As various methods of simply deriving inverse filtering such as the cross-spectrum and long-term Fourier transform methods [2] yielded restored signals with artifacts such as musical noise and echoes, they only created slight improvements in voice quality [3, 4]. Moreover, these methods are difficult to adapt to BC speech's characteristics, which change due to conditions such as BC measurement points, pronounced syllables, and speakers.

Our strategy was to complete a practical framework that would help to restore BC speech. We proposed restoration models using the linear prediction (LP) concept as preliminary models in previous studies. LP-based models [3, 4] were proposed that originated from the concept of the source-filter model. This model could yield restored both voice quality and the intelligibility of BC speech signals. Moreover, we proposed an LP-based model with the ability of blind restoration by predicting parameters [4] from the fact that the LP-based model only depended on the unknown LP coefficients of AC speech (AC-LP coefficients). Machine-learning methods were applied to predicting AC-LP coefficients and some reasonable results were obtained. However, that model [4] still had significant limitations: (a) the LP coefficients were not suitable for prediction with statistical models; (b) even small prediction errors could cause problems with filter instability; and (c) inverse filtering was also determined to remain unchanged for an entire BC speech signal [3, 4].

We improved the model of LP-based blind restoration by (1) extending the processing scheme from long-term to a frame-basis, (2) converting stable parameters of LSF coefficients on LP representation, and (3) predicting parameters using a recurrent neural network. Since LSF coefficients play the same role in the presentation of the spectrum envelope and are limited within a range $(0, \pi)$, they could help alleviate the limitations with LP coefficients in predictions. The processes of restoration on a frame-basis could also be adapted to inverse filtering in real time. A simple recurrent network was applied to predict AC-LSF coefficients to complete the blind restoration system.

2. Blind BC restoration model

2.1. Signal restoration diagram based on LP

Let $x(n)$ and $y(n)$ be discrete signals of AC and its associated BC speech. They are represented [3, 4] as:

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

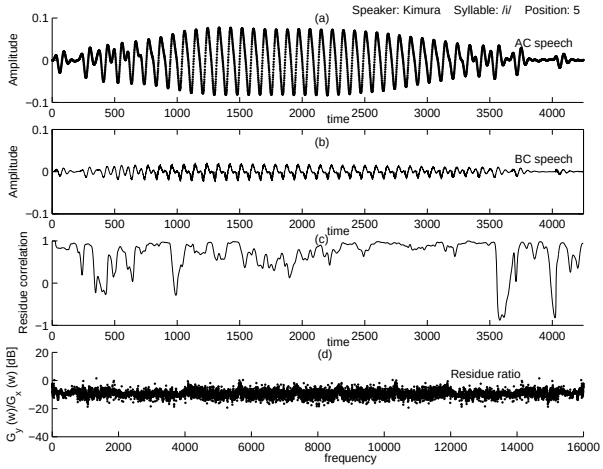

Figure 1: Ratios of AC-BC residues: (a) AC speech, $x(t)$, (b) BC speech, $y(t)$, (c) residue correlation ($g_x(n)$ and $g_y(n)$), and (d) residue ratio, $G_y(\omega)/G_x(\omega)$.

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

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

Figure 1 shows a typical example of the relation between AC and BC speech signals. This suggests that the AC and BC residues are almost the same except for magnitude. We can therefore represent this approximately as a constant factor, k, as

$$G_y(z)/G_x(z) = k. \quad (3)$$

Let us assume that $H(z)$ is the transfer function from AC speech to BC speech in the z-domain. The inverse filter, $H^{-1}(z)$, can be found as the inverse of $H(z)$ and used to easily restore BC to AC speech. We can obtain the equation for $H^{-1}(z)$ simply as [3, 4]

$$H^{-1}(z) = k\sum_{i=0}^{Q} a_y(i)z^{-i} \Big/ \sum_{i=0}^{P} a_x(i)z^{-i}. \quad (4)$$

The constant value, k, can be chosen manually and used to control the magnitude of restored speech. The latter term depends on the LP coefficients of signals. Therefore, these LP coefficients have to be predicted from observed BC speech; LP coefficients, however, are inappropriate parameters for statistical models. Line spectral frequency (LSF) coefficients are thus used as more appropriate parameters in this paper.

