Acoust. Sci. & Tech. **26**, 6 (2005)

PAPER

Formant frequency estimation of high-pitched speech by homomorphic prediction

M. Shahidur Rahman* and Tetsuya Shimamura†

Department of Information and Computer Sciences, Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama, 338–8570 Japan

(*Received 20 December 2004, Accepted for publication 19 April 2005*)

Abstract: The conventional model of the linear prediction analysis suffers from difficulties in estimating vocal tract characteristics of high-pitched speakers. This is because the autocorrelation function used by the autocorrelation method of linear prediction for estimating autoregressive coefficients is actually an "aliased" version of that of the vocal tract impulse response. This "aliasing" occurs due to the periodic nature of voiced speech. Generally it is accepted that homomorphic filtering can be used to obtain an estimate of vocal tract impulse response which is free from periodicity. Thus linear prediction of the resulting vocal tract impulse response (referred to as homomorphic prediction) is expected to be free from variations of fundamental frequencies. To our knowledge any experimental study, however, has not yet appeared on the suitability of this method for analyzing high-pitched speech. This paper presents a detail study on the prospects of homomorphic prediction as a formant tracking tool especially for high-pitched speech where linear prediction fails to obtain accurate estimation. The formant frequencies estimated using the proposed method are found to be accurate by more than an order of magnitude compared to the conventional procedure. The accuracy of formant estimation is verified on synthetic vowels for a wide range of pitch periods covering typical male and high-pitched female speakers. The validity of the proposed method is also examined by inspecting the spectral envelopes of natural speech spoken by high-pitched female speakers. We noticed that almost all the previous methods dealing with this limitation of linear prediction are based on the covariance technique where the obtained AR filter can be unstable. The solutions obtained by the current method are guaranteed to be stable which makes it superior for many speech analysis applications.

Keywords: Linear prediction, Autocorrelation "aliasing," Minimum-phase cepstrum, Liftering, Fundamental frequency effect

PACS number: 43.72.Ar, 43.72.Gy, 43.72.Ja [DOI: 10.1250/ast.26.502]

1. INTRODUCTION

Formant frequencies are the principal analytical features in speech processing. This is because they are clearly related to the articulatory act and the perception of speech. Formant information is used extensively in coding, analysis/synthesis applications, and recognition of speech [1,2]. Linear predictive analysis [1] is one of the most powerful techniques to extract formant frequencies. The importance of this method lies in its ability to provide accurate estimates and its relative speed of computation. However, the conventional linear prediction technique is not free from limitations [2]. The basic formulation of the linear prediction seeks to find an optimal fit to the envelope of the speech spectrum. Since the source of voiced speech is of a quasi-periodic nature with spiky excitations, those impulsive periodic innovations sometimes result in inaccuracy in spectrum estimation, especially, in case of high-pitched speech. In this paper, we briefly illustrate the cause of inaccuracy of formant frequency estimation in case of pitch-asynchronous autocorrelation method and propose a solution based on homomorphic deconvolution.

In the conventional autocorrelation method of linear prediction (CALP) when a finite segment is extracted over multiple pitch periods, the obtained autocorrelation sequence is actually an "aliased"‡ version of the true autocorrelation of vocal tract system impulse response. This is because the replica of autocorrelation of vocal tract

*rahmanms@sie.ics.saitama-u.ac.jp
†shima@sie.ics.saitama-u.ac.jp

‡Unlike the fold-over phenomena in the frequency domain, the term "aliased" intends to mean the distortion due to the periodic repetition of autocorrelation function.

impulse response is repeated periodically with the periodicity equivalent to pitch period, which overlaps and distorts the underlying autocorrelation of the speech waveform. The true solutions of the autoregressive (AR) coefficients can be obtained only if the autocorrelation sequence equals that of the vocal tract system impulse response. This true solution, however, is approximately achieved for a periodic waveform having a long pitch period. As the pitch period of high-pitched speech is small, the periodic replicas cause "aliasing" of the autocorrelation sequence. Thus the low order autocorrelation coefficients are considerably different from those of system impulse response. This leads to the fact that the accuracy of CALP decreases as the fundamental frequency (F_0) of speech increases.

For voiced speech, one approach to avoid this problem is to analyze only the interval included within a duration of glottal closure of a pitch period [3,4] where the covariance method is applied. However, it is very difficult to find such an interval of appropriate length on natural speech especially on speech uttered by females or children. Even if such an interval is found, the duration of the interval may be very short. For example, for a high-pitched female of $F_0 = 400$ Hz at 10 kHz sampling rate this interval consists of 2.5 ms minus the duration of the glottal open phase, which is very short. The closed-phase method has been shown to give smooth formants contours in cases where the glottal close phase is at least 3 ms in duration [3]. If the covariances are computed from an extremely short interval, they could be in error, and the resulting spectrum might not accurately reflect the vocal tract characteristics [5]. Another main concern of the covariance methods is vulnerability to the stability of estimated AR filter which reduces the scope of such methods for many practical applications.

An improvement to these methods have been proposed in [6] where the idea has been modified and extended to multiple pitch periods. The linear prediction has been applied to only the speech samples with nearly zero excitations in each frame. These samples are selected by referring to the residual signals obtained by conventional covariance method. The selection criterion is based on a threshold logic which may in turn causes instability of the synthesis filter when the number of selected prediction equations become small. Attempts have also been made to decouple the F_0 effects from spectrum estimation by using weighting function of the prediction residuals [7,8]. Both the methods have been presented to perform quite well for formant estimation of synthetic vowels. The methods, however, can also easily be subject to the instability of the resulting AR filter. Thus if stability is required, methods based on autocorrelation function should be privileged.

