Linear-Prediction-Based Accurate Spectrum Estimation with Pitch Extension for Bone-Conducted Speech

Md Arifour Rahman and Tetsuya Shimamura

Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama 338-8570, Japan
E-mail: {arifour, shima}@sie.ics.saitama-u.ac.jp

Abstract This paper proposes an approach to pitch-synchronous linear prediction (LP) for bone-conducted (BC) voiced speech. A combination of the spectral compensation (SC) method with pitch extension LP is used to obtain a more accurate power spectrum of the BC speech signal. Simulation experiments show that the proposed method provides better performance than the conventional autocorrelation and original SC methods.

Keywords: linear prediction, autocorrelation method, prediction error filter, spectrum compensation, air-conducted speech, bone-conducted speech, pitch-synchronous analysis.

1. Introduction

Recently, a great deal of attention has been paid to the use of bone-conducted (BC) speech in speech processing [1]-[9]. This is because BC speech is obtained through a BC microphone touching the body, resulting in a noise-robust speech signal even in noisy environments. It is known that BC speech usually has amplitude attenuation at high frequencies of more than 1 kHz [2]. For this reason, much effort has been devoted to restoration of the high-frequency components of BC speech [1]-[7]. On the other hand, although pitch detection was discussed in [8], there are few works related to the analysis of BC speech.

Several spectral analysis techniques exist such as the periodogram, linear prediction (LP), and cepstrum method [10]. In particular LP has been widely used in a variety of speech-processing systems such as for coding, recognition and enhancement [10]. LP can be applied to BC speech as well as air-conducted (AC) speech obtained through a standard microphone. (normal speech is sometimes called air-conducted (AC) speech to contrast it with BC speech). Actually, LP has been utilized to restore the high-frequency components of BC speech [7]. However, when the autocorrelation method for LP is applied to the BC speech input, the condition of the autocorrelation matrix is an important factor determining the accuracy of the method. An ill-conditioned matrix can cause numerical problems in the process of LP [11]. The spectral dynamic range is one measure of the ill-conditioning of the correlation matrix [12]. BC speech has an expanded spectral dynamic range compared with the corresponding AC speech [9]. We should consider the ill-conditioned problem to solve for the LP coefficients in the normal equation. To this end, in [9], a spectral compensation (SC) method was derived and its satisfactory performance for BC speech was demonstrated.

In this paper, a pitch extension version of the SC method is developed and applied to BC voiced speech. The idea of pitch extension originates from Paliwal and Rao[13] and Barnwell [14]. The pitch extension approach requires information of the pitch (inverse of fundamental frequency) of speech signals. While we cannot avoid pitch detection errors when implementing pitch detectors in normal speech [15] [16], pitch could be detected accurately from BC speech as suggested in [8]. This property of BC speech is very beneficial for performing pitch-synchronous analysis such as a pitch extension approach to BC voiced speech. From this viewpoint, a pitch-extension-based SC method for BC voiced speech is derived and discussed in this paper. In the proposed pitch-extension-based SC method, prediction error filtering is first implemented, which provides two main advantages of the proposed method: (1) avoiding the ill-conditioning problem of the correlation matrix of LP and (2) improving the detection accuracy of the pitch. Both advantages result in accurate power spectrum estimation of BC speech. This may be directly and indirectly useful for applications such as spectral analysis [8], speech enhancement [7], speech recognition [17] [18], and speaker recognition [19] [20].
This paper is organized as follows. We describe the proposed pitch-extension-based SC method and discuss its properties in Sect. 2. Section 3 is devoted to experimental results on synthetic and real BC speech, where the performance of the proposed method is investigated in a comparative manner. Finally, a conclusion is drawn in Sect. 4.

2. Proposed Method

In this section, the detail of the algorithm for the proposed pitch-extension-based SC method is first described in Part A. Then, in Part B the properties of the proposed method are discussed.

2.1 Algorithm

The algorithm consists of the following steps.

Step 1: Detect a frame with $N$ length of rectangular window as $s(n), n = 0, ..., N - 1$.

