Multi-closure-interval Linear Prediction Analysis Based on Phase Equalization

Sadao Hiroya*, Nobuhiro Miki† and Takemi Mochida*
* NTT Communication Science Laboratories, Kanagawa, Japan
E-mail: {hiroya.sadao, mochida.takemi}@lab.ntt.co.jp
† Future University Hakodate, Hokkaido, Japan
E-mail: miki@fun.ac.jp

***Abstract*—This paper presents a multi-closure-interval linear prediction (MCLP) analysis based on phase equalization in order to remove the effect of subglottal resonance in speech signals for estimating a vocal-tract spectrum. The validity of this method is evaluated by using a vocal-tract simulator that models vocal-tract losses and the subglottal system. Results show that the proposed method improves the estimation accuracy of a vocal-tract spectrum compared with the conventional MCLP method.**

I. Introduction

Linear prediction coding (LPC) is a fundamental technique for estimating a vocal-tract spectrum from speech signals. However, the estimated vocal-tract spectrum of voiced speech is affected by imperfect cancellation of source characteristics, such as harmonics and subglottal resonance. To overcome the problem, a pitch synchronous analysis [1], multi-closure-interval LP (MCLP) [2], and discrete all-pole (DAP) modeling [3] have been proposed. In these methods, a periodic impulse excitation is assumed for the voiced speech, and it is therefore possible to estimate a vocal-tract spectrum with small harmonics effect.

The subglottal cavity consists of the trachea and lungs. Usually, the analysis window length of 10-20 msec is used for LPC. However, it is known that a vocal-tract spectrum varies in the open/closed phase of the glottis due to the subglottal resonance [4][5]. Miki et al. [4] have proposed a very short-time analysis of speech signals based on the Fejér kernel mapping and shown that the vocal-tract spectrum of Japanese vowel /a/ varies depending on the closure/opening intervals of the glottis. Kitamura et al. [5] have analyzed vocal-tract MRI data of Japanese vowels and found that the vocal-tract spectrum in the frequency region from 3.0 to 3.7 kHz disappears when the glottis is open, but appears when it is closed. Recently, the second subglottal resonance was used for speaker adaptation since the subglottal resonance varies depending on the speaker [6].

In MCLP, several time-segments of speech signals corresponding to the closed phase of the glottis were utilized in order to get stability in vocal-tract spectrum estimation as well as to remove the effect of subglottal resonance. However, because natural speech signals include complex phase characteristics, there is some doubt whether speech signals corresponding to the closed phase of the glottis can be determined properly.

We have proposed an LP method based on phase equalization [7]. The phase equalization was proposed by Moriya and Honda to compensate for the phase characteristics of speech signals using a matched filter [8]. They found both the speech spectrum and the quality of the phase-equalized speech are almost equivalent to those of the original speech. The phase-equalized speech signals can be considered to be the output of the LPC filter whose input is the impulse train spaced at the pitch period. Therefore, we used the almost perfect impulse excitation properties of phase-equalized speech to remove the effects of harmonics in voiced speech. From our findings, we expected that phase equalization would be beneficial even for extracting speech signals corresponding to the closed phase of the glottis.

In this paper, we propose an MCLP based on a phase equalization from speech signals. This method consists of phase equalization for original speech signals and MCLP for the phase-equalized speech signals.

II. Speech synthesizer with subglottal system

To verify the proposed method, we need to know the 'true' vocal-tract and subglottal resonances in speech signals. A 'true' vocal-tract (supra-glottal) resonance for natural speech can be obtained by MRI measurements, but it is still difficult to measure a 'true' subglottal resonance. Thus, in this study, we used a speech synthesizer with the subglottal system proposed by Hayashi and Miki [9]. The vocal-tract shape data used are Japanese vowels /a/, /i/, and /u/ and extracted from MRI data of a Japanese male subject. A high-order FIR filter with frequency-dependent losses, such as viscous friction, heat conduction and wall admittance, was used for the vocal-tract model. The transmission line model was used for subglottal system. The infinite baffle model [10] was used as the radiation impedance and the two-mass model [11] as the vocal folds model. The synthesizer can generate steady-state vowels at a sampling rate of 179 kHz since the length of one section of the vocal tract is 0.2 cm. The results showed that the approximation accuracy is sufficient up to 6 kHz.

