Pitch-Synchronous Linear Prediction Analysis of High-Pitched Speech Using Weighted Short-Time Energy Function

Liqing Liu\(^1\) and Tetsuya Shimamura\(^2\)

Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama 338-8570, Japan
E-mail: \{leisei\(^1\),shima\(^2\}\}@sie.ics.saitama-u.ac.jp

Abstract Conventional linear prediction (LP) analysis is known to suffer from problems in estimating the formant frequencies (vocal tract resonances) of high-pitched speech signals. The performance of conventional LP analysis deteriorates due to the harmonic structure of the glottal excitation source, especially in the case of high-pitched speech signals. Attempting to resolve this problem, a pitch-synchronous analysis technique based on a short-time energy function is presented. The proposed method has been verified to reduce the effect of the harmonic structure of the glottal excitation source. Experiments were carried out using synthetic vowels and real vowels. The results show that the proposed method yields a better performance in the estimation of formant frequencies than some previous LP analysis methods.

Keywords: glottal excitation, glottal closure, high-pitched, weighted LP, peak picking

1. Introduction

Linear prediction (LP) \cite{1} is well-suited for speech analysis due to its ability to model the speech production process approximately. Hence, LP analysis has been widely used for estimating formant frequencies (vocal tract resonances), pitches, glottal flows and so forth. In addition, LP can guarantee the stability of the estimated all-pole filter when utilizing the stationary (autocorrelation) formulation. However, the performance of the conventional LP method is influenced by several factors, such as the pitch harmonic structure of the glottal excitation source (fundamental frequency), the analysis segment size and so forth. The pitch period causes aliasing to take place in the autocorrelation domain, resulting in the estimation of formant frequency being degraded due to the neighbouring harmonics. The degradation becomes especially severe in the case of high-pitched speech.

In order to eliminate or decrease the influence of the pitch harmonic structure of the glottal excitation source, several modifications to conventional LP have been investigated. Miyoshi et al. \cite{2} presented a sample-selective linear prediction (SSLP) method for the analysis of high-pitched speech signals, which discards speech samples whose residual values exceed a threshold to decrease the effect of the pitch period.

Rahman and Shimamura \cite{3} proposed an improvement to the LP technique by employing homomorphic deconvolution in the autocorrelation domain to eliminate the aliasing effect, which is caused by the autocorrelation of the vocal tract impulse response being repeated periodically due to the pitch harmonic. In the case of high-pitched speech, the pitch is very short and the increased overlapping causes severe spectral distortion. This technique is known as LP using refined autocorrelation (LPRA). In the last few decades, weighted LP (WLP) has received considerable attention. The basic concept of WLP is to estimate the all-pole filter by applying temporal weighting of the square of the residual signal. The temporal weighted function aims to emphasize the speech samples during the glottal closed phase and attenuates the effect of the glottal excitation source. Several weighted functions have been designed. Yanagida et al. \cite{4} devised a weighted function based on an exponential function. Ma et al. \cite{5} chose the short-time energy (STE) as the weighted function. The WLP method based on STE weighting has been verified to enable spectral models to be less vulnerable to the influence of the pitch period than conventional LP. Alku and Pohjalainen \cite{6} proposed a weighted function called the attenuated main excitation (AME) function. The WLP-AME method requires the instants of the glottal closure to be identified to determine the locations of the main excitations by using either an electroglottogra-
phy (EGG) signal or epoch extraction techniques [7].

Another way to remove the influence of the pitch harmonic structure of the glottal excitation source is to extract only an interval included within the duration of the closed phase of a glottal cycle. This is known as pitch-synchronous analysis. For the pitch-synchronous analysis of voiced speech, the duration of the analysis segment is less than or equal to one pitch period (glottal cycle) [8]. In [9], a Fourier transform was applied to perform pitch-synchronous analysis using successive approximations to find the poles and zeros characterizing glottal excitation to approximate the spectra. However, the resulting spectra with a harmonic structure adversely affected the extraction accuracy of formants, especially for high-pitched speech. A well-known speech system named TANDEM-STRAIGHT [10], which is a speech analysis, modification and synthesis framework, provides a stable power spectrum and can eliminate the temporal and spectral variance caused by time window positioning and the harmonic structure, respectively. The process for extracting the \( F_0 \) (pitch) adaptive spectral envelope, which is based on consistent sampling theory and consists of the procedures of smoothing, anti-aliasing and compensating for spectral envelope recovery, is complicated. It is well known that the exclusion of areas known to correspond to an open glottis will lead to more accurate estimation of the vocal tract. The key point is how to find such intervals. Wong et al. utilized the minimum of the normalized total squared error to locate the instants of glottal closure and opening [11].