Step 2: Calculate the autocorrelation function of $s(n)$ as

$$R_s(m) = \frac{1}{N} \sum_{n=0}^{N-1} s(n)s(n + m) \quad (1)$$

Step 3: Considering the Toeplitz matrix with a size of $L + 1$ given as

$$R_s = \begin{bmatrix} R_s(0) & R_s(1) & ... & R_s(L - 1) & R_s(L) \\ R_s(1) & R_s(0) & ... & R_s(L - 2) & R_s(L - 1) \\ ... & ... & ... & ... & ... \\ R_s(L - 1) & R_s(L - 2) & ... & R_s(1) & R_s(0) \end{bmatrix}$$

where $L$ is the prediction order, we implement the Levinson-Durbin algorithm and obtain the prediction coefficients $a_i, i = 1, 2, ..., L$, and the prediction error filter output as

$$x(n) = \sum_{i=0}^{L} a_i s(n - i) \quad (3)$$

where $a_0 = 1$ is assumed.

Step 4: Find the pitch period $T_0$ by applying any pitch detection algorithm to $x(n), n = 0, ..., N - 1$.

Step 5: With the estimate of $T_0$, obtain a one-pitch-period signal, as shown in Fig. 1, where it is assumed that the maximum sample point in the frame is the starting sample point of one pitch period. If the number of samples after the maximum sample point is not sufficient for $T_0$, the previous pitch period is also used.

The samples of the resulting one-pitch-period signal, $x(n_0), x(n_0 + 1), ..., x(n_0 + T_0 - 1)$, are expressed as $y(n), n = 0, 1, ..., T_0 - 1$, here. Thus, $y(n)$ has pitch period length $T_0$.

Step 6: Calculate the autocorrelation function of $y(n)$ as

$$R_y(m) = \frac{1}{T_0} \sum_{n=0}^{T_0-1} y(n)y(n+m) \quad (4)$$

where $y(n)$ is the corresponding two pitch-length-signal of $y(n)$ and represented as

$$y(n) = \begin{cases} y(n), & n = 0, 1, ..., T_0 - 1 \\ y(n - T_0), & n = T_0, T_0 + 1, ..., 2T_0 - 1 \end{cases} \quad (5)$$

In Fig. 2, $y(n)$ is shown graphically.

Step 7: Considering the Toeplitz matrix with a size of $M + 1$ given as

$$R_y = \begin{bmatrix} R_y(0) & R_y(1) & ... & R_y(M - 1) & R_y(M) \\ R_y(1) & R_y(0) & ... & ... & ... \\ ... & ... & ... & ... & ... \\ R_y(M - 1) & R_y(M - 2) & ... & R_y(1) & R_y(0) \end{bmatrix} \quad (6)$$

Fig. 1 Extraction of pitch period from the framed signal

Fig. 2 Two-pitch-length signal
where \( M \) is the prediction order, we implement the Levinson-Durbin algorithm and obtain the prediction coefficients \( b_i, i = 1, 2, ..., M \), and the prediction error power \( \sigma_y^2 \).

Step 8: Find an estimate of the compensated power spectrum as the final output

\[
P_c(\omega) = \frac{P_y(\omega)}{A(\omega)}
\]  

(7)

where

\[
P_y(\omega) = \frac{\sigma_y^2}{|1 + \sum_{i=1}^{M} b_i e^{-j\omega}|^2}
\]  

(8)

\[A(\omega) = |\sum_{i=0}^{L} a_i e^{-j\omega}|^2.
\]  

(9)

In Eqs. (8) and (9), \( \omega \) is the continuous angular frequency, to be divided into frequency bins by a discrete Fourier transform in practice. The parameters in Eq. (8), \( \sigma_y^2 \) and \( b_i \), and those in Eq. (9), \( a_i \), are obtained in Steps 7 and 3, respectively.

### 2.2 Properties

The algorithm described in Sect. 2.1 is an extended version of the SC method derived in [9], where the pitch extension LP is used instead of the autocorrelation method. In [9], the autocorrelation method for LP is used twice to obtain both the numerator and denominator in Eq. (7). In the proposed method, the numerator in Eq. (7) is obtained by the pitch extension LP method. The properties of the proposed pitch-extension-based SC method are given below.