2.2. LSF representation

Let $A(z)$ be an 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.

$$A(z) = \sum_{i=0}^{P} a(i)z^{-i}, \qquad a(0) = -1, \quad (5)$$

$$U(z) = A(z) + z^{-(P+1)}A(z^{-1}), \quad (6)$$

$$V(z) = A(z) - z^{-(P+1)}A(z^{-1}). \quad (7)$$

$U(z)$ 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 are interlaced with each other in the interval, $(0, \pi)$.

$$0 < \phi_1 < \theta_1 < \phi_2 < \theta_2 < \cdots < \pi. \quad (8)$$

The interlacing properties of LSF coefficients help to exclusively determine $U(z)$ and $V(z)$, then $A(z)$. Substituting Eqs. (5)-(7) into Eq. (4), we can obtain:

$$H^{-1}(z) = k\frac{U_y(z) + V_y(z)}{U_x(z) + V_x(z)}. \quad (9)$$

Here, the inverse filtering depends on the LSF coefficients of speech signals, instead of the LP coefficients.

Figure 2 is a block diagram of the LP-based blind-BC speech restoration model. We will explain how to predict AC-LSF coefficients in this section.

2.3. Prediction of AC-LSF coefficients

`Problem:` Let $\mathbf{V}_Y$ be the observed vector of BC-LSF coefficients $\mathbf{V}_Y(l_y(1), l_y(2), \cdots, l_y(q))$, and let $\mathbf{V}_X$ be the associated vector of AC-LSF coefficients $\mathbf{V}_X(l_x(1), l_x(2), \cdots, l_x(p))$. We need to approximately predict the best match series of output vector $\mathbf{V}_X$ from a series of input vectors $\mathbf{V}_Y$. Since the characteristics of LSF coefficients are as in Eq. (8), LSF differentials have positive values in the range of $(0, \pi)$. Instead of using LSF coefficients directly, using LSF differentials can help simplify the requirements for predicting problems. Let $\Delta\mathbf{V}_Y$ be the observed vector of BC-LSF differences $\Delta\mathbf{V}_Y(\Delta_y(1), \Delta_y(2), \cdots, \Delta_y(q))$, and let $\Delta\mathbf{V}_X$ be the predicted vector of AC-LSF differential $\Delta\mathbf{V}_X(\Delta_x(1), \Delta_x(2), \cdots, \Delta_x(p))$. We need a model Ω that can approximately predict the best match series of output vectors $\delta\mathbf{V}_X$ from a series of input vector $\Delta\mathbf{V}_Y$ as: $\overline{\Delta}\mathbf{V}_X \leftarrow \Omega(\Delta V_Y)$.

The Elman network, which is also called a simple recurrent network (SRN), has one hidden layer with connections from its hidden layer back to a special copy layer. The special copy layer is treated as just another set of inputs and so standard back-propagation learning techniques, i.e., common supervised learning technique, can be used for training network [5, 6].

Since the function learnt by the network depends on the current inputs and previous states of the network, this model should be a good choice for solving our problem. We chose $k = 1$ and set $P = Q = 20$ in this paper. This means that the input and output vectors have 20 dimensions. There were 20 nodes for each layer: the input layer, the output layer and the hidden

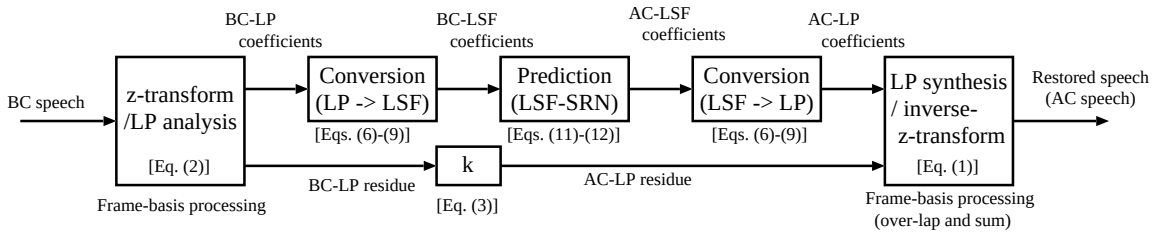

Figure 2: Block diagrams of proposed model.