Since the conventional autocorrelation method also suffers from "aliasing" due to speech periodicity, this effect needs to be removed from the autocorrelation function. One such approach is to deconvolve the vocal tract impulse response from speech signal using homomorphic filtering (real or complex cepstrum) which is free from the waveform periodicity. The deconvolved impulse response can result in good autocorrelation estimates.

In this paper, for the purpose of formant estimation we propose a use of the autocorrelation method with homomorphic filtering. The use of cepstrum analysis in combination with linear prediction, called homomorphic prediction, is not actually first in this paper. In [9], a homomorphic vocoder employing predictive coding has been proposed to reduce the bit rate with respect to its predecessor [10] from 7800 to 4,000 bits/s. In the analysis phase, predictive coefficients are estimated from the vocal tract impulse response (obtained through homomorphic deconvolution) and in the synthesis phase these coefficients are used to synthesize speech. In [11], additionally it has been shown that homomorphically estimated vocal tract impulse response can be used to estimate moving average parameters as well using the residual signal obtained by linear predictive inverse filtering. In this paper, we aim at formant estimation and elaborate the application of homomorphic prediction in a perspective of eliminating effects due to high F_0 of speech signal.

As mentioned earlier, true solutions for AR coefficients can be obtained only if the autocorrelation function equals that of the vocal tract impulse response. To accomplish this, the proposed method first obtains a minimum phase estimate of the vocal tract impulse response using homomorphic filtering. The autocorrelation function of the resulting impulse response estimate is thus free from the distortions caused by the overlapping of the periodic replicas (as occurred when autocorrelation is directly calculated from speech waveform). The AR coefficients obtained using the newly estimated autocorrelation sequence are very close to the true solutions. In fact, the estimated errors of formant frequency obtained using this approach are found to be negligibly effected even for very high F_0. We present experimental results on both synthetic and real speech and we observe that a suitable liftering window can provide very good robustness against higher F_0 of speech.

We organize the paper as follows. We define the problem in Section 2 and we propose our method in Section 3. Sections 4 and 5 illustrate results obtained using synthetic and natural speech respectively. Finally, Section 6 is on the concluding remarks.

2. PROBLEM IDENTIFICATION

Let us consider an all-pole impulse response $h[n]$. The z-transform of $h[n]$ is given by

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{1}$$

so that

$$h[n] = \sum_{k=1}^{p} \alpha_k h[n-k] + \delta[n] \tag{2}$$

where $\delta[n]$ is an impulse. Here, the gain of the impulse input is considered to be unity. By multiplying both sides of the above equation by $h[n-i]$, summing over n and noting that $h[n]$ is causal, it can be shown that

$$\sum_{k=1}^{p} \alpha_k r_h[i-k] = r_h[i], \quad 1 \le i \le p \tag{3}$$

where $r_h[i]$ is the autocorrelation of $h[n]$. Suppose now that a periodic waveform $s[n]$ is constructed by running a periodic impulse train through $h[n]$, thus

$$s[n] = \sum_{k=-\infty}^{\infty} h[n - kP] \tag{4}$$

where P is the pitch period. The normal equations associated with $s[n]$, windowed over multiple pitch periods, for an order p predictor, are given by

$$\sum_{k=1}^{p} \alpha_k r_n[i-k] = r_n[i], \quad 1 \le i \le p \tag{5}$$

where $r_n[\tau]$ is the autocorrelation function of the windowed $s[n]$ and can be shown to equal periodically repeated replicas of $r_h[\tau]$, i.e.,

$$r_s[\tau] = \sum_{k=-\infty}^{\infty} r_h[\tau - kP] \tag{6}$$

with decreasing amplitude due to the windowed version. Equation (6) implies that as the pitch period P decreases, the spacing between impulses decreases and $r_n[\tau]$ suffers from increasing distortion. Thus $r_n[\tau]$ can be thought of as "aliased" version of $r_h[\tau]$. When the "aliasing" is minor, the two solutions of Eqs. (3) and (5) are approximately equal. The accuracy of this approximation, however, decreases as the pitch period decreases, because the autocorrelation functions, repeated every P samples, overlap and distort the underlying desired $r_h[\tau]$. This effect becomes potentially more severe for higher-pitched speakers. This is illustrated by an example shown in Fig. 1. The low order autocorrelation coefficients in case of $F_0 = 100$ Hz in Fig. 1(b) look very similar with those of $r_h(\tau)$ in Fig. 1(a). In fact, this is the reason why CALP performs relatively better in estimating spectrum at lower F_0. The autocorrelation coefficients in case of $F_0 = 250$ Hz in Fig. 1(c), however, are considerably different from those of $r_h[\tau]$. This illustrates that accuracy of linear prediction