1) The power spectrum of \( s(n) \), \( P_s(\omega) \), has a wide dynamic range of amplitude when \( s(n) \) is a BC speech signal. In such a case, the direct use of the autocorrelation method for LP suffers from numerical problems. Ill-conditioning of the correlation matrix, which is induced by the wide dynamic range of the power spectrum of the input signal, leads to an inaccurate estimate of LP coefficients. Therefore, the LP results for BC speech may be inaccurate. However, utilizing the proposed method, the problem is solved. For the prediction error filtering Eq. (3) in the proposed method, the prediction coefficients \( a_i \), are decided in a numerically stable form because the prediction order \( L \) is set to a low order. Then the prediction error filter significantly compresses the dynamic range of the power spectrum of the input BC speech. Therefore, the following LP analysis does not suffer from the ill-conditioning problem of the correlation matrix, resulting in an accurate estimate of the power spectrum \( P_s(\omega) \).

2) Usually, in the spectrum analysis of speech signals based on LP, one uses a window function, \( w(n) \), such as a Hamming or Hanning window, not a Rectangular window. This is because an overlapping operation is required to deal with a continuous speech signal. A Hamming window (or Hanning window) produces an almost flat time domain waveform in the case of half overlapping and sufficiently suppresses the side-lobe of the spectrum. However, a convolved spectrum such as \( W(\omega) \ast S(\omega) \) in the frequency domain, where \( W(\omega) \) and \( S(\omega) \) correspond to the Fourier transforms of \( w(n) \) and \( s(n) \), respectively, and \( \ast \) denotes convolution, has to be considered. We cannot obtain the true power spectrum of \( s(n) \) from \( W(\omega) \ast S(\omega) \).

In the proposed method, a rectangular window is used in Step 1. The one-pitch-period signal \( y(n) \) is detected and utilized in Step 6. Here let us assume that the periodic signal includes \( y(n) \) as follows:

\[
y_{(r)}(n) = \begin{cases} 
y(n), & n = 0, 1, ..., T_0 - 1 
y(n - T_0), & n = T_0, T_0 + 1, ..., 2T_0 - 1 
\vdots 
y(n - (r - 1)T_0), & n = (r - 1)T_0, 
(r - 1)T_0 + 1, ..., rT_0 - 1 
\end{cases}
\]

(10)

where \( r \) is an integer that can include positive and negative values. The autocorrelation function of \( y_{(r)}(n) \) is calculated for an infinite length of data as

\[
R(m) = \lim_{r \to \infty} \frac{1}{2rT_0} \sum_{n=-(rT_0-1)}^{rT_0-1} y_{(r)}(n)y_{(r)}(n + m)
\]

(11)

However, Eq. (11) reduces to

\[
R(m) = \frac{1}{T_0} \sum_{n=0}^{T_0-1} y_{(1)}(n)y_{(2)}(n + m)
\]

(12)

\[
= \frac{1}{T_0} \sum_{n=0}^{T_0-1} y(n)y_{(2)}(n + m)
\]

(13)

where \( y(n) = y_{(1)}(n) \). Equation (13) is equivalent to Eq. (4). This means that in the proposed method, an infinite number of data samples are considered, resulting in an accurate estimate of \( P_s(\omega) \). Therefore, the proposed method has the potential to provide the true power spectrum of \( s(n) \), \( P_s(\omega) \), if \( s(n) \) is a stationary periodic signal, that is, \( P_s(\omega) = P_s(\omega) \). This property of unbiased estimation for the power spectrum of the input signal was not discussed in the original papers on the pitch extension method [13][14].

In [21], the use of a rectangular window for the autocorrelation method for LP was discussed in a comparative manner with that of a Hamming or Hanning window. A rectangular window does not decrease the effective
number of data samples but suffers from bias propagation in the Levinson-Durbin algorithm, resulting in a poor performance for spectral estimation. In the proposed method, rectangular windowing is used. However, the proposed method does not suffer from bias propagation in the Levinson-Durbin algorithm by utilizing the periodicity of the input speech signal.

