SELECTED PAPER AT NCSP’20

Improvement in Bone-Conducted Speech Restoration Using Linear Prediction and Long Short-Term Memory Model

Huy Quoc Nguyen and Masashi Unoki

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa 923–1292, Japan
E-mail: s1910410, unoki@jaist.ac.jp

Abstract

Bone-conducted (BC) speech has a significant advantage as a solution for speech communication in an extremely noisy environment because of its stability against surrounding noise. However, the quality and intelligibility of BC speech degrade, making BC speech difficult to restore. To solve this problem, we propose a method for restoring BC speech with a combination of a linear prediction (LP) model using line spectral frequencies (LSFs) and a long short-term memory (LSTM) model. We evaluated the method using three objective measurements: log-spectrum distortion, LP coefficient distance, and a perceptual evaluation of speech quality. The results of all these measurements show that our method is better than the previous method, which used a simple recurrent network. These results also show that the model can yield speech with better quality when the LP gain is estimated more accurately.

1. Introduction

Speech communication in noisy environments has always been a challenge. Although many noise-cancellation models and algorithms have been introduced, they are not efficient for application in extremely noisy environments. Another solution for this problem is using a bone-conducted (BC) microphone, which records the speech signals transmitted through the speaker’s head and face. The recorded speech signals are called BC speech, and they are stable against surrounding noise. However, the quality and intelligibility of BC speech degrade in a complex manner during transmission. Therefore, attenuated BC speech signals need to be restored.

Some research has been conducted to analyze and restore BC speech. Among these studies, one milestone finding was that BC speech can be attenuated with a low-pass filter having a cut-off frequency of 1 kHz [1]. Hence, the general approach is to design an inverse filter of the attenuating low-pass filter to emphasize the deteriorating frequency components.

The earliest methods for this design were using the cross-spectrum method [2] and long-term Fourier transform method [3]. Nevertheless, these methods yield artifacts besides speech signals [1]. In later years, the modulation transfer function (MTF) and linear prediction (LP) methods [1] produced better results. However, these methods require the corresponding air-conducted (AC) speech to be simultaneously recorded in a clean environment, which is not practical. Therefore, an LSF-SRN model combining line spectral frequencies (LSFs) and a simple recurrent network (SRN) was proposed by Vu et al. [1] based on the LP method to overcome this requirement of AC speech. An improvement in the LSF-SRN model proposed by Phung et al. [4] is an LSF-GMM model, which replaces the SRN with a Gaussian mixture model (GMM).

In this paper, we propose an LSF-LSTM model, one that still uses the same kind of inverse filter as [5] and [4]. However, it uses a long short-term memory (LSTM) network instead of a GMM or SRN to estimate the LSFs of AC speech from the LSFs of BC speech. The LSTM model keeps the sequential relationship of LSF values that may be omitted by the LSF-GMM model and avoids the long-term problem of dependencies in the SRN [6].

2. Previous Method

The LP-based model and SRN model in Vu et al.’s method [5] were used to restore BC speech. The idea of the LP model is that the information corresponding to the source characteristics has the same LP residue for both AC and BC speech, while the intelligibility of speech is contained in the spectral envelope as LP coefficients. This inverse filter can be derived from the LP coefficients of AC speech and BC speech by assuming that the conversion of BC speech to AC speech includes an infinite impulse response (IIR). Thus, it is called an LP-based filter. However, the information on the AC speech is not provided in practice, so Vu et al. proposed using the SRN model to predict the LSF presentation of LP coefficients.

2.1 LP-based filter

The LP model assumes that the future values of a discrete-time signal can be estimated as a linear function of a specific
number of previous samples:

\[ s_n = g_n + \sum_{i=1}^{P} a_i s_{n-i} \]  

(1)

where \( s_n \) is the speech signal, \( P \) is the order of prediction, and \( a_1, a_2, \ldots, a_P \) are the LP coefficients (LPCs). They are chosen such that the error \( E = \sum g_n^2 \) is minimal.