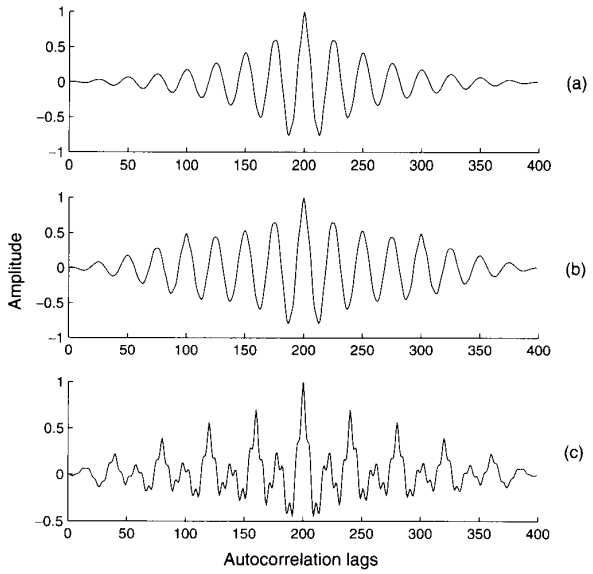

Fig. 1 Autocorrelation "aliasing" for windowed speech waveform. a) autocorrelation of the impulse response, $r_h[\tau]$; b) autocorrelation of a periodic waveform at $F_0 = 100$ Hz; c) same as (b) at $F_0 = 250$ Hz.

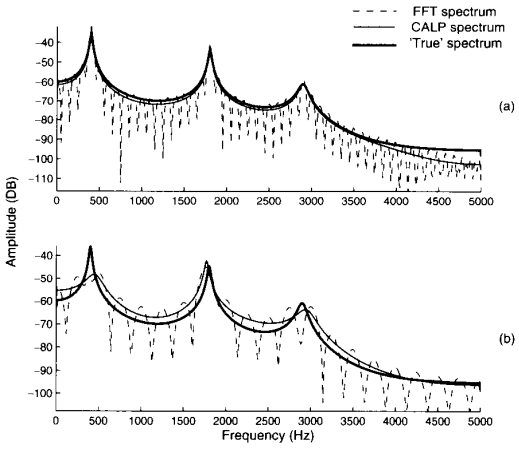

Fig. 2 Comparison of the spectra estimated using CALP with the 'true' spectrum. a) at $F_0 = 100$ Hz; b) at $F_0 = 250$ Hz.

analysis decreases with increasing F_0. In this example, $h[n]$ consists of three poles (at 400, 1,800, and 2,900 Hz) and $s[n]$ is the convolution of $h[n]$ with a periodic impulse train. The spectra estimated by CALP using the autocorrelation sequence of Fig. 1 are shown in Fig. 2 along with the FFT spectrum. By 'true' spectrum in Fig. 2, we mean the spectrum obtained from the autocorrelation of pure vocal tract impulse response $r_h(\tau)$. It is seen that the CALP spectrum estimated at $F_0 = 100$ Hz in Fig. 2(a) approximately overlaps with the 'true' spectrum. However, a clear deviation is observed in the CALP spectrum estimated at $F_0 = 250$ Hz in Fig. 2(b), which ascertains the effect of autocorrelation "aliasing" at higher pitch frequencies.

To summarize, from Eq. (3), we can say that if the autocorrelation function in the normal equations equals that of $h[n]$, then the solution must equal the α_k's of $H(z)$ in Eq. (1) and which are thus the true solutions.

3. THE PROPOSED SOLUTION

From Section 2, we understand that true solutions can be obtained if we have an estimate of $r_h[\tau]$. It means that $r_h[\tau]$ is the key to obtain solutions free from F_0 variations. In order to obtain $r_h[\tau]$ we, in turn, need a good estimate of the vocal tract impulse response $h[n]$. Though it is generally accepted that $h[n]$ can be separated from $s[n]$ using homomorphic filtering, unfortunately accuracy of the straightforward deconvolution approximation is limited by distortion which is induced from the repeated nature of the vocal tract contribution. The conditions for an absolute separation are still unknown. Nevertheless, a reasonable deconvolution approximation can be obtained by employing some means.

From Eq. (4), we can write

$$s[n] = h[n] * p[n] \tag{7}$$

where $p[n]$ is an impulse train with period P. Our target is to achieve the estimation of $h[n]$ as good as possible.

3.1. Deconvolution Method

We note that not all formulations of cepstral deconvolutions are appropriate for formant extraction. The discrete complex cepstrum using a N-point DFT is defined by [12]:

$$\hat{x}[n] = \frac{1}{N}\sum_{k=0}^{N-1} \log[X(k)] e^{j\frac{2\pi}{N}kn},\ 0 \le n \le N-1$$

where $X(k)$ is the Fourier transform of the speech signal $x[n]$ and $\log[X(k)] = \log(|X(k)|) + j\angle X(k)$. This formulation is not suitable for formant estimation because of its high sensitivity to phase [13]. Estimation of the complex cepstrum varies significantly depending on the positioning of analysis window.

Real cepstrum, on the other hand, defined by

$$c[n] = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)| e^{j\frac{2\pi}{N}kn},\ 0 \le n \le N-1$$

sets the phase terms effectively to zero and the sole magnitude is indeed enough for the resonant information to be restored. A 1,024-point DFT is sufficient to avoid DFT aliasing. The minimum phase counterpart of the real cepstrum can be obtained by multiplying the real cepstrum by the right-sided window $w[n]$ as:

$$\hat{x}[n] = w[n]c[n]$$

where,

$$w[n] = \begin{cases} 1, & n = 0; \\ 2, & 1 \le n \le \dfrac{N}{2}; \\ 0, & \dfrac{N}{2} + 1 \le n \le N-1; \end{cases}$$

In case of formant tracking application, both the real and minimum phase cepstrum result in similar estimates because the *log* magnitudes are same in both cases. When the vocal tract impulse response is extracted from the speech signal, however, the minimum phase counterpart of the real cepstrum makes the impulse response right-sided, whose energy is compressed toward origin. For the analysis/synthesis application minimum phase definition is also preferred for producing more natural sounding speech [12]. For these reasons, we use the minimum phase cepstrum in the current work.