3) The power spectrum of the input BC speech is accurately obtained by (7) in a spectrum compensation form. The pre-emphasis approach [11], which is often used in conventional spectral analysis, usually does not consider spectrum compensation, in contrast to the proposed method, to obtain the power spectrum of the input speech signal.

4) In the proposed method, pitch detection is required in Step 4. Here, accurate pitch information must be obtained. BC speech includes fewer high-frequency components than the corresponding AC speech. This feature of BC speech is inherently suitable for pitch detection because the pitch detection results are often affected by high-frequency components. Accurate estimate of the pitch is expected from BC speech. Owing to the above feature of BC speech, the pitch information from BC speech can be used as the true pitch for AC speech [8].

In addition to the above feature of BC speech, in the proposed method, the prediction error filter is used as a pre-processing filter for pitch detection. Usually, the formants of the vocal tract alter the harmonic structure of the input speech signal, making the actual pitch period difficult to detect [22]. The use of a prediction error filter before the pitch detection inherently improves pitch detection by producing a flattened spectrum of the input speech signal [23], [24]. In the proposed method, this merit is directly obtained.

3. Experiments

To validate the performance of the proposed pitch-extension-based SC method, we conducted experiments. First, we used synthetic BC speech signals. Then, we evaluated the method using real BC speech signals.

A. Synthetic BC speech

We generated BC synthetic speech by low-pass filtering AC synthetic speech. The AC synthetic speech signal $s_a(n)$ was generated using the all-pole filter [13]:

$$V(z) = \frac{A_v}{1 + \sum_{k=1}^{K} c_k z^{-k}}$$

(14)

where the gain factor $A_v$, and the filter coefficients $c_k$, $k = 1, 2, ..., K$, are known a priori. $K$ corresponds to the filter order. Table 1 shows the list of the filter coefficients for the five vowels we used. The gain factor was fixed 0.1354. The sampling frequency was 10 kHz and a periodic train of impulses was used for the excitation of the source. The AC synthetic speech signal $s_a(n)$ was then passed through an infinite impulse response (IIR) low-pass filter to obtain the BC synthetic speech.

$$s_0(n) = 0.8s_a(n - 1) + 0.25s_a(n)$$

(15)

The coefficients of 0.8 and 0.25 in (15) were decided by considering the spectrum gap between the AC and BC speech as shown in [9]. The length of each synthetic BC speech was 5 s. Figure 3 shows the power spectrum of the synthetic BC speech /u/ with $F_0 = 100$ [Hz] as an example. In Figure 3, we can observe that the spectral dynamic range of the synthetic BC speech signal is wider than that of the synthetic AC speech signal.

![Fig. 3 Power spectra of synthetic AC and BC speech signals](image)

We compared the performance of the proposed method to validate the performance of the proposed pitch-extension-based SC method, we conducted experiments. First, we used synthetic BC speech signals. Then, we evaluated the method using real BC speech signals.

| $c_k$ | /a/ | /i/ | /u/ | /e/ | /o/ |
|------|-----|-----|-----|-----|-----|
| $c_1$ | -0.22181 | 0.82995 | -0.99116 | -0.92965 | -0.87015 |
| $c_2$ | 0.38305 | -0.33920 | 0.52552 | 1.21137 | 0.67042 |
| $c_3$ | 0.12125 | -1.61389 | -0.92721 | -0.88498 | -0.57341 |
| $c_4$ | 0.14342 | -1.23332 | 1.12357 | 0.65533 | 0.70213 |
| $c_5$ | -0.01168 | 0.00932 | -1.09906 | -0.74714 | -0.88725 |
| $c_6$ | 0.15431 | 1.31354 | 1.14719 | 0.46214 | 0.64263 |
| $c_7$ | 0.19194 | 1.08109 | -0.81144 | -0.29582 | -0.39441 |
| $c_8$ | 0.18992 | 0.12361 | 0.05872 | 0.42319 | 0.32445 |
| $c_9$ | 0.05973 | -0.55380 | -0.26298 | -0.13810 | -0.39001 |
| $c_{10}$ | 0.64174 | -0.17653 | 0.58394 | 0.54478 | 0.64581 |