In this paper, our purpose is to find the duration of glottal closure for accurate estimation of the vocal tract by exploiting the STE function. Rather than estimating the instants of glottal closure and opening exactly, we utilize the simple STE computation of the speech signal and the predicted residual signal to extract the interval of the glottal closed phase during a glottal cycle. Since nonstationarity is a better assumption than stationarity [12] for pitch-synchronous analysis, we apply a nonstationarity formulation of LP to the extracted interval.

2. Short-Time Energy (STE) Function

In [5], the temporal weighted function was calculated from the speech signal using the STE function

\[
w_n = \sum_{i=1}^{M} s_{n-i}^2
\]  

where \( s_n \) is a speech signal and \( M \) is the length of the STE window. The use of STE weighting based on a speech signal has been successfully verified for feature extraction in automatic speech recognition [13], glottal flow estimation [14] and speaker verification [15].

\[w'_n = \sum_{i=1}^{D} e_{n+i-1}^2\]

From Eq. (1), the STE function emphasizes the duration where the speech samples have a large amplitude. A speech signal with a larger amplitude appears in a glottal closed phase interval. Hence, the STE weighted function can be used to focus on the glottal closed phase and emphasize the contribution of the speech samples in the glottal closed phase interval [13].

However, the purpose of the method is not to try to define the glottal closed phase interval precisely. Hence, sometimes the STE weighted function calculated directly from the speech signal cannot completely attenuate the influence of the glottal excitation source as shown in Fig. 1. Figure 1(a) illustrates that the STE weighted function calculated from a speech signal using Eq. (1), where \( M \) is 10, emphasizes the duration where the speech waveform has a larger amplitude. However, a drawback appears in Fig. 1(b). The STE weighted function cannot completely attenuate the contribution of the residual peaks. The remainders of the residual peaks produce a biased spectrum and affect the estimation of the formant frequency, especially in the case of high-pitched speech. In order to resolve this problem, we devise an STE function based on the predicted residual signal in the next section.

3. Proposed Method

3.1 STE function based on residual signal

An STE function based on a residual signal is computed as

\[w'_n = \sum_{i=1}^{D} e_{n+i-1}^2\]
where $e_n$ is a residual signal (prediction error) and $D$ is the length of the STE window. The residual signal $e_n$ is computed as

$$e_n = s_n - \hat{s}_n = s_n - \sum_{i=1}^{p} a_i s_{n-i}$$  \hspace{1cm} (3)

where $\hat{s}_n$ is the estimated value, $p$ is the linear predictor order and $a_i$ are the predictive coefficients. In Eq. (3), $a_i$ are obtained by the covariance method as follows:

$$\sum_{i=1}^{p} a_i \phi_{ji} = \phi_{j0}$$  \hspace{1cm} (4)

where

$$\phi_{ji} = \sum_{n=p}^{N-1} s_{n-j} s_{n-i}$$  \hspace{1cm} (5)

Here, $N$ denotes time sequence samples of each speech frame.

Comparing Eq. (2) with Eq. (1), there is a difference. In Eq. (2), the value of the current weighted function is computed from the future residual signal. Our proposal is to avoid a computation delay. For a high-pitched speech signal, the duration of the glottal closed phase interval is very short. A computation delay will thus cause a severe extraction error. The use of the STE function based on the residual signal is justified by two aspects. Firstly, unlike the speech signal, the residual signal has a direct relationship with the glottal excitation source. The residual signal, $e_n$, is an approximation of the second derivative of the glottal waveform [16]. Secondly, the prediction error will be large in the main excitation interval. In particular, at the instant of glottal closure, the amplitude of the speech signal has the largest increase [17]. In the glottal closed phase interval, the values of the residual signal are assumed to be small [18]. Hence, calculating the STE function from the residual signal can emphasize the main excitation interval. As shown in Fig. 2, the STE weighted function calculated from the residual signal using Eq. (2), where $D = 8$, includes the residual peaks even for the main glottal excitation duration. It can be considered that the STE weighted function calculated from the residual signal can be used to locate the interval of the main excitation.

3.2 Pitch-synchronous analysis based on weighted STE function

Here we normalize the STE function $w'_n$ by $w'_{nn} = \frac{w'_n}{\max_{i} w'_i}$ for each frame, where $\max_{i} (w'_i)$ denotes the maximum value in each current frame. Then we subtract it from 1 to give

$$W'_{n} = 1 - w'_{nn}$$  \hspace{1cm} (6)

For the following computation, a small value of $d$ (e.g., 0.01) is introduced here and Eq. (6) is rewritten as

$$W'_n = \begin{cases} W'_n & W'_n \geq d \\ d & W'_n < d \end{cases}$$  \hspace{1cm} (7)

so that $W'_n$ can be a positive real nonzero value.