3.2. Liftering Window

Since the liftering process contributes significantly to the degree of deconvolution of $h[n]$, it requires careful inspection. The length of the traditional liftering window (which is simply less than the pitch period) can introduce errors in formant estimation because the cepstrum coefficients closer to the pitch period location get distorted [14]. Moreover, there exists possibilities of pitch estimation errors. A liftering window having the length close to one pitch period can cause the inclusion of samples at or outside the actual pitch period within the liftering window. In that case we could not achieve the estimation free from F_0 variations. A liftering window with the length of $0.5P$ (50% of the pitch period) has been proposed in [14] with some compensating factors that minimize the distortions due to the windowing of speech waveform. In case of high-pitched speech, we found that a slight increase in the window length increases accuracy of formant estimation. We observe that a lifter of length $0.6P$ is best suited for analyzing synthetic speech signal with F_0 value up to 250 Hz and $0.7P$ for larger F_0 values. Tapering at the end of higher coeffieients can also be useful. It helps underweight the less important higher coefficients. The general form of the liftering window is given as in [15]:

$$l(n) = \begin{cases} 1, & n \le L_1 \\ 0.5(1 + cos[\pi(n - L_1)/\Delta L]), & L_1 < n < L_1 + \Delta L \\ 0, & n \ge L_1 + \Delta L \end{cases} \tag{8}$$

where $L = L_1 + \Delta L$ is the length of the liftering window. Equation (8) implies that only the higher part ΔL are tapered. Typical length of ΔL can be 25% of the total lifter length L. Tapering is, however, optionally used. In this paper, results are presented without using any tapering.

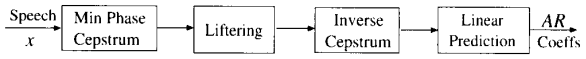

Fig. 3 Block diagram of the proposed method.

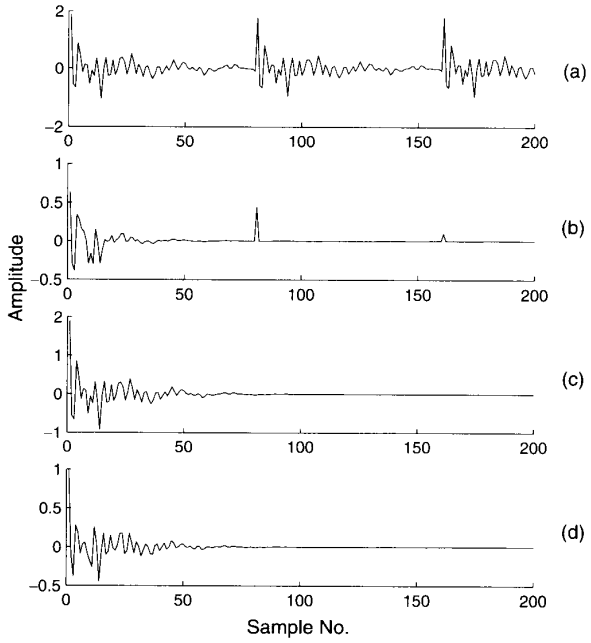

Fig. 4 Steps required by the proposed method. a) speech segment; b) obtained cepstrum; c) vocal tract impulse response, $h[n]$, in time-domain; d) autocorrelation function $r_h[\tau]$ of $h[n]$.

3.3. Definition of the Proposed Method

Based on the discussions presented so far, we propose our method as shown in the block diagram of Fig. 3.

Figure 4 illustrates graphically the steps required to obtain the AR coefficients of speech signal by using the proposed method. Figure 4(a) shows a segment of synthesized Japanese vowel /e/ sampled at 10 kHz. The synthetic speech is generated by convolving the train of impulses of different pitch frequencies with the vocal tract impulse response. Figures 4(b) and 4(c) show the obtained cepstrum and vocal tract impulse response which is transferred back to the time domain, respectively. Finally, Fig. 4(d) represents the autocorrelation of vocal tract impulse response. Clearly now the estimation is free from the distortion due to periodic replicas of $r_h[\tau]$ (as discussed in Section 2). Thus the estimation of AR coefficients from the resulting autocorrelation function is expected to be robust to F_0 variations. Figure 5 shows the spectra estimated using CALP and the proposed method from speech waveform synthesized at eight different values of F_0 (125, 150, 200, 250, 275, 300, 330, and 380 Hz) for the vowel sound /e/. It clearly depicts that the proposed method estimates the spectra very accurately even for $F_0 = 380$ Hz. Variations in the formant peaks are completely absent in all the spectra when compared with the 'true' spectrum. The 'true'