280 Journal of Signal Processing, Vol. 22, No. 6, November 2018
Fig. 4 Power spectra of synthetic BC speech /a/

Fig. 5 Power spectra of synthetic BC speech /i/

Fig. 6 Power spectra of synthetic BC speech /u/

Fig. 7 Power spectra of synthetic BC speech /e/

with that of the SC method [9] as well as the standard autocorrelation method [25]. The comparison was first made by visual inspection. Figures 4 to 8 show the power spectra of the synthetic BC speech (for the five vowels). In each figure, the FFT spectrum (periodogram), the true power spectrum transformed from the generated model, and the estimates by the autocorrelation, SC, and proposed methods from one frame are shown, which are denoted by Synthetic BC Speech, True Spectrum, Autocorrelation Method, SC Method, and Proposed Method, respectively.

The true spectrum was computed as

$$P_{tr}(\omega) = |H(e^{j\omega})|^2 P_a(\omega)$$  \hspace{1cm} (16)

where $P_a(\omega)=|V(e^{j\omega})|^2$ is the power spectrum of the generated AC speech signal ($A_v=0.1354$ and $c_k$ in Table 1 are used here) and $H(e^{j\omega})$ corresponds to the frequency response of the transfer function $H(z)$ of the IIR filter in (15). Each frame consisted of 400 samples. For the proposed method, the pitch period was estimated using the autocorrelation function method [26]. The prediction orders $M$ and $L$ were 12 and 2, respectively, for the proposed method. We used $M=14$ for the autocorrelation method to make a fair comparison with the proposed method. For the true spectrum and the autocorrelation, SC, and proposed methods, we calculated the power spectrum from the prediction coefficients using the fast Fourier transform (FFT) whose number of FFT points was 1024. For the FFT spectrum, the same number of FFT point was also used. From these figures, it is clearly observed that the proposed method generally produces the closest spectral shape to the true one. Note that the true spectrum calculated by (16) is generally most closely matches to the envelope of the FFT spectrum in Figures 4-8.
Fig. 8 Power spectra of synthetic BC speech /o/

Next, log-spectrum distortion (LSD) was used for the objective evaluation of the proposed method. In the experiments, the LSD was computed, which is given as

$$LSD = \sqrt{\frac{1}{Q} \sum_{t=1}^{Q} \left[ 10 \log_{10} \left( \frac{P_{tr}(\omega)}{\hat{P}(\omega)} \right) \right]^2}$$ (17)

where $Q$ is the upper-frequency bin number (half of the number FFT points, 512) and $P_{tr}(\omega)$ and $\hat{P}(\omega)$ are the true power spectrum and estimated power spectrum, respectively. We calculated the average for 30 frames of each vowel. Table 2 lists the improvement of the proposed method relative to the autocorrelation and SC methods, where a lower value of LSD indicates that the estimated spectrum is closer to the true one. From Table 2, it can be observed that for all vowels the proposed method provides lower LSD values than the autocorrelation and SC methods.

In order to optimize the order $L$ of the prediction error filter in the proposed method, we investigated the average LSD distances of five vowels from 30 frames of each vowel for the proposed method using different values of $L$. Table 3 shows the resulting LSD distances. From Table 3, it can be seen that the case of $L = 2$ provides the minimum distances. This result indicates that the performance of the proposed method is best at $L = 2$. We also investigated the LSD distances of the autocorrelation method for different orders $M$ as shown in Table 4. The autocorrelation method provided almost the smallest LSD distances for $M = 14$, which did not change significantly for $M \geq 12$. Also, the LSD distance was very high for $M < 12$. Therefore $M = 14$ was chosen for the autocorrelation method.

| Vowel | $L=1$ | $L=2$ | $L=3$ | $L=4$ |
|-------|-------|-------|-------|-------|
| /a/   | 0.25  | 0.25  | 0.27  | 0.28  |
| /i/   | 0.34  | 0.30  | 0.33  | 0.33  |
| /u/   | 0.31  | 0.29  | 0.28  | 0.30  |
| /e/   | 0.28  | 0.27  | 0.29  | 0.31  |
| /o/   | 0.27  | 0.25  | 0.25  | 0.28  |
| Average| 0.29  | 0.27  | 0.28  | 0.30  |