Figure 1(A), (E), and (F) show examples of the synthesized speech signals, the volume velocity at the glottis, and open/closed phase of the glottis of Japanese vowel /a/, respectively. The closed phase was delayed 0.45 msec from the closing instance of the glottis since the distance between

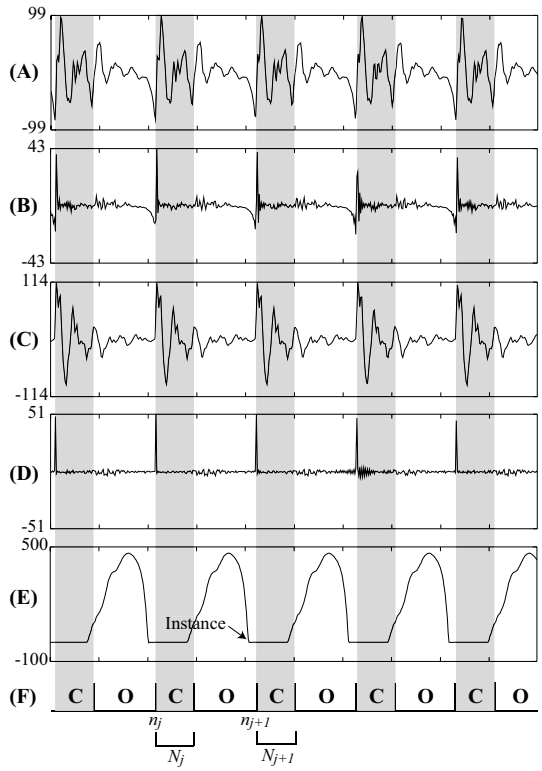

Fig. 1. *Examples of synthesized speech signals for the Japanese vowel /a/. (A) Original (synthesized) speech signals; (B) LPC residual signals; (C) phase-equalized speech signals; (D) phase-equalized LPC residual signals; (E) Volume velocity at the glottis; (F) open/closed phase of the glottis. O and C indicate open and closed phase, respectively.*

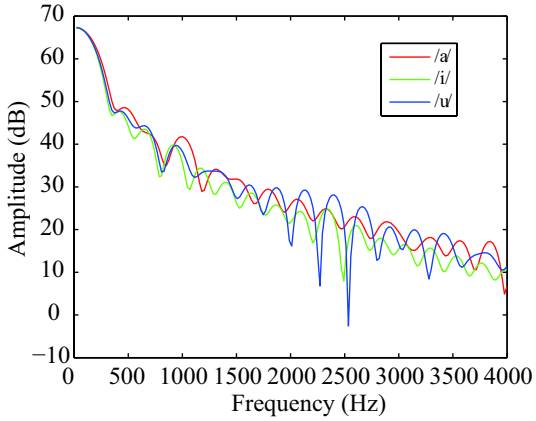

Fig. 2. *Spectra of the volume velocity at the glottis for vowels /a/ (red), /i/ (green) and /u/ (blue).*

the glottis and lips is 16 cm. The sampling rate was converted to 16 kHz. In MCLP, speech signals corresponding to the closed phase of the glottis are used, but large amplitudes in the open phase in Fig. 1(A) are seen for synthesized speech. Figure 2 shows spectra of the volume velocity at the glottis for Japanese vowels /a/, /i/, and /u/. These were obtained for a one-pitch period. Spectral peaks can be seen around 1000 Hz for /a/ and 2300 Hz for /i/ and /u/. These peaks would affect

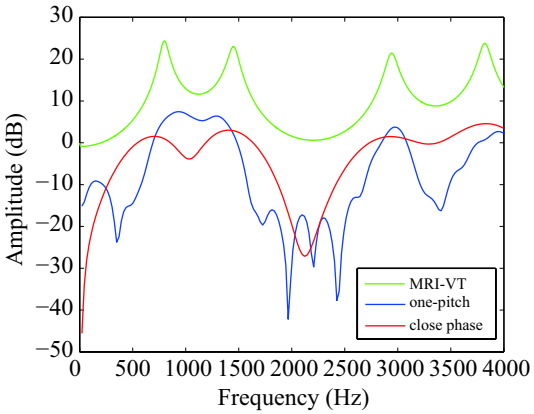

Fig. 3. *Vocal-tract spectra of the Japanese vowel /a/. 'True' spectrum from MRI data (green). FFT spectra from one-pitch speech signals (blue) and from closed phase (red).*

the estimation of the vocal-tract spectrum. Figure 3 shows the 'true' vocal-tract spectrum computed with MRI data (MRI-VT) and FFT spectra of synthesized speech signals for a one-pitch period and for a closed phase period of Japanese vowel /a/. The first two formant frequencies of the one-pitch period seem to be affected by the subglottal resonance. Therefore, multi-closure-interval analysis would be useful for estimating the vocal-tract spectrum.

