Journal of Signal Processing, Vol.16, No.5, pp.409-417, September 2012

PAPER

A Study on Restoration of Bone-Conducted Speech in Noisy Environments with LP-based Model and Gaussian Mixture Model

Phung Nghia Trung, 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: {ptnghia, unoki, akagi}@jaist.ac.jp

Journal of Signal Processing

信号処理

Journal of Signal Processing, Vol. 16, No. 5, pp. 409-417, September 2012

PAPER

A Study on Restoration of Bone-Conducted Speech in Noisy Environments with LP-based Model and Gaussian Mixture Model

Phung Nghia Trung, 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: {ptnghia, unoki, akagi}@jaist.ac.jp

Abstract **The restoration of bone-conducted speech is a very important issue that enables robust speech communication in extremely noisy environments. We proposed a method of blind restoration in our previous studies based on a scheme of linear prediction with a method of training and prediction based on the simple recurrent neural network. However, prediction based on neural networks is not suitable for training with large corpora, which is necessary for real applications. The over-training problem with simple recurrent neural networks makes it difficult to train various kinds of bone-conducted speech in one session. In addition, it is difficult to adapt the neural network model to bone-conducted speech in unknown noisy environments to build an open dataset restoration of bone-conducted speech. Thus, a method of training and prediction based on the Gaussian mixture model was used in this research, instead of a neural network. A method of re-estimating the residual ratio in the scheme of linear prediction is also proposed. We also investigated how the proposed method works to restore bone-conducted speech in extremely noisy environments. Objective and subjective evaluations were carried out to evaluate the improvements in sound quality and the intelligibility of restored speech. The results revealed that our proposed method outperformed previous methods in both human hearing and automatic speech recognition systems even in extremely noisy environments.**

Keywords: bone-conducted speech, Gaussian mixture model, linear prediction, speech intelligibility

1. Introduction

Speech communication in noisy environments still remains a challenge. There have been many models and algorithms to reduce noise in noisy speech, but there is still a lack of efficient models and algorithms in extremely noisy environments. Bone-conducted (BC) speech in extremely noisy environments is stable against surrounding noise so that it is able to be efficiently used for communication instead of air-conducted (AC) speech [1].

However, there are two main drawbacks to BC speech, i.e., degradation due to bone conduction and changing speaker pronunciations due to surrounding noise, which is referred to as the Lombard effect. While the Lombard effect is a typical problem, which is the same as that in AC speech in noisy environments, another is its critical effect on the quality of speech. When signals are transmitted through bone conduction, they are complexly affected by a loss of sound quality and intelligibility of speech. The degradation varies for different pick-up points (i.e., BC microphone positions), speakers, and the way syllables are pronounced. This is because the characteristics of bone conduction vary for different measuring positions and the distribution of frequency components varies with speakers who pronounce syllables differently.

There have been many studies on the restoration of BC speech in the literature to overcome the degradation in BC speech caused by bone conduction. However, the results have still been limited. For example, a model for restoring BC speech based on cross spectrum has been proposed [2], and long-term Fourier transform (LTF) has been applied to restore BC speech [3]. These methods seem the simplest and most straightforward methods of restoring BC speech, but they have yielded restored signals with artifacts such as musical noise and echoes and only achieved slight improvements in voice quality [4]. In addition, these methods have been difficult to apply to blind