We combine Eqs. (1) and (7) to derive a new weighted function for locating the glottal closed phase interval. The new weighted function is expressed as

$$W_n = w_n \times W'_n$$  \hspace{1cm} (8)

The new weighted function has two advantages. Firstly, $W_n$ inherits the merit of $w_n$ of emphasizing the speech signal occurring during the glottal closed phase. Secondly, by multiplying by the $W'_n$ function, $W_n$ can also avoid the influences of the main glottal excitation source.

We considered how to extract the glottal closed phase interval by using the proposed weighted function $W_n$. Actually, instead of extracting the glottal closed phase interval directly from the speech signal domain, we search for the corresponding duration in the proposed weighted function domain. A threshold is introduced here to assist in extracting the most suitable duration. The threshold $\theta$ in this paper is experientially designed as

$$\theta = \begin{cases} \text{mean}(W_n) & F_0 \geq 200 \text{ Hz} \\ \frac{1}{2} \times \text{mean}(W_n) & F_0 < 200 \text{ Hz} \end{cases}$$  \hspace{1cm} (9)

where mean$(\cdot)$ denotes an average value. The value of 200 Hz is set as the boundary between male and female speech. For female speech with a high $F_0 \geq 200$ Hz, it is known that the wide spacing of harmonics leads to degradation of the formant estimation. However, the influence of the harmonic structure can basically be ignored for male speech. It has been shown that the relative estimation error of the formant is not affected significantly when $F_0$ is less than 200 Hz [3]. Since the formant estimation of male speech, whose $F_0$ is less than 200 Hz, can ignore the influence of
the harmonic structure, we reduced the value of $\theta$ so that more speech samples could be extracted. For the pitch-synchronous analysis of LP, a long interval can provide greater temporal stability of formant estimation.

Then, in the $W_n$ domain we locate the intervals whose values exceed the threshold $\theta$. Next, we compute the length of each located interval and choose the interval with the largest length, $L$, as the most suitable duration. We extract the speech signal duration corresponding to the location of the most suitable duration in the $W_n$ domain as the glottal closed phase interval. The extracted glottal closed phase interval is denoted as $\{s_q, s_{q+1}, \cdots, s_{q+L-1}\}$, where $q$ is the position number of the first sample in the extracted glottal closed phase interval, which corresponds to the location of the original speech frame in the $s_n$ domain. Figure 3 illustrates this procedure. In Fig. 3, the proposed weighted function $W_n$ in the upper panel corresponds to the speech signal $s_n$ for a frame in the lower panel. From $W_n$, we locate the intervals whose values exceed the threshold $\theta$, corresponding to the thick horizontal line, and select the interval with the largest length $L$ as the most suitable duration. The speech signal $\{s_q, s_{q+1}, \cdots, s_{q+L-1}\}$ corresponding to the location of the most suitable duration is obtained. Then, based on the extracted speech interval, the LP parameters can be computed.

The computation steps in the proposed method are summarized below:

**Step 1** Perform the STE computation of the speech signal and residual signal. The residual signal is estimated using the covariance method for frames.

**Step 2** After normalizing and implementing Eqs. (6), (7) and (8), a new weighted function $W_n$ is obtained.

**Step 3** Extract intervals for which the amplitudes of the proposed weighted function $W_n$ exceed the threshold $\theta$ and calculate the length of each extracted interval.

**Step 4** Select the interval with the largest length as the glottal closed phase interval, that is, $\{s_q, s_{q+1}, \cdots, s_{q+L-1}\}$. Actually, the length, $L$, may be smaller than the prediction order $p$ in the case of high-pitched speech. Hence, it is necessary to compare their sizes. If $L < p$, we extend the length of the glottal closed phase interval by adding speech samples in the reverse direction of time until $L \geq p$ is satisfied.

**Step 5** Compute the LP parameter by the following formulation [2]:

$$Y^T \hat{Y} \hat{a} = Y^T \delta$$  \hspace{1cm} (10)

where

$$Y = \begin{bmatrix}
  s_{q-1} & s_{q-2} & \cdots & s_{q-p} \\
  s_{q} & s_{q-1} & \cdots & s_{q-p+1} \\
  s_{q+1} & s_{q} & \cdots & s_{q-p+2} \\
  \vdots & \vdots & \ddots & \vdots \\
  s_{q+L-2} & s_{q+L-3} & \cdots & s_{q+L-p-1}
\end{bmatrix}$$  \hspace{1cm} (11)
Fig. 4 Block diagram of the proposed method

\[
\hat{a} = [a_1, a_2, a_3, ..., a_p]^T \quad (12)
\]

\[
\delta = [s_q, s_{q+1}, s_{q+2}, ..., s_{q+L-1}]^T \quad (13)
\]

and \(T\) denotes transposition.

The LP parameter, \(\hat{a}\), is obtained by

\[
\hat{a} = [Y^T Y]^{-1} Y^T \delta \quad (14)
\]

A block diagram of the proposed method is depicted in Fig. 4.