Table 2 | LSD distance for five synthetic vowels

| Vowel | Autocorrelation Method | SC Method | Proposed Method |
|-------|------------------------|-----------|-----------------|
| /a/   | 0.41                   | 0.35      | 0.25            |
| /i/   | 0.45                   | 0.42      | 0.30            |
| /u/   | 0.60                   | 0.40      | 0.29            |
| /e/   | 0.45                   | 0.36      | 0.27            |
| /o/   | 0.43                   | 0.32      | 0.25            |
| Average| 0.47                 | 0.37      | 0.27            |

Table 3 | LSD distances of proposed method for different orders of prediction error filter

| Vowel | $M=10$ | $M=12$ | $M=14$ | $M=16$ |
|-------|--------|--------|--------|--------|
| /a/   | 1.07   | 0.43   | 0.41   | 0.41   |
| /i/   | 0.52   | 0.46   | 0.45   | 0.43   |
| /u/   | 0.71   | 0.61   | 0.60   | 0.60   |
| /e/   | 1.05   | 0.43   | 0.45   | 0.46   |
| /o/   | 1.85   | 0.47   | 0.43   | 0.43   |
| Average| 1.04   | 0.48   | 0.47   | 0.47   |

Table 4 | LSD distances of autocorrelation method for different LP orders

| Vowel | $M=12$ | $M=14$ | $M=16$ |
|-------|--------|--------|--------|
| /a/   | 48.33  | 36.05  | 23.12  |
| /i/   | 80.09  | 66.95  | 52.03  |
| /u/   | 103.71 | 88.81  | 62.19  |
| /e/   | 101.50 | 82.17  | 51.72  |
| /o/   | 82.69  | 70.03  | 40.95  |
| Average| 83.26  | 68.80  | 46.00  |

Table 5 | Condition number for different methods

| Vowel | Autocorrelation Method | SC Method | Proposed Method |
|-------|------------------------|-----------|-----------------|
| /a/   | 48.33                  | 36.05     | 23.12           |
| /i/   | 80.09                  | 66.95     | 52.03           |
| /u/   | 103.71                 | 88.81     | 62.19           |
| /e/   | 101.50                 | 82.17     | 51.72           |
| /o/   | 82.69                  | 70.03     | 40.95           |
| Average| 83.26                 | 68.80     | 46.00           |
When we implemented the proposed method, the number of samples in one period was varied from 100 to 101 for the fundamental frequency $F_0=100$ Hz. The variation was 7% among the above five vowels and usually the number of samples was 100, which was the true pitch length. Also, the average LSD distance was 0.27. For $F_0=250$ Hz, the number of samples was varied from 39 to 40 and the average LSD distance was 0.35. The variation was 9%. Although the variation in the number of samples was the same for both $F_0=100$ Hz and $F_0=250$ Hz, actually the difference was only one sample. In the frequency domain the pitch estimation error and the resulting LSD distance were higher for high-pitch speech. The LSD distances of the proposed method were increased for high-pitch speech. For $F_0=400$ Hz, the number of samples was varied from 24 to 25 and the average LSD distance was 0.39. This variation was 12%. The performance of the proposed method, however, may be
increased by using other estimation methods such as those in [27][28], which are robust against high-pitch speech.