By applying LP to both AC speech signal \( x_n \) and BC speech signal \( y_n \), the residual signals of AC speech and BC speech in the \( z \)-domain are

\[-G_X(z) = X(z) \sum_{i=0}^{P} a_X z^{-i} \]  

(2)

and

\[-G_Y(z) = Y(z) \sum_{i=0}^{P} a_Y z^{-i} \]  

(3)

The transfer function of inverse filter converting \( y_n \) to \( x_n \) can be derived as

\[ H^{-1}(z) = \frac{X(z)}{Y(z)} \frac{G_X(z) \sum_{i=0}^{P} a_Y z^{-i}}{G_Y(z) \sum_{i=0}^{P} a_X z^{-i}} \]  

(4)

By assuming that \( H^{-1}(z) \) is an IIR filter, \( k = \frac{G_X(z)}{G_Y(z)} \) can be assumed to be constant with respect to \( z \). Also, \( k \) is assigned to 1 in [5].

2.2 SRN predictor on LSF presentation

2.2.1 LSF presentation of LPC

The LSFs proposed by [7] are a sequence of frequency values that presents LPCs without loss and are better than the LPCs in quantization [7]. LSFs are derived from the roots of two \( (P+1) \)-th order polynomials \( U(z) \) and \( V(z) \) defined as

\[ U(z) = \sum_{i=0}^{P} a_i z^{-i} + z^{-(P+1)} \sum_{i=0}^{P} a_i z^{-i} \]  

(5)

\[ V(z) = \sum_{i=0}^{P} a_i z^{-i} - z^{-(P+1)} \sum_{i=0}^{P} a_i z^{-i} \]  

(6)

Because the roots of \( U(z) \) and \( V(z) \) are located in an interlaced order on the unit circle and symmetric for the real axis in the \( z \)-domain [7], the roots of \( U(z) \) and \( V(z) \) can be presented as a sequence of angles in range \( (0, \pi) \)

\[ \theta = (\theta_1, \theta_2, \ldots, \theta_P) \]  

(7)

which are the phases of the roots of \( U(z) \) and \( V(z) \) except the trivial roots \( z = 1 \) and \( z = -1 \).

The coefficients \( a_X \) and \( a_Y \) can be presented as \( \theta_X \) and \( \theta_Y \), respectively, in applying this technique to the BC speech restoration problem.

2.2.2 SRN predictor

In Vu et al.’s method, a simple recurrent network (SRN) with an Elman structure [8] is used to train the prediction model to estimate \( \theta_X \) when \( \theta_Y \) is given. An SRN is a neural network architecture belonging to a recurrent neural network family, which allows predicting the sequential output from the sequential input. However, a recurrent neural network has something called the long-term dependencies problem. We propose the LSF-LSTM model presented in the next section to solve it.

3. Proposed Method: LSF-LSTM Model

We propose the LSF-LSTM model, which is based on the method of Vu et al. [5]. The block diagrams of the LSF-LSTM model are described in figure 1. We improved Vu et al.’s method by focusing on the prediction phase. First, the SRN model was replaced with an LSTM model. LSTM [6] is a special kind of recurrent network that can solve the long-term dependencies problem of the SRN using a cell state that only changes slightly through sequence processing. Elman’s structure is used in the same way as [5] with one hidden layer of \( L \) units. Second, LSF is directly used instead of the LSF differentials \( \Delta \theta_k = \theta_k - \theta_{k-1} \) because the prediction of LSF differentials can cause cumulative error in the prediction of high frequencies.

The structure of the LSF-LSTM model to apply in the BC speech restoration problem is illustrated in figure 2. The model receives input sequence \( \theta_Y k \) and estimates the output sequence \( \theta_X k \), where \( \theta_Y k \) and \( \theta_X k \) are the LSF of the BC speech and AC speech normalized to range \( (0, 1) \). Each LSTM unit of hidden layer is computed as follows.