Table 1: Objective evaluations (Japanese data).

Objective Measure	BC	Non-blind		Blind
		LTF	LSF	LSF-SRN
LSD	12.08	11.33	10.38	11.21
MCD	20.52	19.37	17.53	19.39
LCD	2.80	2.51	1.83	2.58

Table 2: Objective evaluations (English data).

Objective Measure	BC	Non-blind		Blind
		LTF	LSF	LSF-SRN
LSD	14.28	13.57	8.92	9.64
MCD	21.72	18.15	12.55	15.96
LCD	3.04	2.91	1.79	2.43

layer. We chose 250 ms as the length of frames, and 125 ms as the overlap of two neighbors. These values were to keep the frame-length sufficiently short, and also reduce the number of training vectors for a small prediction model.

3. Evaluations

The aim of our evaluations was to investigate whether the proposed model could adequately restore BC speech to attain better voice quality and speech intelligibility and whether this would work well blindly.

We built two AC/BC databases for the experiments, the first for Japanese and the second for English. The Japanese database included 101 syllables and 100 words recorded from ten speakers. These words were selected from Japanese word lists for the intelligibility test by NTT-AT (2003) [8]. The English database included 300 words of the modified rhyme test (MRT), recorded from six speakers. BC and clean AC speech signals were recorded simultaneously. The BC speech was collected at five different positions on the head and face: (1) the mandibular angle, (2) the temple, (3) the philtrum, (4) the forehead, and (5) the calvaria.

Using both objective and subjective measurements, we evaluated a long-term Fourier transform model (LTF) [3] and the two LP-based models (the first was a non-blind model with LSF coefficients (LSF) and the second was a blind model with SRN applied to LSF (LSF-SRN)).

3.1. Objective evaluations

We used log-spectrum distortion (LSD), LP distance (LCD), and mel frequency cepstral coefficient distance (MCD) for the Japanese and English databases used to evaluate the methods. These objective measurements were computed as:

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

$$\text{LCD} = \sqrt{\frac{1}{P}\sum_{i=1}^{P}\left(a_x(i)-a_y(i)\right)^2}, \quad (11)$$

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

where W is the upper frequency (8 kHz in this case), and $S(\omega)$ and $\hat{S}(\omega)$ are the amplitude spectra obtained by the 1024-points FFT calculation of 25-ms frames with 15-ms overlap. $a_x(i)$ 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 mel frequency cepstral coefficients (MFCCs) of the signals.

Tables 1 and 2 show that the distances of the three objective measurements between the clean AC speech signal and the observed BC speech and restored speech signals for Japanese and English datasets, respectively. The results of comparison revealed that the LSF model was the best for all measurements. Even for blindly restored BC speech, the LSF-SRN was almost the same as LTF on the Japanese database and closely followed the best LSF model on the English database.

3.2. Subjective evaluation

Word intelligibility tests (WITs) were carried out on the Japanese database with 40 Japanese subjects and the modified rhyme tests (MRTs) were carried out on the English database with six English native speakers. All the subjects had normal hearing.

The speech signals of 80 words were played in random order in the WITs. The subjects, who did not know these words previously, were asked to listen to each word only once and write down what they heard in Hiragana. We intended to evaluate the intelligibility of these signals in four different familiarity ranges [8]. Since all subjects listened to a word only once, we divided the 40 subjects into five listening groups to listen to 400 stimuli. Then, subjects in each group listened to 80 distinct words. Intelligibility could generally be evaluated by the average recognition accuracy, which

Table 3: Word 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
Average	10.0	15.9	**46.4**	*30.0*	69.5

Table 4: Modified rhyme test (correction (%)).

BC	LTF	LSF	LSF-SRN	AC
69.7	76.3	**88.0**	*82.5*	95.7

was scored for all subjects. Table 3 lists the recognition accuracy scores of the WITs.