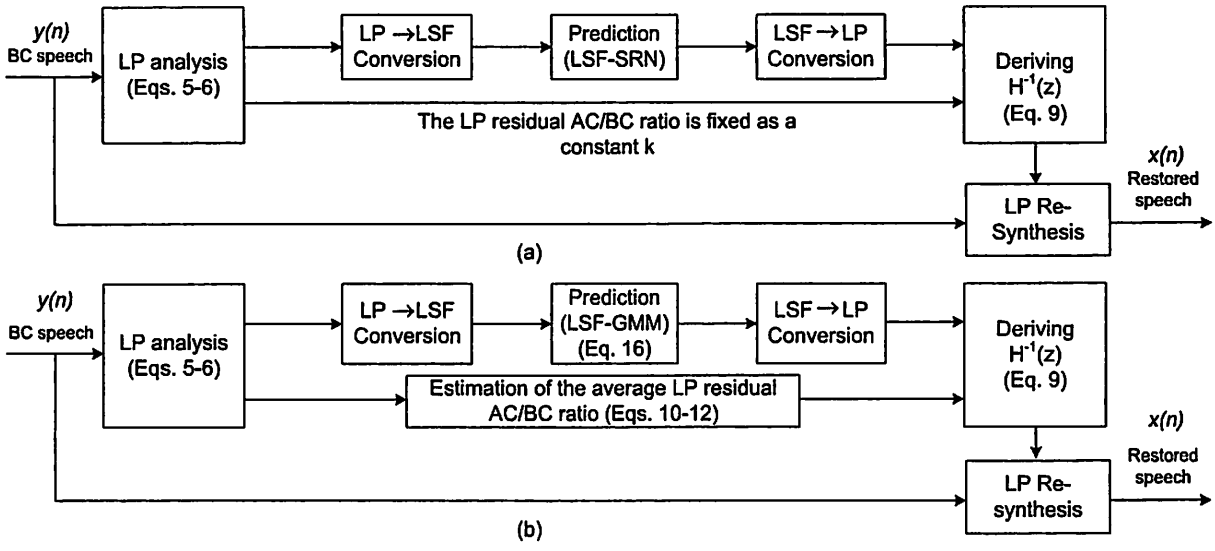


Fig. 1 Block diagrams of BC-speech blind restoration based on LP scheme: (a) Our previous model and (b) our proposed model

restoration.

The approach of using the modulation transfer function (MTF) to restore BC speech has been proposed [4] to overcome the drawbacks with previous methods. The MTF-based model has better restoration abilities and yields restored signals with better intelligibility than the cross-spectrum and LTF methods. However, the quality of speech is still limited and it is still difficult to predict the model's parameters in blind restoration.

Body-transmitted speech (which is like BC speech) restoration based on the Gaussian mixture model (GMM) has been proposed [5], which has been adopted from a technique of voice conversion. In general, GMM is flexible and available for training with huge amounts of data. Therefore, it is easy to train GMM under various conditions and to adapt the trained models under various conditions to those of other unknown conditions. This is an advantage of the use of GMM in the voice conversion. However, due to the difficulty of estimating the fundamental frequency (F0) from these signals, this approach has only been efficiently applied to unvoiced speech such as whispered speech.

We proposed a scheme of linear prediction (LP) to restore BC speech in previous studies [6]. Instead of long-term processing as in traditional methods, we used short-term frame-based processing in this model, which might be used in real-time practical applications. The inverse filter was built only based on line spectral frequency (LSF) coefficients without F0 estimation. We used a simple recurrent neural network (SRN) to blindly predict the LSF of AC speech from BC speech after a training process.

The experimental results presented in our previous paper revealed that our method of LSF-SRN, based on the LP scheme, could adequately improve the quality and intelligibility of BC speech, and it could also efficiently be applied to blind-restoration and real-time applications. However, this method also had three outstanding problems that needed to be solved to further improve the quality and intelligibility of speech.

Vu et al. [6] assumed that the residual ratio was a constant. However, these values changed from frame to frame in the time domain in our current analysis, and thus they should be optimized for each frame.

The learning method we used in our previous restoration model was SRN but it is impractical to use SRN for training huge corpora. Due to the overtraining problem with SRN, it is not suitable to train various kinds of BC speech in one training session. Another problem with SRN training is that it is difficult to adapt the SRN model to BC speech in unknown noisy environments. This problem makes it difficult to build an open dataset BC speech restoration in noisy environments.

We only evaluated the performance of an LP scheme for BC speech restoration in a clean environment in previous studies. Thus, we needed to confirm whether the LP scheme (both LSF-SRN and our currently proposed LSF-GMM) was useful for restoring BC speech in noisy environments.

We solved all three outstanding problems to improve the performance of the LP scheme in the restoration of BC speech in this study. Instead of using the fixed residual ratio for all frames, we approximated these ratios as the ratios of the averaged LP residuals of AC/BC frame pairs. To overcome the drawbacks with SRN, we used GMM to train the joint vector of LSF and the average LP residual of BC and associ-

ated AC speech, since the GMM-based approach can be used to adapt the trained models of BC speech under various noisy conditions to those of other unknown noisy conditions and then to restore BC speech in unknown noisy environments. The LSF and averaged residual of AC speech were then predicted by using those of BC speech and the trained parameters.

The rest of this paper is organized as follows. The next section describes the restoration of BC speech based on the LP scheme. Section 3 explains the problems with the current LP scheme for the restoration of BC speech. Section 4 explains our improved method. Section 5 presents evaluations and Section 6 concludes with a summary and mentions future work.

2. Restoration of BC Speech Based on LP Scheme

2.1 Definition of LP scheme

The flow of the LP scheme for the blind restoration of BC speech [6] is outlined in Fig. 1. We compute the LSF parameters of BC speech and corresponding AC speech to train SRN in the training phase. We have to predict the LSF parameters of AC speech based on those of BC speech in the restoration phase and the corresponding trained SRN parameters. The residual ratio is fixed as a constant, k. This residual ratio and the LP parameters restored from LSF parameters are used to derive the inverse filtering function to convert BC speech to associated AC speech.