- Compute the amount of information for the cell state to forget:
  \[ f_k^{(l)} = \sigma \left( W_f^{(l)} \left[ \theta_X k_{l-1} \theta_Y k \right] + b_f^{(l)} \right) \]  

(8)

- Compute the new information generated at step \( k \):
  \[ i_k^{(l)} = \sigma \left( W_i^{(l)} \left[ \theta_X k_{l-1} \theta_Y k \right] + b_i^{(l)} \right) \]  

(9)

- Compute the amount of new information to be added to the cell state:
  \[ \tilde{C}_k^{(l)} = \sigma \left( W_C^{(l)} \left[ \theta_X k_{l-1} \theta_Y k \right] + b_C^{(l)} \right) \]  

(10)

- Compute the new value of the cell state at step \( k \):
  \[ C_k^{(l)} = C_{k-1}^{(l)} f_k^{(l)} + \tilde{C}_k^{(l)} i_k^{(l)} \]  

(11)

- Compute the amount of information in the cell state that will be used for the output:
  \[ o_k^{(l)} = \sigma \left( W_o^{(l)} \left[ \theta_X k_{l-1} \theta_Y k \right] + b_o^{(l)} \right) \]  

(12)
2.2.1 LSF presentation of LPC

Because the roots of $X(z)$ and $Z(z)$ can be presented as polynomials of $z^{-1}$, a sequence of frequency values $\theta \in \mathbb{C}^N$, where $N$ is the number of previous samples, can be assumed to be constant with respect to $\theta$. Therefore, $X(z)$ can be computed as follows:

$$X(z) = \sum_{n=0}^{N-1} X_n z^{-n}$$

Using Elman’s structure, the output $\theta_{X_k}$ combines all of the following outputs of the hidden units:

$$\theta_{X_k} = \sigma \left( W_e \left[ \theta_{X_1}^{(1)} \ldots \theta_{X_k}^{(L)} \right] + b \right)$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

is the sigmoid function. The sigmoid function is used because its range of values is from 0 to 1, which is suitable for the normalized output sequence.

4. Evaluations

To evaluate our method, we applied the restoration methods to a database of AC and BC speech recorded simultaneously. The detailed information on the data is described in Table 1. In addition, the LP order $P = 20$ was used for both the LP-based and LSF-LSTM methods, the length of each frame was 128 ms, the overlapping time was 64 ms, and the number of units in the hidden layer was 20 for the LSF-LSTM method.

![Figure 1: Block diagrams of LSF-LSTM model](image)

![Figure 2: LSTM predictor for LSF with Elman’s structure](image)

Table 1: Data information

| Measurement site       | Soundproof room   |
|------------------------|-------------------|
| Recorder               | SONY, TCD-D10 Proll |
| Sampling frequency     | 16 kHz            |
| Mic. A for AC speech   | SONY, CS36P       |
| Mic. B for BC speech   | Temco, HG-17      |
| Measurement position   | Calvarium         |
| Number of words        | 100 words         |
| Total number of speakers | 11 people       |
| Number of speakers for training | 10 people     |
| Number of speakers for testing | 1 person        |

In this experiment, we evaluated the results using three objective measurements: log-spectrum distortion (LSD), linear prediction coefficient distance (LCD) and perceptual evaluation of speech quality (PESQ) [9]. LSD and LCD are computed as follows:

$$\text{LSD} = \left( \frac{1}{W} \sum_{\omega} \sum_{n} 20 \log_{10} \left( \frac{|X(\omega)|}{|Y(\omega)|} \right) \right)^2$$

and

$$\text{LCD} = \frac{1}{P} \sum_{i=1}^{P} (a_{X_i} - a_{Y_i})^2$$

where $X(\omega)$ and $Y(\omega)$ are correspondingly the 1024-point fast Fourier transform calculation of each 25-ms frame of AC speech and BC speech and where the time these frames overlap is 15 ms; $a_{X_i}$ and $a_{Y_i}$ are the LPCs with 20 coefficients. PESQ is a well-known speech quality measurement introduced by Hu and Loizou [9], and it enables assessing the quality of speech by comparing it with the original reference speech signal (AC speech signals in this context). The range value of the PESQ score is from $-0.5$ (bad) to $4.5$ (no distortion).