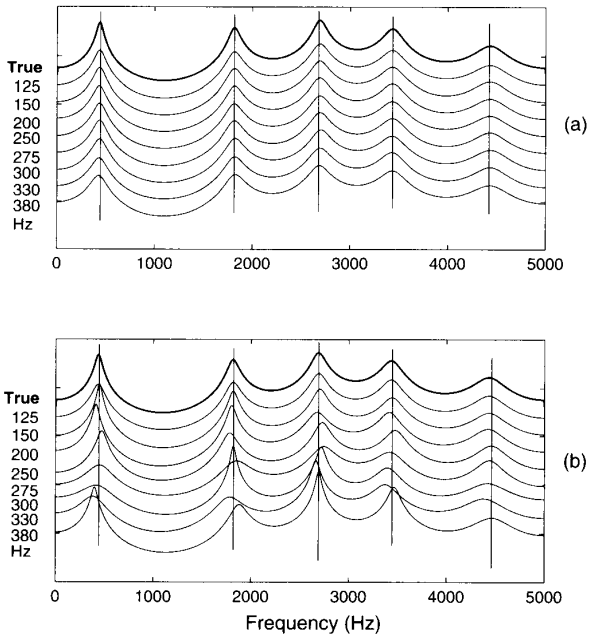

Fig. 5 Spectra estimated at 8 different pitch frequencies. (a) using proposed method; (b) using CALP.

spectrum has been obtained using the pure vocal tract impulse response which is used for synthesizing the vowel sound. In contrast, variations in the formant peaks are clearly evident in the spectra estimated using CALP. Though impulse train used in the above demonstration does not exactly represent the glottal volume velocity, it is a good representative to show the goodness of the method. In the next section, we present the results in more detail taking the glottal effects into consideration.

4. RESULTS ON SYNTHETIC SPEECH

The accuracy of the proposed method in estimating formant frequencies is verified on five synthetic Japanese vowels over a wide range of F_0. The popular Liljencrants-Fant glottal model [16] is used to simulate the source and the synthetic speech is sampled at 10 kHz. Since the purpose of this paper is to study the estimation against F_0 variations, all the parameters of the glottal model (open phase, close phase, and slope ratio) are kept constant for all values of F_0. The formant frequencies used here are shown in Table 1. Bandwidths of the five formants are set fixed to

Table 1 Formant frequencies used to synthesize vowels.

vowel	F_1	F_2	F_3	F_4	F_5 Hz
/a/	813	1,313	2,688	3,438	4,438
/i/	375	2,188	2,938	3,438	4,438
/u/	375	1,063	2,188	3,438	4,438
/e/	438	1,813	2,688	3,438	4,438
/o/	438	1,063	2,688	3,438	4,438

60, 100, 120, 175, and 281 Hz, respectively. A sample differencing operation is employed on the output of the formant synthesizer to simulate the radiation characteristics from lip. The analysis order is set to 12. The window used here is Hamming of length 20 ms. The speech is preemphasized by the filter $1-z^{-1}$. A cepstrum window of length $0.6P$ is used for speech waveform with the F_0 value up to 250 Hz and $0.7P$ for larger values of F_0. Formant values are obtained by the root-solving method of AR coefficients.

4.1. Estimating Formant Frequencies

In order to obtain a reasonable estimation of the formants concerning relative positions between the analysis window and the excitation point, analysis is conducted on different window positions. The final result is the arithmetic mean of results taken from all window positions. This is accomplished by shifting the frame by 0.1 ms (one speech sample) over the duration of one pitch period as in Eq. (9):

$$\hat{F} = \frac{1}{n}\sum_{d=1}^{n}\hat{F}(d) \tag{9}$$

where $\hat{F}(d)$ implies a formant estimated at d displacement from the excitation point and n is the number of different window positions. Finally, estimation error of the ith formant, EF_i, is calculated by averaging the individual F_i errors of all the five vowels. Thus we can express the estimation error EF_i as:

$$EF_i = \frac{1}{5}\sum_{j=1}^{5}|\hat{F}_{ij} - F_{ij}|/F_{ij} \tag{10}$$

where F_{ij} denotes the ith formant frequency of the jth vowel and $\hat{F}_{ij}$ is the corresponding estimated value.

The first and second formants are observed to be effected mostly by F_0 variations. The estimation errors of the first and second formant frequencies are shown in Figs. 6(a) and 6(b). It is seen that F_1 estimation error using CALP can be about 15% depending on the pitch frequency. Using the proposed method, on the otherhand, it is very smaller and looks very robust against even very high F_0. A more objective evaluation has been made by averaging the estimation errors of all the first three formants of the five vowels. We express it by

$$E = \frac{1}{15}\sum_{j=1}^{5}\sum_{i=1}^{3}|\hat{F}_{ij} - F_{ij}|/F_{ij} \tag{11}$$

Figure 7 shows the estimation error using Eq. (11). The estimation error using the proposed method is much smaller than JND (Just Noticeable Difference) which is 3–5%, reported by Flanagan in [17]. A close inspection of Fig. 6 and Fig. 7 suggests that the proposed method can be used