Let $x(t)$ and $y(t)$ be AC and associated BC speech signals. Using LP analysis, the discrete signals, $x(n)$ and $y(n)$, can be represented as:

$$\hat{x}(n) = -\sum_{i=1}^{P} a_x(i)x(n-i) \quad (1)$$

$$\hat{y}(n) = -\sum_{i=1}^{P} a_y(i)y(n-i) \quad (2)$$

where $\hat{x}(n)$ and $\hat{y}(n)$ are the predicted signals, P is the LP order, $x(n-i)$ and $y(n-i)$ are the previous observed values, and $a_x(i)$ and $a_y(i)$ are the i-th LP coefficients where $i = 1, 2, \cdots, P$. The residual is obtained by using the error between the current and the predicted samples.

$$g_x(n) = x(n) - \hat{x}(n) \quad (3)$$

$$g_y(n) = y(n) - \hat{y}(n) \quad (4)$$

Here, $x(n)$ and $y(n)$ are represented by the LP model in the z-domain as:

$$-G_x(z) = X(z)\sum_{i=0}^{P} a_x(i)z^{-i}, \quad a_x(0) = -1 \quad (5)$$

$$-G_y(z) = Y(z)\sum_{i=0}^{P} a_y(i)z^{-i}, \quad a_y(0) = -1 \quad (6)$$

where $X(z)$ and $Y(z)$ are the z-transforms of $x(n)$ and $y(n)$. Here, $G_x(z)$ and $G_y(z)$ are the z-transforms of the LP residuals $g_x(n)$ and $g_y(n)$. In Vu et al. [6], the residual ratio of $x(n)$ and $y(n)$ in z domain (or frequency domain) was defined as gain k:

$$k = G_x(z)/G_y(z) = G_x(e^{j\omega})/G_y(e^{j\omega}) \quad (7)$$

2.2 LSF inverse filtering in LP scheme

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

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{i=0}^{M} h(i)z^{-i} \quad (8)$$

Vu et al. [4] demonstrated that the inverse filter could be represented using LSF parameters.

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

Here, $(U_y(z), V_y(z))$ and $(U_x(z), V_x(z))$ are a symmetric polynomial and an anti-symmetric polynomial for BC and AC speech that are determined from the LSF coefficients. Inverse filtering therefore depends on the LSF coefficients of AC and BC speech and gain k.

3. Problems with Current LP Scheme for Restoration of BC Speech

The latest method using the LP scheme to restore BC speech is known as the method of LSF-SRN as in Vu et al. [6]. This approach is based on the supposition that the LP residual is related to the source information (glottal information) of speech, and this kind of information may remain unchanged in both AC and BC speech signals. Therefore, the inverse restoration function is built up with a fixed value of the averaged LP residuals ratio of AC and BC speech. However, in our current analysis shown in Fig. 2, these average values change from frame to frame in the time domain.

The learning method used in the previous LP scheme was SRN. SRN and other neural-network-based training techniques can be used efficiently with small corpora, but when the size of the training corpus increases, the time taken for training will greatly increase. This makes it impractical to use SRN for training huge corpora. In addition, we had to separately train the joint LSF vectors of AC/BC speech for each specific condition in our previous blind restoration of BC speech due to the over-training problem with SRN. When we extended the method of blind BC speech restoration to various kinds of noisy environments, SRN did not seem to be suitable for training. Another problem with SRN training is that it is difficult to adapt the SRN model to BC speech in unknown noisy environments. This problem makes it

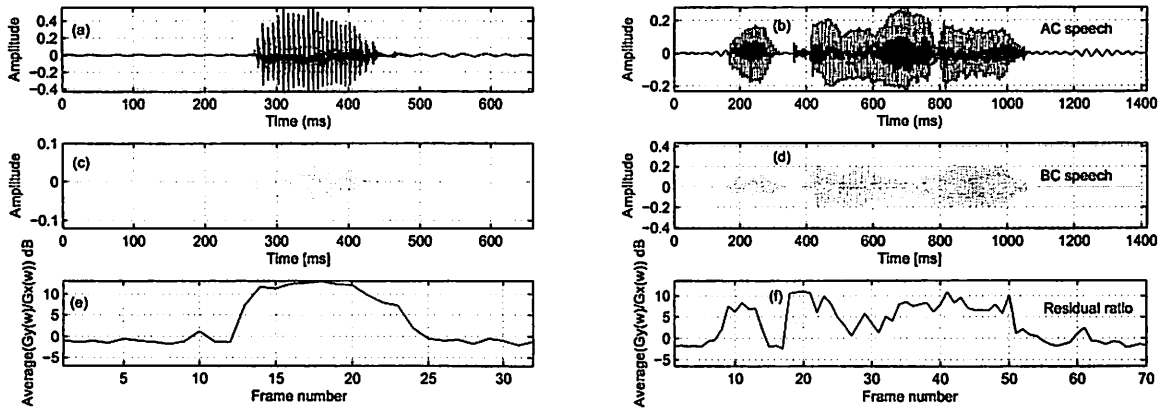


Fig. 2 (a) and (b): AC speech, (c) and (d): BC speech, and (e) and (f): Variations in k for /a/ (left panel) and for /nukumori/ (right panel) of a male speaker

difficult to build an open dataset BC speech restoration in noisy environments. Here, statistical models such as GMM might be a better solution for training the joint LSF vectors of AC/BC speech for various kinds of clean and noisy speech.