These objective measurements were applied to compare the speech restored using our method with ones using the LSF-SRN model, the LP non-blind restoration model, and with the original BC speech. The evaluation results presented in Table 2 show that the restored speech using the LSF-LSTM model...
Table 2: Objective evaluations results for $k = 1$

|                      | LSD   | LCD   | PESQ  |
|----------------------|-------|-------|-------|
| Original BC Speech   | 15.4263 | 0.1635 | 2.3703 |
| LP (non-blind, $k = 1$) | 16.3750 | 0.1069 | 2.2795 |
| LSF-SRN              | 16.7704 | 0.2480 | 1.7823 |
| LSF-LSTM             | 14.8796 | 0.1969 | 2.0133 |

Table 3: Objective evaluations results for estimated $k$

|          | LSD   | LCD   | PESQ  |
|----------|-------|-------|-------|
| LP       | 8.6662 | 0.1131 | 2.7917 |
| LSF-LSTM | 13.8679 | 0.1974 | 2.0921 |

is better than that using the LSF-SRN model in all three measurements. Also, it is closer to the LP model, which is the baseline model. However, the results of our method are still worse than the original BC speech, and we suspect the reason comes from the value of $k$. To verify this, we did an analysis experiment, applying a non-blind estimation of $k$ as

$$k = \frac{1}{W} \sum_{\omega} W \frac{G_X(e^{i\omega})}{G_Y(e^{i\omega})}$$

The results of this experiment are in table 3. They show that the distortion was reduced and that the quality of speech was improved when $k$ was more precisely estimated.

5. Conclusions

In this paper, we proposed the LSF-LSTM model, which derives the inverse transfer function to transform BC speech to AC speech based on the LP model. The method does not require the AC speech signal during the prediction process. We found that the LSF-LSTM method gives better performance than the LSF-SRN method by applying small-scale data including 11 speakers and 100 words. In addition, the results also show that the model improves as $k$ is estimated more precisely. These results are the first steps toward building a BC speech restoration system using a neural-network-based generative model, which is the LSTM in this paper. As our future work involves designing a better restoration system, we need to build a better generative model that reaches the baseline results of the LP model and that yields a better estimation of $k$. In addition, because neural networks are data-driven models, we must extend the experiment using much more data.

Acknowledgment

This work was supported by a Grant in Aid for Innovative Areas (No. 16H01669, No. 18H05004) from MEXT, Japan. This work was also supported by JST-Mirai Program (Grant Number: JPMJMI18D1).

References

[1] 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.
[2] S. Ishimitsu, H. Kitakaze, Y. Tsuchibushi, H. Yanagawa and M. Fukushima: A noise-robust speech recognition system making use of body-conducted signals, Acoustical Science and Technology, Vol. 25, No. 2, pp. 166–169, 2004.
[3] T. Tamiya and T. Shimamura: Reconstruction filter design for bone-conducted speech, Proc. ICSP2004, pp. 1085–1088, 2004.
[4] N. Phung, M. Unoki and M. Akagi: A study on restoration of bone-conducted speech in noisy environments with LP-based model and Gaussian mixture model, Signal Processing, Vol. 16, No. 5, pp. 409–417, 2012.
[5] 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.
[6] S. Hochreiter and J. Schmidhuber: Long short-term memory, Neural Computation, Vol. 9, No. 8, pp. 1735–1780, 1997.
[7] F. Itakura: Line spectrum representation of linear predictor coefficients of speech signals, J. Acoustical Society of America, Vol. 57, No. 5, pp. 533–535, 1975.
[8] J. L. Elman: Finding structure in time, Cognitive Science, Vol. 14, No. 2, pp. 179–211, 1990.
[9] Y. Hu and P. Loizou: Subjective comparison and evaluation of speech enhancement algorithms, Speech Communication, Vol. 49, pp. 588–601, 2007.