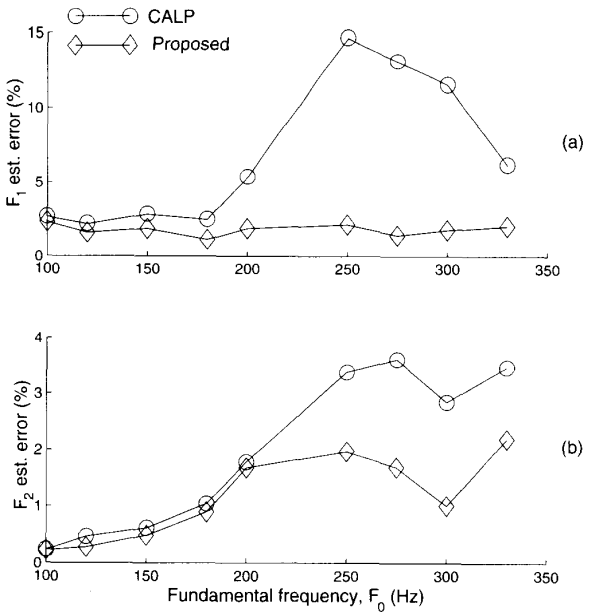

Fig. 6 Comparison of the estimated errors of formant frequencies at different values of F_0. (a) F_1 estimation error; (b) F_2 estimation error.

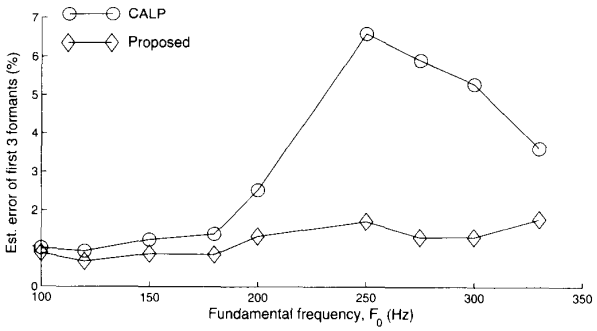

Fig. 7 Estimation error of first three formants.

for formant estimation for female and children speech with considerable improvement in estimation accuracy.

In Fig. 5(a), it is seen that the proposed method estimates the formant peaks perfectly without being affected by the pitch periods. In Fig. 7, however, we observe that the proposed method introduces error though very slightly. Actually this is due to the different synthesizing conditions of vowel sounds. The excitation source used for synthesizing sound of Fig. 5 is pure impulse train which is the ideal condition for cepstrum deconvolution. In case of Fig. 7, however, the glottal and radiation effects have been taken into consideration in synthesis process and as a result the source is no longer a pure impulse train. Thus the deconvolved vocal tract impulse response is not completely perfect.

In Fig. 7, it is seen that the estimation error obtained using CALP drops around $F_0 = 300$ Hz. This is somewhat unexpected but still can be explained by taking the philosophy of linear prediction into account. It is well

known that the linear prediction seeks to find an optimal fit to the speech spectrum. For voiced speech, accuracy of the fitness depends mostly on the frequency of the harmonics of speech spectrum and on the prediction order as well. An optimal fit can only be expected for relatively lower F_0 because more harmonics are available in the frequency domain. For higher F_0, however, the number of harmonics decreased and as a result formant peaks shift irregularly from its original position depending on the value of prediction order. Thus at some F_0, the estimation error can be slightly smaller, which is subjected to increase at an adjacent F_0.

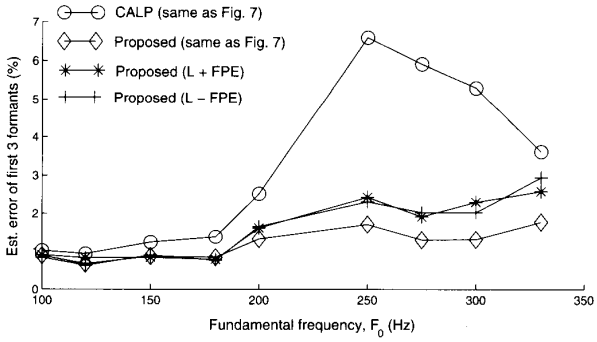

Fig. 8 Effect of fine pitch estimation error on the proposed method.

4.2. Effect of Pitch Estimation Errors on the Proposed Method

One disadvantage of the homomorphic filtering methods is that an estimate of the pitch period of the underlying speech waveform is required. Fortunately, many methods [18,19] have been proposed for estimating pitch period accurately based on cepstrum analysis. In case when the pitch period is not estimated exactly, the length of the liftering window will vary, that can affect the estimation of formant frequencies. Two error parameters GPE (Gross Pitch Error) and FPE (Fine Pitch Error) are commonly used as a measure of errors in estimating pitch period. The possible sources of GPE is pitch doubling, tripling, halving, inadequate suppression of formants as to affect the estimation, etc. One effective way to deal with the above situations is to verify the consistency of estimation. The estimated pitch period of nth frame, for example, can be compared with that of the $(n-1)$th and $(n+1)$th frame. The FPE, on the other hand, is attributed to the bias of measurement technique. The mean value of FPE is reported for seven pitch detection algorithms (including a cepstral method [18]) by Rabiner *et al.* in [20] as on the order of ± 0.5 samples across all utterances and speakers. Figure 8 demonstrates the effect of FPE as ± 1 sample for all F_0 used in Fig. 7.

It is seen that even if the pitch period can not be obtained to the utmost of precision, the proposed method still provides a good estimation accuracy in compared with CALP.