We only evaluated the performance of the LP scheme to restore BC speech in clean environments in previous studies. The goal of BC speech communication is especially to use it in noisy speech environments. Therefore, we should confirm whether the LP scheme (both LSF-SRN and our proposed LSF-GMM) is useful for restoring BC speech in noisy environments.

4. Improved Method

4.1 Re-estimation of residual ratio

Vu et al. [4] investigated change in gain k in the frequency domain and assumed gain k would be constant in the time domain. We investigated the change in gain k in the time domain in this research. The results we obtained from analysis are given in Fig. 2, where we can see the change in k in the time domain, especially in long syllables, is considerable.

The residuals ratio in the frequency domain can be fixed as an average constant factor over all frequencies [4]. Therefore, instead of using constant gain k for all frames [4], we compute the average LP residuals for each frame of AC/BC speech to better estimate gain k. Gain k is computed as:

$$k \approx \overline{G_x(e^{j\omega})} / \overline{G_y(e^{j\omega})} \tag{10}$$

where

$$\overline{G_x(e^{j\omega})} = \frac{1}{N}\sum_{i=1}^{N} G_x(e^{j\omega_i}) \tag{11}$$

$$\overline{G_y(e^{j\omega})} = \frac{1}{N}\sum_{i=1}^{N} G_y(e^{j\omega_i}) \tag{12}$$

Here, N is the number of frequency bins in FFT analysis.

4.2 LSF-GMM training and prediction

GMM is one of the most efficient methods of training in voice conversion [7]. GMM is suitable for training with huge amounts of data and is used to adapt the noise model in other unknown models in speech recognition [8]. Thus, GMM can be used to adapt the trained model of BC speech under various specific conditions to those of other unknown conditions.

We used GMM for the training phase in this study instead of SRN, which was used in the original LP scheme [4].

This section presents the procedure for training and prediction in our proposed LSF-GMM BC speech blind restoration.

4.2.1 Procedure for training

The source (BC speech) and target (clean AC speech) vector are presented in two time sequences, i.e., $X = [x_1, x_2, \cdots, x_N]$ and $Y = [y_1, y_2, \cdots, y_N]$, where N is the number of frames. The x_i and y_i are D-dimensional feature vectors for the i-th frame. For each frame of AC/BC speech, we add one average LP residual coefficient computed with Eq. (10) to the LSF vectors to compute the joint AC/BC vector. The source and target vector of each frame are therefore replaced as:

$$X_i = \left[\mathrm{LSF}_{x1}, \mathrm{LSF}_{x2}, \cdots, \mathrm{LSF}_{xP}, \overline{G_x(e^{j\omega})}\right]^T \tag{13}$$

$$Y_i = \left[\mathrm{LSF}_{y1}, \mathrm{LSF}_{y2}, \cdots, \mathrm{LSF}_{yP}, \overline{G_y(e^{j\omega})}\right]^T \tag{14}$$

Joint source-target vector $Z = [z_1, z_2, \cdots, z_n]$ where $z_q = [x_i^T, y_j^T]^T$. The distribution of Z is modeled by GMM, as in Eq. (15).

$$p(z) = \sum_{m=1}^{M} \alpha_m N(z; \mu_m, \Sigma_m) = p(x, y) \quad (15)$$

where M is the number of Gaussian components. $N(z; \mu_m, \Sigma_m)$ denotes the 2D dimension normal distribution with the mean μ_m and the covariance matrix Σ_m. α_m is the prior probability of z having been generated by component m-th, and this satisfies $0 \leq \alpha_m \leq 1$, $\sum_{m=1}^{M} \alpha_m = 1$. The parameters $(\alpha_m, \mu_m, \Sigma_m)$ for the joint density $p(x, y)$ can be estimated using the expectation maximization algorithm [7].