4. Experimental Results

4.1 Results for synthetic speech

To verify the effectiveness of the proposed method, several experiments have been conducted for synthetic vowels and real vowels.

4.1.1 Impulse train excitation

A synthetic vowel /o/ [8] was generated from an impulsive sequence excitation with a known value of the pitch period using the gain and autoregressive parameters \(G=0.1354, a_1=-1.53527, a_2=0.97789, a_3=-1.48396, a_4=1.78023, a_5=-0.71704, a_6=-0.73514, a_7=-0.76348, a_8=-0.12135, a_9=0.15552, a_{10}=0.17814\). The sampling frequency was 10 kHz. Depending on the pitch, various synthetic vowels /o/ can be generated. We utilized the generated vowels to evaluate the proposed method. Figure 5 shows the average LP power spectra of 95 consecutive frames estimated by three methods: the covariance method, WLP [5] and the proposed method. The four vertical lines represent the true formant values. The prediction order \(p\) of the all-pole model is set to 10 so that it is equal to the system order of the generated model. The frame length is set to 25.6 ms to include 256 samples and the frame shift is half of the frame length. The lengths of the STE window, \(M\) and \(D\), in Eqs. (1) and (2) are set to \(p\) and 8, respectively.

Figure 5 illustrates some noteworthy features. Firstly, the shapes of the LP power spectra estimated by the proposed method are stable and invariant regardless of the fundamental frequency \(F_0\). This means that the performance of the proposed method is basically not affected by the value of \(F_0\). Secondly, with increasing \(F_0\), the shapes of the LP power spectra estimated by the covariance method vary. It is observable that the performance of the covariance method is easily influenced by the value of \(F_0\). Although the first formant peak estimated by the covariance method occurs at nearly the true value at a high pitch with \(F_0 = 400\) Hz, it cannot be claimed that the formant estimation is not influenced by the value of \(F_0\). The reason why the first formant estimated by the covariance method is close to the true value is that the high-pitched \(F_0\) at 400 Hz happens to be close to the true \(F_1\) at 410 Hz. The LP power spectra estimated by the WLP can be seen to be less affected by the value of \(F_0\).

The average estimated prediction parameters for 95 consecutive frames at \(F_0 = 400\) Hz are summarized in Table 1. It can be seen from Table 1 that the estimation accuracy is highest for the proposed method.

4.1.2 Realistic excitation

Rather than using impulse trains, we utilized more realistic excitation waveform that can be used as an approximate human glottal source model to generate synthetic speech. Here, the Liljencrants-Fant (LF) model [19] was introduced to generate synthetic vowels. The LF model has been proved to be suitable for describing glottal area functions and capable
Table 1 Estimated parameters for synthetic vowel /o/ at $F_0=400$ Hz

| Parameter | True Value | Covariance Method | WLP     | Proposed Method |
|-----------|------------|-------------------|---------|-----------------|
| $a_1$     | -1.53527   | -1.52722          | -1.37631| -1.52569        |
| $a_2$     | 0.97789    | 0.88255           | 0.46616 | 0.95886         |
| $a_3$     | -1.48396   | -1.42887          | -0.61737| -1.46692        |
| $a_4$     | 1.78023    | 1.86355           | 0.89122 | 1.76031         |
| $a_5$     | -0.71704   | -0.78345          | -0.03677| -0.69407        |
| $a_6$     | 0.73514    | 0.87863           | 0.07798 | 0.71971         |
| $a_7$     | -0.76348   | -1.05215          | -0.22942| -0.75231        |
| $a_8$     | -0.12135   | 0.13832           | -0.29240| -0.13138        |
| $a_9$     | 0.15552    | -0.03390          | 0.18248 | 0.15710         |
| $a_{10}$  | 0.17814    | 0.27303           | 0.15253 | 0.18046         |

Fig. 5 LP power spectra estimated by the covariance method (blue lines), WLP (red lines) and proposed method (black lines) from the synthetic vowel /o/ whose fundamental frequencies $F_0$ range between 100 Hz and 400 Hz in seven steps.

of producing natural-sounding synthetic speech [20], [21]. We utilized the LP model as the glottal excitation source to generate five synthetic vowels whose sampling frequency was 10 kHz. The formant frequencies specified for the five synthetic vowels are listed in Table 2 [3], [16]. Then the generated speech is preemphasized by a $1 - z^{-1}$ filter to simulate the lip radiation characteristics.