5. RESULTS ON REAL SPEECH

We present the result of analyzing two speech signals of a vowel sound /o/ spoken by a male and a high-pitched female speaker. The narrow-band spectrogram of vowel sound /o/ spoken by a male speaker is shown in Fig. 9(a). Some spectra estimated from the speech using the proposed method and CALP are shown in Figs. 9(b) and 9(c), respectively. The frame interval used here is 10 ms. The speech is preemphasized by the same filter as used for synthetic speech (i.e. $1-z^{-1}$). The prediction order is 12 and the frame size is adjusted to 4 pitch periods. The F_0 value of the speech signal is 160 Hz. In Figs. 9(a) and 9(b), it is seen that both methods estimate F_1 and F_2 well. Figure 10, on the other hand, is obtained using very high-pitched speech of the same vowel /o/. The F_0 value of this vowel sound is estimated as 352 Hz. From the spectrogram (Fig. 10(a)), it is seen that F_1 and F_2 become more closer than those in Fig. 9(a). This time CALP misses tracking F_2 which is depicted in Fig. 10(c). In contrast, the proposed method still estimates F_2 correctly as shown in Fig. 10(b).

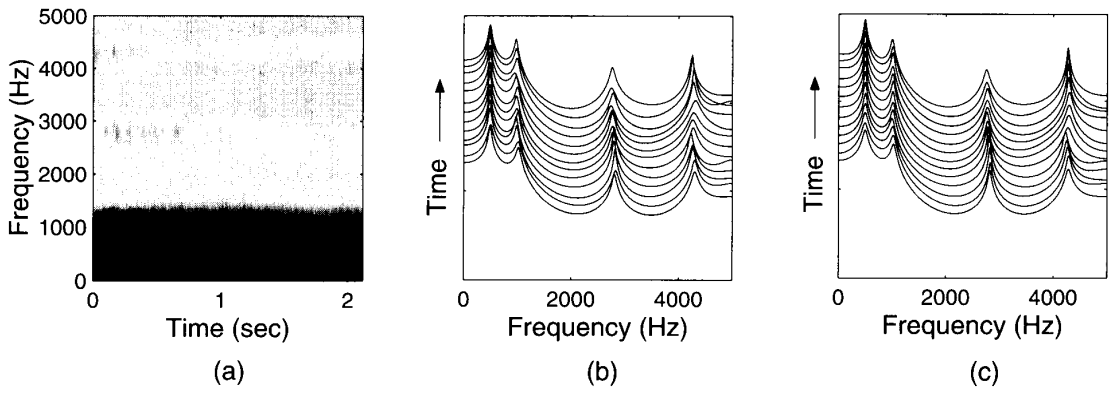

Fig. 9 Analysis of natural vowel /o/ spoken by a male speaker. a) narrow-band spectrogram; b) several consecutive spectra estimated using the proposed method; c) same as (b) estimated using CALP.

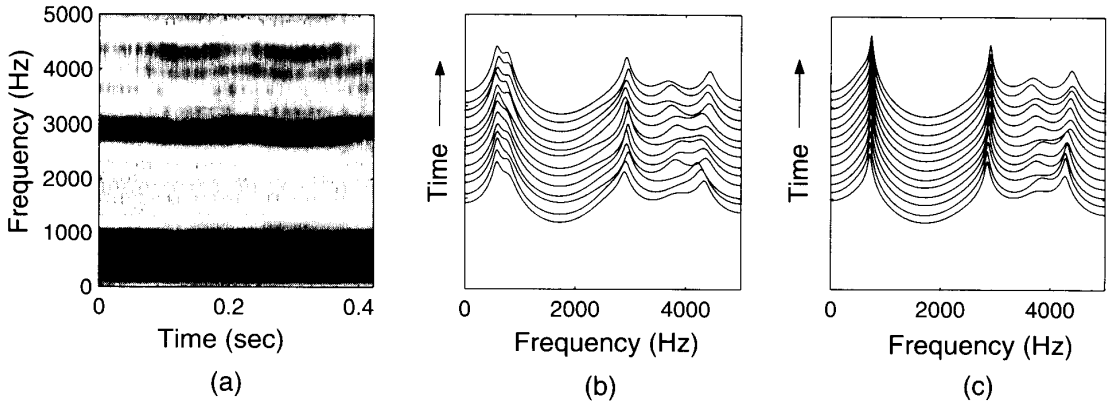

Fig. 10 Analysis of natural vowel /o/ spoken by a high-pitched female speaker. a) narrow-band spectrogram; b) several consecutive spectra estimated using the proposed method; c) same as (b) estimated using CALP.

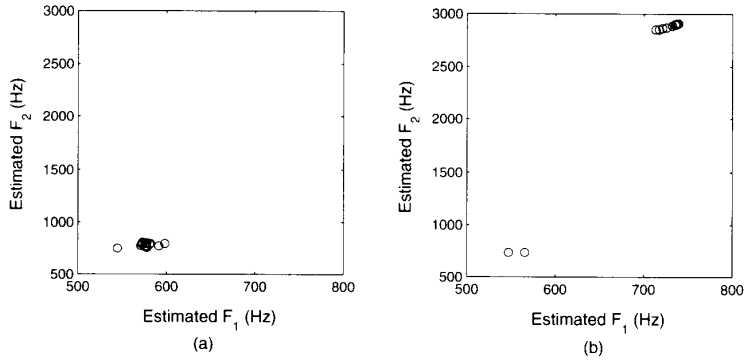

Fig. 11 F_1-F_2 plot of natural vowel /o/ spoken by a high-pitched female speaker. (a) estimated using the proposed method; (b) estimated using CALP.