4.2.2 Procedure for prediction

The transformation function that converts source feature x to target feature y is given by Eq. (16).

$$\begin{aligned} F(x) &= E(y|x) = \int y p(y|x) dy \\ &= \sum_{m=1}^{M} p_m(x)(\mu_m^y + \Sigma_m^{yx}(\Sigma_m^{xx})^{-1}(x - \mu_m^x)) \end{aligned} \quad (16)$$

where

$$p_m(x) = \frac{\alpha_m N(x; \mu_m^x, \Sigma_m^{xx})}{\sum_{m=1}^{M} \alpha_m N(x; \mu_m^x, \Sigma_m^{xx})} \quad (17)$$

$$\mu_m = \begin{bmatrix} \mu_m^x \\ \mu_m^y \end{bmatrix} \quad (18)$$

$$\Sigma_m = \begin{bmatrix} \mu_m^{xx} & \mu_m^{xy} \\ \mu_m^{yx} & \mu_m^{yy} \end{bmatrix} \quad (19)$$

The $p_m(x)$ is the probability of x belonging to the m-th Gaussian component. We use Eq. (16) to predict vector $\mathbf{Y}'$ of clean AC speech from vector $\mathbf{X}'$ of BC speech. After that, we separate the LSF coefficients and the average residuals. Gain k is then computed as in Eq. (10) and the inverse filter to restore BC speech is finally computed as in Eq. (9).

We used diagonal covariance GMM in our experiments. The chosen number of Gaussian components M, which should be selected to be sufficiently large if we have sufficient data for training, was 15. The frame size was set to be large enough at 256 ms and the step was 128 ms. The use of large frames assisted our method in real-time applications. The order of LP analysis P was chosen to be 20 in all experiments.

5. Evaluations

5.1 Data preparation and experimental setup

We evaluated the proposed model for BC speech restoration in clean environment in our previous studies [4, 6]. BC speech is especially used for noisy speech environments. Therefore, we should confirm whether the LP scheme is useful for restoring BC speech in noisy environments.

Table 1 Equipment and setup for recording

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 and C	Handmade
Speakers (4 set)	JBL, CM62

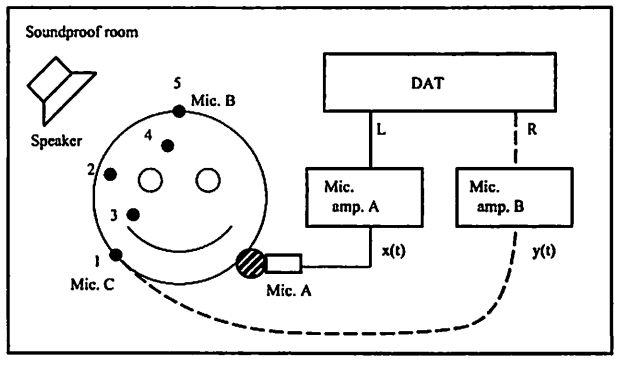


Fig. 3 Environment for recording AC/BC speech

We investigated both our LSF-SRN method [4, 6] and our proposed LSF-GMM model in both clean and noisy environments.

The speech data used in our evaluation is a familiarity-controlled Japanese speech dataset that was recorded from 17 speakers, including 10 males and 7 females. All speakers were native Japanese graduate students.

Figure 3 outlines the environment we used to construct the database. BC speech was collected at five different pick-up points on the head and face (1: mandibular angle, 2: temple, 3: philtrum, 4: forehead, and 5: calvarium). Different microphones were used at pick-up points from 1 to 4 and at pick-up point 5.

In this work, we only used BC speech that was collected at the farthest pick-up point from the mouth, i.e., pick-up point 5 (calvarium).

The microphone was positioned in front of the mouth to record AC speech. Original speech was transmitted from the mouth to the microphone through air, which is the air-conduction process. The

further the distance from the mouth to the microphone, the greater the effect the air-conduction process had on observed AC speech.

It is known that the recording environments for AC and BC speech always differ. To compare the quality of AC and BC speech, we should use AC and BC microphones that have recording environments that are as similar as possible. We defined the term observed noisy AC speech in this paper, which is noisy AC speech dependent on the environment, on the recording quality of the AC microphone, on the position of this microphone in front of the mouth, and on the noise source. We then recorded clean AC speech to train BC speech, and observed noisy AC speech in comparative testing; both were recorded closely in front of the mouth, as seen in Fig. 3.

The list of equipment and the setup are listed in Table 1.

Our dataset contains 100 Japanese words and 100 Japanese syllables from Japanese word lists in four different familiarity ranges (R1, R2, R3, and R4) [10].The noisy data contains three kinds of noise, which are factory, pink, and white noise.

Each kind of noise has three levels, in which the sound pressure level (SPL), called the noise level in this paper, is low (35 dB), medium (55 dB), and high (75 dB). It is known that the widely used signal to noise ratio (SNR) depends on the signal being investigated while SPL, which is used to describe the noise source, is independent of the signal being investigated. We wanted to control the effects of independent noise sources on AC and BC speech in our experiments; therefore, we used SPL instead of SNR.