We experimentally compare the estimation accuracy of the formant frequency of the proposed method with that of the covariance method, WLP [5] and LPRA [3]. The speech is preemphasized by a $1 - z^{-1}$ filter before analysis. The formant frequencies are estimated using the peak-picking technique to extract the peaks of the all-pole spectrum except for LPRA. The peak-picking technique has been shown to be reliable, especially when handling formants that are located at low frequencies or close to each other [22]. For consistency with the original LPRA method in [3], the formant frequencies for LPRA in this paper are also estimated using the root-solving method from the estimated AR parameters. The lowest three formant values in 95 consecutive frames are obtained and used for evaluation. The other experimental specifications are summarized as follows:

- frame length: 25.6 ms
- frame shift: 12.8 ms
- prediction order: $p = 12$
- length of the STE window: $M = p$ (used in WLP and proposed method)
- length of the STE window: $D = 8$ (used in proposed method)
- window function: Hamming (used in LPRA)
- number of FFT points: 1024 (used in LPRA)

The relative estimation error (EE) [3] is introduced to evaluate the performance of the proposed method. EE for five vowels is expressed by

$$ EE_i = \frac{1}{5} \sum_{j=1}^{5} \sum_{k=1}^{K} \left| \frac{F_{ij,k} - F_{ij}}{F_{ij}} \right| \times 100\% $$

(15)
where $\hat{F}_{ij,k}$ is the estimated $i$th formant frequency of the $j$th vowel at the $k$th frame, $F_{ij}$ denotes the true value of the $i$th formant frequency of the $j$th vowel and $K$ is the frame number.

Figures 6 - 8 respectively show the average relative EE of the first, second and third formant frequencies for the five synthetic vowels. Comparing these results, it is worth noting that although $F_0$ varies, the proposed method provides stable and small EE values for each formant. In other words, the influence of the pitch is eliminated in the proposed method. As can be seen in these figures, the frequency estimation error for the first formant is more severe than those for the second and third formants for most values of $F_0$. Namely, frequency estimation of the first formant is much more easily influenced by $F_0$ than that of the second and third formants.

Furthermore, we averaged the estimation errors of the first three formants of the five vowels as follows:

$$EE = \frac{1}{3} \sum_{i=1}^{3} EE_i$$  \hspace{1cm} (16)

Figure 9 shows the average error for the first three formants of the five vowels obtained from Eq. (16). The results suggest that the proposed method produces the smallest formant estimation error in general. This indicates that the proposed method is capable of eliminating the influence of glottal excitation. In general, the estimation accuracy of WLP and LPRA is better than that of the covariance method, which implies that the WLP and LPRA methods are less vulnerable to changes in $F_0$ than the covariance method.

The proposed method is also applied to analyze speech signals including both poles and dips. Two high-pitched speech signals are synthesized by exciting pole-zero resonators with a glottal waveform generated using the LF model. The frequencies and bandwidths of the poles are set to 800, 1200 and 3500 Hz and 50, 100 and 120 Hz, respectively. The frequency and bandwidth of the zero are set to 2200 and 100 Hz, respectively [23]. Two sounds are synthesized with $F_0 = 300$ Hz and $F_0 = 350$ Hz. The synthesized sounds are pre-emphasized by a $1 - z^{-1}$ filter. The experimental specifications are similar to those of the five synthetic vowels.

Figures 10 and 11 show the average spectra of ten consecutive frames estimated by four methods. The vertical dotted line and thick line represent the locations of $2F_0$ and the first formant, respectively. Since one assumption of LP analysis is that the vocal tract model in LP analysis is an approximation of an all-pole model, all four methods based on the LP formulation
failed to extract the zero location. We compared the formant (pole) estimation performance of these four methods. In Fig. 10, the first formant estimated by WLP, LPRA and the proposed method is close to the true value. However, the first formant estimated by the covariance method clearly deviates from the true location and is close to the second harmonic structure expressed by the vertical dotted line. It can be seen that the performance of the covariance method is more seriously affected by the pitch structure than other methods.

Some similar phenomena can be seen in Fig. 11. Firstly, both the covariance method and the WLP fail to accurately exhibit the first and second formant peaks. In particular, the location of the first formant peak estimated by these methods is near to the second harmonic structure denoted by the vertical dotted line. Secondly, although the LPRA estimates the second formant peak accurately, the first formant peak deviates from the true location represented by the vertical black line. On the other hand, the proposed method exhibits basically accurate formant peaks that are near the true ones due to its ability to exclude the influence of the pitch harmonic structure. These results indicate that even if the proposed method is applied to high-pitched speech consisting of poles and zeros, the proposed method is effective for extracting the formants (poles).