Listeners in the MRTs were given six-word lists and then required to identify which of the six had been spoken. There were 50 six-word lists of rhyming or similar sounding monosyllabic English words. Every word was in consonant-vowel-consonant sound sequence, and the six words in each list only differed in the sound of the initial or final consonant. The MRT results indicated errors in discrimination for both initial and final consonant sounds, and also showed improvements in the intelligibility of restored speech [7]. Table 4 lists the correct scores for the MRTs. The LSF model yielded the best results, the same as for the objective measurements, closely followed by LSF-SRN.

3.3. Discussion

The evaluation results in Tables 1, 2, 3, and 4 demonstrate that the non-blind LP-based model, LSF, and the blind LP-based model, LSF-SRN, restored the BC speech signal significantly, both in terms of intelligibility (LSD, MRT, and WIT) and the spectral distance for the front end of ASR systems (LCD, MCD).

The LSF model improved the average recognition accuracy of BC speech by 36.4 % in the WIT scores. The LSF-SRN model followed closely with an expressed result 20 % greater than that of BC speech. We generally found that it was more difficult to restore the BC speech signal in low familiarity ranges. The LTF model yielded no improvements in low familiarity ranges (R1 and R2). There were more improvements when these was higher familiarity. The LSF model even improved the average recognition accuracy by about 45 % in high familiarity ranges (R3 and R4). At these familiarity ranges, LSF-SRN improved the BC speech by almost the same amount (43 %) in these familiarity ranges.

Even though it is a blind model, LSF-SRN had the ability to improve voice quality and the intelligibility of the BC speech signal. The improvements in intelligibility were evident for both English (MRTs) and Japanese (WITs). This also means that the SRN was adequately trained to predict AC-LSF coefficients and this then helped the LSF-SRN model to achieve good restoration.

4. Conclusion

We proposed an LP-scheme-based restoration model for improving voice quality and the intelligibility of BC speech. In this scheme, we improved the model of LP-based blind restoration in this scheme in three ways by (1) extending the processing scheme from a long-term to a frame-basis, (2) converting stable parameters of LSF coefficients on LP representations, and (3) predicting parameters using a recurrent neural network. We comprehensively evaluated the model we developed on two different AC/BC databases to compare it with other models to find whether it could adequately improve voice quality and the intelligibility of BC speech. We used three objective measures and two subjective tests. The experimental results revealed that the LP-based model was sufficiently practical for blind-BC restoration. The model could especially be applied to improving the intelligibility of BC speech without considering language differences.

We intend to evaluate this model using a larger AC/BC dataset in future work and different measuring positions for recording. We also intend to assess what effect it has on restoring different syllables and utterances. Building a blind restoration model as good as the LSF model poses a real challenge in the future.

Acknowledgments

This work was supported by a Grant Program by the YAZAKI Memorial Foundation for Science and Technology and a scheme for the "21st Century COE Program". It was also partially supported by the SCOPE (071705001) of the Ministry of Internal Affairs and Communication (MIC), Japan.

References

[1] Kitamori, S. and Takizawa, M. "An Analysis of Bone Conducted Speech Signal by Articulation Tests," *IEICE Trans.* **J72-A**(11), 1764–1771, Nov. 1989.

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

[3] Thang, V. T., Kimura, K., Unoki, M., and Akagi, M. "A study on restoration of bone-conducted speech with MTF-based and LP-based models," *J. Signal Processing*, **10**(6), 407–417, Nov. 2006.

[4] Thang, V. T., Unoki, M., and Akagi, M. "A study on an LP-based model for restoring bone-conducted speech," *Proc. ICCE' 2006*, 294–299, Hanoi, Vietnam, Oct. 2006.

[5] Bishop, C. M., Neural networks for pattern recognition, Oxford University Press, Oxford, U.K., 1995.

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

[7] Brungart, D. S. "Evaluation of speech intelligibility with the coordinate response measure," *J. Acoust. Soc. Am.*, **109**(5), 2276–2279, May, 2001.

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