Objective evaluations were carried out for all low, medium, and high noise levels of factory noise and subjective evaluations were only undertaken with the highest noise level, i.e., the factory noise of 75 dB.

5.2 Objective evaluations

We used the log spectral distortion (LSD), linear prediction coefficients distance (LPCD), Mel-frequency cepstral coefficients distance (MFCCD), and perceptual evaluation of speech quality (PESQ) to objectively and comparatively evaluate the proposed method. LSD was defined [4], PESQ was defined [9], LPCD, and MFCCD was defined similarly to those in Vu et al. [4].

$$\text{LPCD} = \sqrt{\sum_{i=1}^{P} (a_x(i) - a_y(i))^2} \quad (20)$$

$$\text{MFCCD} = \sqrt{\sum_{i=1}^{Q} (c_x(i) - c_y(i))^2} \quad (21)$$

where $a_x(i)$ & $a_y(i)$ and $c_x(i)$ & $c_y(i)$ are the LP coefficients and Mel-frequency cepstral coefficients (MFCC) of the source and target speech for evaluation. *The* P is the LP order and Q is the cepstral order. The LSD and PESQ are objective evaluations of voice quality for hearing while LPCD and MFCCD are objective evaluations of the voice quality for speech recognition.

From Fig. 4, we can see that non-blind LSF is basically the best method and non-restored BC is the worst. The two blind restoration methods of LSF-SRN and LSF-GMM approximately reach the restoration of LSF. This might be because we trained enough data and the predicted LSFs in blind restorations approximated the LSFs of clean AC speech.

The proposed LSF-GMM outperformed LSF-SRN as well the LTF methods in the LSD and PESQ tests related to speech quality for human hearing. The proposed LSF-GMM outperformed LSF-SRN as well the LTF methods in the LPCD and MFCCD tests related to performance in automatic speech recognition (ASR). Therefore, the proposed LSF-GMM outperformed both previous LSF-SRN and LTF methods in both human hearing (LSD and PESQ tests) and ASR systems (LPCD and MFCCD tests). Note that in our dataset, the AC and BC speech were recorded with different microphones, as in Fig. 3. The recording microphone for AC speech was a Sony C536P, which was directional and expensive, while the recording microphone for BC speech was a Temco, which was inexpensive and commercially available. The observed signals recorded with the Temco microphones were smeared due to surrounding noise while those recorded with the Sony C536P had much less smear due to surrounding noise. Thus, the quality of observed AC speech was much better than that of BC speech. In addition, our recording position for observed noisy AC speech may have been too close to the mouth and insufficient to emphasize the effect of the air-conduction process on observed AC speech. Therefore, the quality of observed noisy AC speech was better both for original and restored BC speech. This finding does not conflict with the previous results [1], in which BC speech is more stable against surrounding noise than AC speech. It only supports a supplementary ideal that BC speech is only more robust against noise than noisy AC speech under similar recording environment conditions, which have to be carefully chosen and set up.

5.3 Subjective evaluations

Due to time limitations, we only conducted a subjective evaluation, in which we evaluated the recognition scores for the LP-scheme-based methods in only high-level factory noise (75 dB).

The subjective tests were carried out with seven subjects who had normal hearing. All were native Japanese graduate students.

The speech signals of 96 Japanese syllables, extracted from our dataset, were played in random order