4.2 Results for real speech

Real vowels are also used to verify the effectiveness of the proposed method. Two real vowels uttered by female speakers are used to evaluate the performance of the method. The two speech data are as follows:

- /u/ in /bu/ at $F_0 \approx 300$ Hz
- /o/ in /bo/ at $F_0 \approx 340$ Hz

The experimental specifications are similar to those for synthetic vowels in Sect. 4.1.2. Figures 12 and 13 show the spectra of /u/ and /o/ estimated by the covariance, WLP, LPRA and proposed methods for 10 consecutive frames, respectively. In Fig. 12, it can be seen that all the methods extract formants well at the high-pitched vowel. However, the disparity in the performance among these methods is clearly reflected in Fig. 13. From Fig. 13, it is observed that the LPRA and proposed methods succeeded in tracking the first and second formants, while the covariance and WLP methods failed to separate the first and second formants when their frequencies were close.

Since the exact formant frequencies of real vowels are not known, we cannot provide an EE to evaluate the estimation performance of these methods. Here we summarize the estimation performance in terms of the mean and standard deviation, as introduced in [24]. Table 3 shows the estimation performance of /u/ in /bu/ at $F_0 \approx 300$ Hz for 48 consecutive frames in terms of the mean and standard deviation. The mean followed by the standard deviation are shown in the parenthesis.

Unlike the case of /u/, in which all the methods can successfully extract the five formants, in the case of /o/ the methods cannot estimate all five formants. Here we introduce well-estimated numbers 1 to evaluate the performance. Then the mean and standard deviation were computed from the well-estimated formants. The estimation performance of /o/ in /bo/ at $F_0 \approx 340$ Hz in terms of the well-estimated numbers, mean and standard deviation for 52 consecutive frames is shown in Table 4. The mean and standard deviation are shown in the parenthesis. From Table 4, it can be seen that the mean values of the first formants are near the true ones due to its ability to exclude the influence of the pitch harmonic structure. These results indicate that even if the proposed method is applied to high-pitched speech consisting of poles and zeros, the proposed method is effective for extracting the formants (poles).

---

1 Well-estimated numbers are the numbers of well-estimated formants. From the distribution of formants for female speech, the well-estimated formants are defined as follows: the peaks extracted in the region [0, 1500] Hz are defined as the first and second formants in a sequence; the peaks extracted in the region [2000, 4000] Hz are defined as the third and fourth formants in a sequence; the peak extracted in the region [4000, 5000] Hz is defined as the fifth formant.
The experimental specifications are similar to those of the method. The two speech data are as follows: female speakers are used to evaluate the performance of the proposed method. Two real vowels uttered by 4.2 Results for real speech of the formants (poles).

zeros, the proposed method is effective for extracting the influence of the pitch harmonic structure. These results indicate that even if the proposed method is near the true ones due to its ability to exclude the formant peaks. In particular, the location of the first formant deviates from the true location represented by the vertical dotted line. On the other hand, the proposed method exhibits basically accurate formant peaks that fail to accurately exhibit the first and second formant location and is close to the second harmonic structure denoted by the vertical dotted line. It can be seen that the performance of the covariance method clearly deviates from the true value. However, the first formant estimated by the covariance method is more seriously affected by the pitch structure than the other methods. In Fig. 10, the first formant estimated by the covariance method clearly deviates from the true location and is close to the second harmonic structure denoted by the vertical dotted line. It can be seen that all the methods extract formants well at the high-pitched vowel. However, the disparity in the performance among these methods is clearly reflected in the formant (pole) estimation performance of these four methods. Here we introduce well-estimated numbers of /o/ the methods cannot estimate all five formants. Some similar phenomena can be seen in Fig. 11. Since the exact formant frequencies of real vowels are not known, we cannot provide an EE to evaluate the performance. Then the mean and standard deviation are shown in the parenthesis. From Table 4, the well-estimated formants are defined as follows: the peaks extracted in the region $[0, 1500]$ Hz are defined as the first and second formants in a sequence; the peaks extracted in the region $[2000, 4000]$ Hz are defined as the third and forth formants in a sequence; the peak extracted in the region $[4000, 5000]$ Hz is defined as the fifth formant.

Fig. 12 Spectra of /u/ in /bu/ at $F_0 \approx 300$ Hz estimated by covariance (a), WLP (b), LPRA (c) and proposed method (d).

Fig. 13 Spectra of /o/ in /bo/ at $F_0 \approx 340$ Hz estimated by covariance (a), WLP (b), LPRA (c) and proposed method (d).
The LP analysis tends to focus on the stability of the resulting all-pole filter. The proposed method, which is based on a nonstationary formulation, cannot guarantee the stability of the resulting all-pole filter. However, the stability of the resulting all-pole filter is an important issue, especially when applied to applications such as speech synthesis. The transfer function of the all-pole filter can be modeled as

$$H(z) = \frac{G}{1 + \sum_{i=1}^{p} a_i z^{-i}}$$

where $a_i$ represents the $i$th pole. For a stable all-pole filter, all poles must be strictly inside a unit circle, which means that $|a_i| < 1$. As a nonstationary formulation for LP analysis, the resulting all-pole filter estimated from the proposed method may become unstable. When $|a_i| > 1$, $a_i$ is replaced with $p_i / |p_i|^2$ so that the all-pole filter becomes stable [25].