B. Real BC speech

After the investigating the synthetic BC speech, the proposed method was applied to real BC speech. The AC and BC speech vowels (/a/, /i/, /u/, /e/, and /o/) were uttered by a female speaker and simultaneously recorded in a sound-isolated room. Sony ECM-J3M and Temco HG-17 microphones were used for recording the AC and BC speech, respectively. To observe the spectral dynamic range, we plotted the FFT spectrum of the real BC speech signal and the corresponding AC speech signal in Figure 9. It can be seen from Figure 9 that the high-frequency components of the BC speech signal are attenuated, increasing the dynamic range. As the true values of the spectral parameters are unavailable in the case of natural speech, it is difficult to evaluate the performance of spectrum estimation [13]. However, we plotted the FFT spectrum of BC speech as the reference and observed the power spectrum estimated by the autocorrelation method and the proposed method. The prediction orders $M$ and $L$ were 12 and 2, respectively, for the proposed method, and the order $M=14$ was used for the autocorrelation method. As an example, Figure 10 shows the spectrum of real BC speech for the vowel /a/. From Figure 10, it can be observed that the proposed method provides a closer spectrum than the autocorrelation and SC methods. In particular, above 1 kHz the power spectrum for the autocorrelation method is highly attenuated but the proposed method provides an emphasized spectrum that matches the envelope of the FFT spectrum of BC speech. As we saw for the synthetic BC speech, especially in Figures 4-8, the difference between the power spectrum estimate and the envelope of the FFT spectrum is a useful indicator of spectrum estimation accuracy. Figure 11 also shows the same phenomenon as that in Fig. 10 for the vowel /e/ (although omitted here, we observed the same tendency for the vowels /i/, /u/, and /o/). Figures 10 and 11 reveal that the proposed method provides better performance than the autocorrelation and SC methods.

A sentence uttered by a female speaker was also used to verify the effectiveness of the proposed method here. The experimental specifications are similar to those for the vowels. Figures 12-15 show the power spectra of phonemes detected from parts of the sentence (about 10 s length). In each figure, the FFT spectrum and power spectrum estimated by the autocorrelation, SC, and proposed methods are shown. It is observed from Figures 12-15 that the proposed method provides a closer
spectrum to the envelope of the FFT spectrum than the other methods. We also observed the same phenomenon for other phonemes included in the sentence. These results show that the proposed method can be effectively applied to BC voiced speech.

4. Conclusion

In this paper, we have proposed a pitch extension version of the SC method to accurately estimate the power spectrum of a BC voiced speech signal. Experiments on synthetic and real BC speech have demonstrated that the proposed method provides more accurate power spectrum estimation than the autocorrelation and original SC methods.

Acknowledgment

The authors would like to thank the reviewers for their useful comments, which have significantly improved the paper.