The F_1-F_2 plot obtained using the proposed method and CALP are shown in Figs. 11(a) and Fig. 11(b) respectively. Except one point around (550,750) Hz in Fig. 11(a), the accuracy of the proposed method is satisfactory. In Fig. 11(b), it is observed that the third formant F_3 is treated as the second formant F_2. It implies that for high-pitched speech, CALP can easily be error prone in estimating the formant frequencies while the proposed method has the potential for much better estimation.

One of the greatest concerns for speech synthesis is the stability of the linear prediction synthesis filter. In [6], the estimated solutions, however, have been observed to be unstable at the rate of 18 percent out of 210 frames of natural vowel speech. The method described in [8] also incorporates a module for stability checking and in case of the filter being unstable, CALP is again referred to use. In contrast, the proposed method is guaranteed to produce a stable synthesis filter.

6. CONCLUSION

Practically, it is indeed very difficult to obtain a speech analysis method which is absolutely free from F_0 effects. We, however, expect from the discussion that the proposed technique can be applied to analyze speech data when the conventional model of linear prediction is only an approximation to speech signal uttered by female and children speakers. Though the method is intended for analyzing high-pitched speech signal, the results demonstrate that it can also be used for analyzing typical male speech with better accuracy. The estimation accuracy of the proposed technique depends on the degree of separation obtained through the deconvolutional algorithm. Thus, development of a better deconvolutional algorithm can lead to further enhancement of the accuracy of the proposed method.

REFERENCES

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, **50**, 637–655 (1971).

[2] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, **63**, 561–580 (1975).

[3] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust. Speech Signal Process.*, **34**, 730–743 (1986).

[4] H. W. Strube, "Determination of the instant of the glottal closure from the speech wave," *J. Acoust. Soc. Am.*, **56**, 1625–1629 (1974).

[5] N. B. Pinto, D. G. Childers and A. L. Lalwani, "Formant

speech synthesis: Improving production quality," *IEEE Trans. Acoust. Speech Signal Process.*, **37**, 1870–1887 (1989).
[6] Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida and O. Kakusho, "Analysis of speech signal of short pitch period by a sample-selective linear prediction," *IEEE Trans. Acoust. Speech Signal Process.*, **35**, 1233–1240 (1987).
[7] M. Yanagida and O. Kakusho, "A weighted linear prediction analysis of speech signals by using the Given's reduction," *Digital Signal Processing*, M. H. Hamza, Ed., *IASTED Int. Symp. Appl. Signal Processing and Digital Filtering*, Paris, pp. 129–132 (1985).
[8] C. H. Lee, "On robust linear prediction of speech," *IEEE Trans. Acoust. Speech Signal Process.*, **36**, 642–650 (1988).
[9] C. J. Weinstein and A. V. Oppenheim, "Predictive coding in a homomorphic vocoder," *IEEE Trans. Audio Electroacoust.*, **19**, 243–248 (1971).
[10] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Am.*, **45**, 458–465 (1969).
[11] G. E. Kopec, A. V. Oppenheim and J. M. Tribolet, "Speech analysis by homomorphic prediction," *IEEE Trans. Acoust. Speech Signal Process.*, **25**, 40–49 (1977).
[12] T. F. Quatieri, *Discrete-Time Speech Signal Processing* (Prentice Hall, Upper Saddle River, NJ, 2002).
[13] T. F. Quatieri, Jr., "Minimum and mixed phase speech analysis-synthesis by adaptive homomorphic deconvolution," *IEEE Trans. Acoust. Speech Signal Process.*, **27**, 328–335 (1979).
[14] W. Verhelst and O. Steenhaut, "A new model for the short-time complex cepstrum of voiced speech," *IEEE Trans. Acoust. Speech Signal Process.*, **34**, 43–51 (1986).
[15] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Am.*, **47**, 634–648 (1970).
[16] G. Fant, J. Liljencrants and Q. G. Lin, "A four parameter model of glottal flow," *Q. Progr. Stat. Rep.*, Speech Transmission Lab., Royal Inst. Technol. Oct.–Dec., pp. 1–13 (1985).
[17] J. L. Flanagan, "A difference limen for vowel formant frequency," *J. Acoust. Soc. Am.*, **27**, 613–617 (1955).
[18] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, **41**, 293–309 (1967).
[19] H. Kobayashi and T. Shimamura, "A modified cepstrum method for pitch extraction," *Proc. IEEE Asia-Pacific Conf. Circuits and Systems*, pp. 299–302 (1998).
[20] L. R. Rabiner, M. J. Cheng, A. E. Rosenburg and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust. Speech Signal Process.*, **24**, 399–418 (1976).

M. Shahidur Rahman received the B.Sc.(Hons) and M.Sc. degree in electronics and computer science from Shah Jalal University of Science and Technology, Sylhet, Bangladesh, in 1995 and 1997, respectively. In 1997, he joined Shah Jalal University as a junior faculty. Since October, 2003 he has been with Saitama University, Saitama City, Japan, to pursue Ph.D. degree in mathematical information systems. His current research interests include speech analysis, speech synthesis, and digital signal processing. He is a student member of IEEE.

Tetsuya Shimamura received the 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 City, Japan, where he is currently as Associate Professor. His research interests are in digital signal processing and applications to speech and communication systems. He is a member of IEEE and EURASIP.