5. Discussion

There are two important parameters, $D$ and $\theta$, in this paper. The parameter $D$ is considered as the length of the main glottal excitation areas. The length of the main glottal excitation areas varies with $F_0$ and is a certain proportion of a glottal cycle. For real speech, it is difficult to extract the interval exactly. Even for an equivalent glottal cycle length, some factors such as the stress accent will also affect the length of the main glottal excitation areas. The optimal proportion of the main excitation areas for the glottal cycle used in [6], 32% of $T$, is introduced here to compute the length of $D$, where $T$ denotes the length of a glottal cycle. For a female speech signal with the upper limit $F_0 = 400$ Hz sampled at a rate of 10 kHz, the length of the main glottal excitation areas is calculated as 32% of $(10000/400) = 8$. General speaking, $D$ increases when $F_0$ decreases. However, the setting of $D = 8$ was used in all the above experiments. This is because the final proposed weighted function $W_n$ is the product of $w_n$ and $W_n^\prime$. $w_n$ can compensate the influence of $W_n^\prime$ caused by the insufficient length of $D$ so that the performance of the desired duration extracted by $W_n$ is not affected.

The other important parameter $\theta$ is the threshold, which is used to determine the interval of glottal closed phase areas. Under the most common condition, the length of the glottal closed phase interval is assumed to be equal to that of the glottal open phase interval for a glottal cycle. Since the amplitude of speech samples in a glottal closed phase interval is larger than that in a glottal open phase interval, the average value can be utilized as a threshold to extract the glottal

| $F_0$ | Cov. | WLP | LPRA | Prop. |
|-------|------|-----|------|-------|
| $F_1$ | (428,17) | (440,23) | (433,38) | (429,23) |
| $F_2$ | (1563,10) | (1576,10) | (1556,38) | (1609,16) |
| $F_3$ | (2712,30) | (2640,15) | (2618,129) | (2619,17) |
| $F_4$ | (3720,31) | (3685,32) | (3676,63) | (3680,53) |
| $F_5$ | (4223,23) | (4235,19) | (4232,33) | (4213,25) |

Table 3 Estimated means and standard deviations for /u/ in /bu/ at $F_0 \cong 300$ Hz

| $F_0$ | Cov. | WLP | LPRA | Prop. |
|-------|------|-----|------|-------|
| $F_1$ | 52 | (685,9) | 52 | (689,11) |
| $F_2$ | 4 | (979,8) | 7 | (953,14) |
| $F_3$ | 52 | (2910,46) | 52 | (2908,35) |
| $F_4$ | 13 | (3758,39) | 9 | (3804,49) |
| $F_5$ | 52 | (4382,52) | 52 | (4390,36) |

Table 4 Estimated well-estimated numbers, means and standard deviations for /o/ in /bo/ at $F_0 \cong 340$ Hz

The LPRA basically failed to extract the second formant, while the LPRA and the proposed method could successfully estimate the second formant in most cases. Note that the interval applied to compute the LP parameters by the proposed method is short, especially in the case of high-pitched speech.
closed interval easily. This is why $\theta$ is experientially fixed to the average value of $W_n$ for female speech ($F_0 \geq 200$ Hz) in this paper. The experimental results show that setting the parameter $\theta$ to the average value of $W_n$ for female speech is effective. The setting of $\theta$ is certainly under a trade-off condition between formant estimation accuracy and temporal stability. A more effective and automatic way of setting the threshold $\theta$ is under investigation.

6. Conclusions

In this paper, a pitch-synchronous analysis technique for linear prediction was proposed by employing the STE function based on speech and residual signals. The proposed method locates a duration of glottal closure that excludes the speech samples when the glottis is open and leads to the more accurate frequency estimation of formants. Based on the experimental results, the proposed method is shown to be suitable for analyzing high-pitched speech signals and robust against changes in the glottal excitation source.