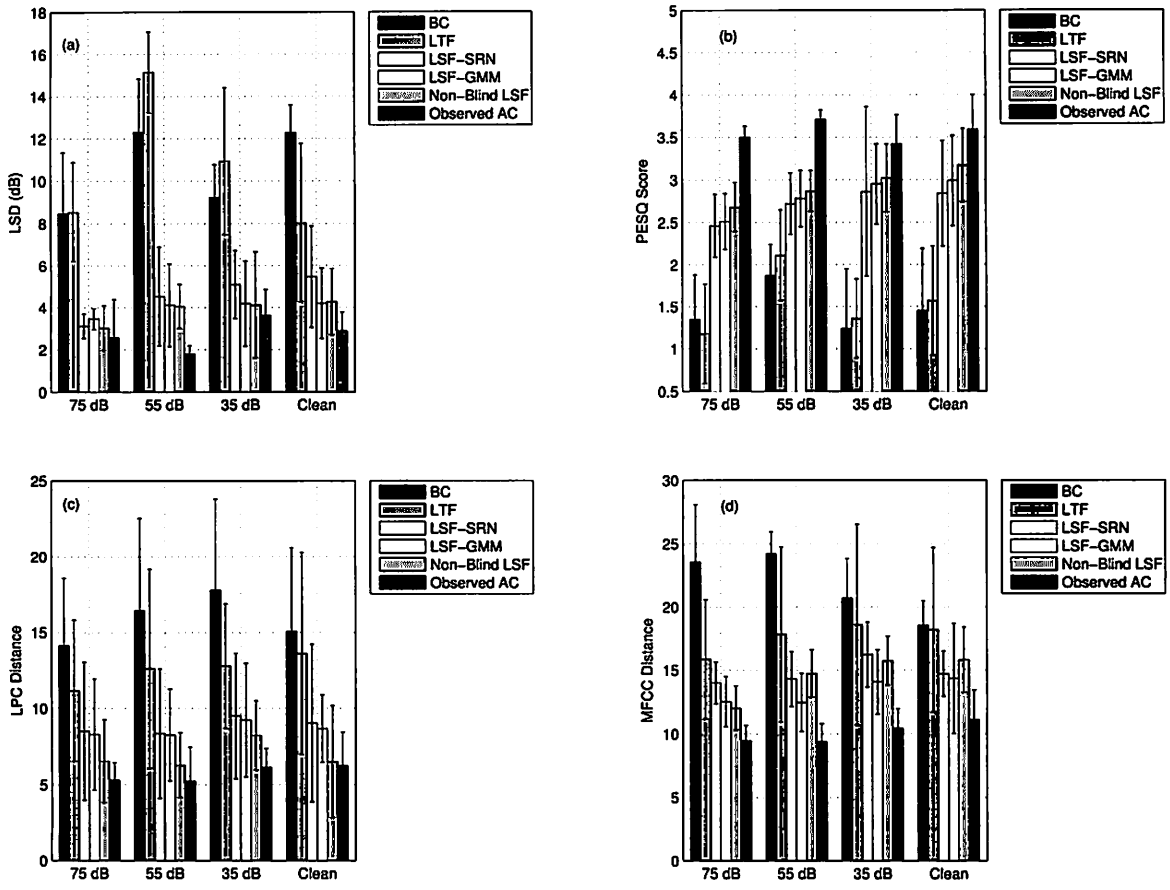


Fig. 4 Results of objective evaluation of factory noise: (a) LSD, (b) PESQ, (c) LPC distance (LPCD), and (d) MFCC distance (MFCCD)

in the tests. The subjects had not heard these syllables previously and they 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 syllables with lower familiarity. We used six types of audio: AC speech, BC speech, and four types of restored signals using the four models (LTF, blind LSF-SRN, blind LSF-GMM, and non-blind LSF). Intelligibility could generally be evaluated using the average recognition accuracy scored by all subjects.

Figure 5(a) shows the average scores for recognition accuracy under clean conditions and Figure 5(b) shows those under the noisy factory conditions of 75 dB. The non-blind LSF model was also the best for the subjective evaluation followed by the blind LSF-GMM model. The subjective evaluation confirmed that our improved method of restoring BC speech, LSF-GMM, outperformed LSF-SRN and the other previous methods of restoring BC speech.

As mentioned in previous sub-section, we comparatively evaluated our proposed method and observed noisy AC speech rather than actual high noisy AC speech. The observed noisy AC speech was recorded with an extremely high quality microphone that was too close to the mouth. Therefore, the effectiveness of BC speech in comparison with AC speech was not demonstrated in the results we obtained from our evaluation. The effectiveness of BC speech was maintained as in previous results [1], in which BC speech was more stable against surrounding noise than AC speech. If we had used actual noisy AC speech, which had been recorded with a general (non-directional) microphone far from the speaker, instead of the observed noisy AC speech in this study, the recognition rate for noisy AC would have been drastically reduced.

It is easy to see that BC speech restored with our improved method was more intelligible than the original BC speech as well as BC speech restored by other previous methods. Consequently, our improved method was robust against both degradation due to bone conduction and changing speaker pronunciations due to surrounding noise, which is referred to as the Lombard effect.