References

[1] Z. Liu, Z. Zhang, A. Acero, J. Droppo and X. Huang: Direct filtering for air- and bone-conductive microphones, Proc. IEEE Workshop Multimedia Signal Processing, pp.363-366, 2004.
[2] T. Shimamura and T. Tamiya: A reconstruction filter for bone-conducted speech, Proc. Midwest Symp. Circuits and Systems, pp.1847-1850, 2005.
[3] K. Kondo, T. Fujita and K. Nakagawa: On equalization of bone conducted speech for improved speech quality, Proc. IEEE Int. Symp. Signal Processing and Information Technology, pp.426-431, 2006.
[4] T. Dekens, W. Verhelst, F. Capman and F. Beaugendre: Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection, Proc. European Signal Processing Conf., pp.1978-1982, 2010.
[5] B. Huang, Y. Xiao, J. Sun, G. Wei and H. Wei: Speech enhancement based on FLANN using both bone- and air-conducted measurements, Proc. Signal and Information Processing Association Annual Summit and Conf. (APSIPA), 5 pages, 2014.
[6] M. Li, I. Cohen and S. Mousazadeh: Multisensory speech enhancement in noisy environments using bone- and air-conducted microphones, Proc. Signal and Information Processing (ChinaSIP), pp.1-5, 2014.
[7] P. N. Trung, M. Unoki and M. Akagi: A study on restoration of bone-conducted speech in noisy environments with LP-based model and Gaussian mixture model, J. Signal Process., Vol.16, No.5, pp.409-417, 2012.
[8] M. S. Rahman and T. Shimamura: Pitch determination from bone conducted speech, IEICE Trans. Vol.E 99-D, No.1, pp.283-287, 2016.
[9] M. A. Rahman, Y. Sugiyura and T. Shimamura: Spectrum compensation method for speech signals based on prediction error filtering, WSEAS Trans. Syst. and Control, Vol.12, pp.213-220, 2017.
[10] L. R. Rabiner and R. W. Schafer, Theory and Applications of Digital Speech Processing, Prentice Hall, 2011.
[11] J. Makhoul : Linear prediction: A tutorial review, Proc. IEEE, Vol.63, No.4, pp.561-580, 1975.
[12] P. Kabal: Ill-conditioning and bandwidth expansion in linear prediction of speech, Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP), pp.824-827, 2003.
[13] K. K. Palivawal and P. V. S. Rao: A modified autocorrelation method of linear prediction for pitch-synchronous analysis of voiced speech, Signal Process., Vol.3, No.2, pp.181-185, 1981.
[14] T. P. Barnwell: Windowless technique for LPC analysis, IEEE Trans. Acoust. Speech Signal Process., Vol.28, pp.421-427, 1980.
[15] A. Camacho and J. G. Harris: A sawtooth waveform inspired pitch estimator for speech and music, J. Acoust. Soc. Am., Vol.124, No.3, pp.1638-1652, 2008.
[16] J. W. Xu and J. C. Principe: A pitch detector based on a generalized correlation function, IEEE Trans. Audio Speech Lang., Vol.16, No.8, pp.1420-1432, 2008.
[17] Y. Yamada and Y. Hijikata: Development of the bone conduction microphone for voice recognition, Denso Tech. Rev., Vol.8, No.1, pp.60-65, 2003.
[18] Z. Zhang et al: Multi-sensory microphones for robust speech detection, enhancement and recognition, Proc. Int. Conf. Acoustics, Speech and Signal Processing, pp.781-784, 2004.
[19] S. Iijima and T. Shimamura: Bone-conducted speech for speaker verification, Proc. Int. Workshop on Nonlinear Circuits and Signal Processing, pp.172-175, 2008.
[20] K. Amino, T. Osanai, T. Kamada, H. Makinae and T. Arai: Effects of the phonological contents and transmission channels on forensic speaker recognition, Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism (A. Neustein, H. Patil eds.), Springer-Verlag, pp.275-308, 2011.
[21] J. S. Erkelens and P. M. T. Broersen, Bias propagation in the autocorrelation method of linear prediction, IEEE Trans. Speech Audio Process., Vol.5, No.2, pp.116-119, 1997.
[22] L. R. Rabiner et al.: A comparative performance study of several pitch detection algorithms, IEEE Trans. Speech Process., Vol.24, No.5, pp.399-418, 1976.
[23] J. D. Markel: The SIFT algorithm for fundamental frequency estimation, IEEE Trans. Audio Electroacoust., Vol.20, No.5, pp.366-377, 1972.
[24] F. Itakura and Y. Yohkura: Feature extraction of speech signal and its application to data compression, J. Inf. Process. Soci. Jpn, Vol.19, No.7, pp.644-656, 1978.
[25] J. D. Markel: Digital inverse filtering-a new tool for formant trajectory estimation, IEEE Trans. Audio Electroacoust., Vol. AU-20, No.2, pp.129-137, 1972.
[26] L. R. Rabiner: On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoust. Speech Signal Process., Vol.25, No.1, pp.24-33, 1977.
autocorrelation function on the log spectrum, Electroni. Communi. Jpn, Part 3, Vol.83, No.1, pp.90-98, 2000.

[28] S. Gonzalez and M. Brookes: PEFAC - A pitch estimation algorithm robust to high levels of noise, IEEE/ACM Trans. Audio Speech Lang. Process., Vol.22, No.2, pp.518-530, 2014.

Md Arifour Rahman received his B.E., and M.E., degrees in applied physics and electronic engineering from Rajshahi University, Rajshahi, Bangladesh, in 2003 and 2004, respectively. He is currently pursuing the Ph.D. degree at the Graduate School of Science and Engineering, Saitama University, Saitama, Japan. His current research interests include speech analysis, synthesis and enhancement of bone-conducted speech.

Tetsuya Shimamura received his B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1986, 1988, and 1991, respectively. In 1991, he joined Saitama University, Saitama, Japan, where he is currently a Professor. His research interests include digital signal processing and its applications to image processing, speech processing, and communication systems.

(Received April 17, 2018; revised August 1, 2018).