References

[1] B. S. Atal and S. Hanauer: Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Am., Vol. 50, No. 2, pp. 637-655, 1971.
[2] Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida and O. Kakusho: Analysis of speech signals of short pitch period by a sample-selective linear prediction, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-35, No. 9, pp. 1233-1240, 1987.
[3] M. Shahidur Rahman and T. Shimamura: Linear prediction using refined autocorrelation function, EURASIP J. Audio Speech Music Process. 45962, pp. 1-9, 2007.
[4] M. Yanagida, S. Tsukada and O. Kakusho: High-order over-determined weighted linear prediction analysis method, IEICE Technical Report, Vol. 87, No. 31, pp. 49-56, 1987.
[5] C. Ma, Y. Kamp and L. Willems: Robust signal selection for linear prediction analysis of voiced speech, Speech Commun., Vol. 12, No.1, pp. 69-81, 1993.
[6] P. Alku and J. Pohjalainen: Formant frequency estimation of high-pitched vowels using weighted linear prediction, J. Acoust. Soc. Am., Vol. 134, No. 2, pp. 1295-1313, 2013.
[7] P. Naylor, A. Kounoudes, J. Gudnason and M. Brookes: Estimation of glottal closure instants in voiced speech using the DYPSA algorithm, IEEE Trans. Speech Audio Process., Vol. 15, No. 1, pp. 34-43, 2007.
[8] K. K. Paliwal 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.
[9] M. V. Mathews, J. E. Miller and E. E. David: Pitch synchronous analysis of voiced sounds, J. Acoust. Soc. Am., Vol. 33, No. 2, pp. 179-186, 1961.
[10] H. Kawahara and M. Morise: Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework, SADHANA, Acad. Proc. Engi. Sci., Vol. 36, Part 5, pp. 713-727, 2011.
[11] D. Y. Wong, J. D. Markel and A. H. Gray, Jr: Least square glottal inverse filtering from the acoustic speech waveform, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-27, No. 4, pp. 350-355, 1979.
[12] S. Chandra and W. C. Lin: Experimental comparison between stationary and nonstationary formulations of linear prediction applied to voiced speech analysis, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-22, No. 6, pp. 403-415, 1974.
[13] C. Magi, J. Pohjalainen, T. Backstrom and P. Alku: Stabilised weighted linear prediction, Speech Commun., Vol. 51, No. 5, pp. 401-411, 2009.
[14] G. F. Kalliantas, Y. Stylianou and P. Alku: Glottal inverse filtering using stabilised weighted linear prediction, Proc. Acoustic. Speech and Signal Processing, ICASSP pp. 5408-5411, 2011.
[15] R. Saeidi, J. Pohjalainen, T. Kinnunen and P. Alku: Temporally weighted linear prediction features for tackling additive noise in speaker verification, IEEE Signal Process. Lett., Vol. 17, No. 6, pp. 599-602, 2010.
[16] M. Shahidur Rahman and T. Shimamura: Speech analysis based on modeling the effective voice source, IEICE Trans., Vol. E89-D, No. 3, pp. 1107-1115, 2006.
[17] H. Strube: Determination of the instant of glottal closure from the speech wave, J. Acoust. Soc. Am., Vol. 56, No. 5, pp. 1625-1629, 1974.
[18] T. V. Ananthapadmanabha and B. Yegnanarayana: Epoch extraction from linear prediction residual for identification of closed glottis interval, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-27, No. 4, pp. 309-319, 1979.
[19] G. Fant, J. Liljencrants and Q. G. Lin: A four parameter model of glottal flow, Quart. Progress and Status Rep., Speech Transmission Lab, Royal Inst. Technol., Vol. 26, No. 4, pp. 1-13, 1985.
[20] H. Fujisaki and M. Ljungqvist: Proposal and evaluation of model for the glottal source waveform, Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, Vol. 4, pp. 1605-1608, 1986.
[21] H. Strik: Automatic parametrization of differentiated glottal flow: comparing methods by means of synthetic flow pulses, J. Acoust. Soc. Am., Vol. 103, No. 5, pp. 2659-2669, 1998.
[22] G. Vallabha and B. Tuller: Systematic errors in the formant analysis of steady-state vowels, Speech Commun., Vol. 38, No.1, pp. 141-160, 2002.
[23] M. Shahidur Rahman and T. Shimamura: Identification of ARMA speech model using an effective representation of voice source, IEICE Trans., Vol. E90-D, No. 5, pp. 863-867, 2006.
[24] S. A. Fattah, W. P. Zhu and M. O. Ahmad: An approach to formant frequency estimation at low signal-to-noise ratio, Proc. Acoustic. Speech and Signal Processing, ICASSP, Vol. 4, pp. 469-472, 2007.
[25] Y. Arima and T. Shimamura: Noise robust speech analysis using system identification methods, IEICE Trans., Vol. J83-A, No. 12, pp. 1455-1466, 2000.
Liqing Liu received his B.E. degree in electronic information engineering from Central South University, Changsha, China, in 2006 and his M.E. degree in mathematics, electronics and informatics from Saitama University, Saitama, Japan, in 2012. He is currently pursuing his Ph.D. degree in the field of speech analysis.

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, as a faculty member, where he is currently a Professor. He was a visiting scholar at Loughborough University (U.K.) in 1995 and at Queen’s University Belfast (U.K.) in 1996. Dr. Shimamura is an author or co-author of six books. He serves on the editorial boards of several international journals and is a member of organizing and program committees of various international conferences. His research interests are in digital signal processing and its applications to speech, image, and communication systems.

(Received June 9, 2014; revised December 10, 2014)