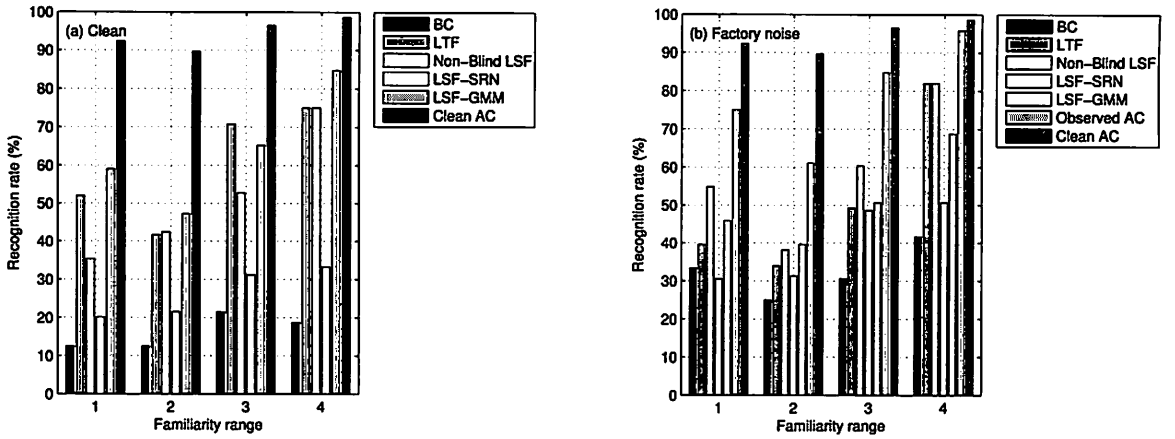


Fig. 5 Results of subjective evaluation: Scores for word recognition

6. Conclusion

We solved all three remaining problems to improve the performance of the LP scheme in restoring BC speech. Instead of using constant residual gain for all frames, we estimated the average gain for each frame. We used GMM for training instead of neural network, which made our model easier to use for training under various conditions. We also conducted experiments under both of clean and noisy environments. The experimental results indicated that our improved approach outperformed the previous methods in both human hearing quality and ASR. It also demonstrated its robustness to both degradation due to bone conduction and changing speaker pronunciations due to surrounding noise, which is referred to as the Lombard effect.

We intend to study how to adapt the trained GMM model to BC speech in various kinds of noisy environments in the future, and therefore, our proposed methods should be able to be applied to restoring open-dataset BC speech in noisy environments. We also plan to rebuild the dataset for AC/BC speech, in which there are many kinds of observed AC speech from near to far from the mouth, recorded with the similar commercially available microphones to those for BC speech, to obtain a balanced evaluation of both BC and AC speech.

Acknowledgments

This work was supported by a Grant Program made by the Yazaki Memorial Foundation for Science and Technology. It was also supported by Scope (071705001) of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1] S. Kitamori and M. Takizawa: An analysis of bone conducted speech signal by articulation tests, IEICE Trans. Vol. J72-A No. 11, pp. 1764–1771, 1989.

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

[3] T. Tamiya and T. Shimamura: Reconstruct filter design for bone-conducted speech, Proc. ICSLP 2004, Vol. II, pp. 1085–1088, 2004.

[4] T.T. Vu, K. Kimura, M. Unoki and M. Akagi: A study on restoration of bone-conducted speech with MTF-based and LP-based models, J. Signal Processing, Vol. 10, No. 6, pp. 407–417, 2006.

[5] M. Nakagiri, T. Toda, H. Kashioka and K. Shikano: Improving body transmitted unvoiced speech with statistical voice conversion, Proc. ICSLP-2006, pp. 2270–2273, 2006.

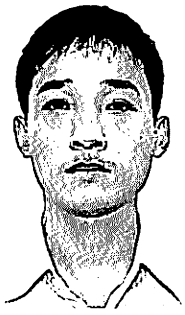
[6] T. T. Vu, G. Seide, M. Unoki and M. Akagi: Method of LP-based blind restoration for improving intelligibility of bone-conducted speech, Proc. Interspeech 2007, pp. 966–969, 2007.

[7] A. Kain and M. W. Macon: Spectral voice conversion for text-to-speech synthesis, Proc. ICASSP-1998, Vol. 1, pp. 285–288, 1998.

[8] M. Ida and S. Nakamura: HMM composition-based rapid model adaptation using a priori noise GMM adaptation evaluation on Aurora2 corpus, Proc. ICSLP-2002, pp. 437-440, 2002.

[9] Y. Hu and P. Loizou: Subjective evaluation and comparison of speech enhancement algorithms, Speech Communication, Vol. 49., pp. 588–601, 2007.

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



Phung Nghia Trung received his B.E. in electronic & telecommunication engineering from the Hanoi University of Technology in 2002 and his M.E. in electronic & telecommunication engineering from the Vietnam National University, Hanoi, in 2007. He has been with the Thai Nguyen University of Information and Communication Technology from 2003. He has also been a Ph.D. candidate at the School of Information Science of the Japan Advanced Institute of Science and Technology (JAIST) since 2009. His main research interest is speech signal processing.

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 auditory motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (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 Centre for the Neural Basis of Hearing (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 and 2010 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 of 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 of the Japan Advanced Institute of Science and Technology (JAIST) and is now a full professor. His research interests include speech perception, the modeling of speech perception mechanisms in human beings, and the signal processing of speech. During 1998, he was associated with the Research Laboratories of Electronics at MIT as a visiting researcher, 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), 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, the Best Paper Award from the Research Institute of Signal Processing in 2009, and the Sato Prize for Outstanding Papers from the ASJ in 1998, 2005, 2010 and 2011. Professor Akagi is currently the president of the ASJ.

(Received November 07, 2011; revised February 14, 2012)