III. Proposed Method

In this section, we describe the method for estimating a vocal-tract spectrum from the phase-equalized speech signals, based on the multi-closure intervals.

A. Phase equalization

In phase equalization, the idea is to convert the phase characteristics of the original speech signals to the minimum phase. This is done by applying an adaptively matched filter and converting the LPC residual signals to a nearly zero phase [8]. In the voiced speech frame, the LPC residual signals $e(t)$ are considered to be the impulse train of the pitch period:

$$e(t) = s(t) - \sum_{i=1}^{p} a(i)s(t-i), \tag{1}$$

where $s(t)$ represents the original speech signals, $a(i)$ represents the LPC coefficients, and p is the dimension of the LPC coefficients. However, the LPC residual signals for natural (and even synthesized) speech are not a zero-phase [Fig. 1 (B)]. Therefore, the impulse train is reconstructed from the filter output using the $M+1$ tap FIR filter $h(t)$ as follows. Provided one pulse exists at a known position t_0 in the frame for the sake of simplicity, the modeled input is represented as $\delta(t-t_0)$ and the reconstructed input $g(t)$ is expressed as

$$g(t) = \sum_{\tau=-M/2}^{M/2} h(\tau)e(t-\tau). \tag{2}$$

The optimum filter coefficients h are derived by minimizing the mean squared error between g and $\delta(t-t_0)$ in the frame:

$$\underset{h}{\operatorname{argmin}} \sum_t \left(\sum_{\tau=-M/2}^{M/2} h(\tau) e(t-\tau) - \delta(t-t_0) \right)^2. \quad (3)$$

If the autocorrelation function of e is a delta function for the time delay up to $M+1$, then

$$h(t) = e(t_0 - t) \Big/ \sqrt{\sum_{\tau=-M/2}^{M/2} e(t_0+\tau)^2}. \quad (4)$$

That is, the LPC residual is converted into a positive impulse train through the FIR filter whose coefficients are the values of the LPC residual itself, which is reversed at a reference position in the time domain. For the obtained h, the phase-equalized speech signals x are computed with

$$x(t) = \sum_{\tau=-M/2}^{M/2} h(\tau) s(t-\tau). \quad (5)$$

Here, the number of filter taps $M+1$ matches the pitch period in the frame. The positions of pitch marks $t_0, t_1, \cdots$ in the frame are detected on the basis of the LPC residual signals as in [8].

Figure 1 shows an example of the results of phase equalization. The phase-equalized LPC residual signals show very sharp pitch spikes at the instant corresponding to the pitch mark [Fig. 1 (D)]. The phase-equalized speech signals are considered to be the output of the LPC filter whose input is the impulse train spaced at the pitch period [Fig. 1(C)]. Large amplitudes are not shown in the open phase of phase-equalized speech.

B. MCLP

MCLP is based on a covariance method for estimating a vocal-tract spectrum from speech signals. In the covariance method, the LPC coefficients $\hat{a}$ are computed by minimizing the following prediction errors:

$$f(t) = x(t) - \sum_{i=1}^{p} \hat{a}(i) x(t-i); \quad (6)$$

that is, by solving the following simultaneous equation:

$$\Phi^T \Phi \hat{a} = \Phi^T \phi, \quad (7)$$

where

$$\phi = (x(p+1), x(p+2), \cdots, x(L))^T \quad (8)$$

$$\Phi = \begin{pmatrix} x(p) & x(p-1) & \dots & x(1) \\ x(p+1) & x(p) & \dots & x(2) \\ \vdots & \vdots & & \vdots \\ x(L-1) & x(L-2) & \dots & x(L-p) \end{pmatrix}. \quad (9)$$

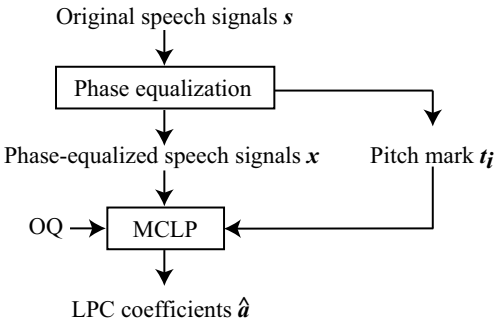

Fig. 4. *Algorithms for the proposed method.*

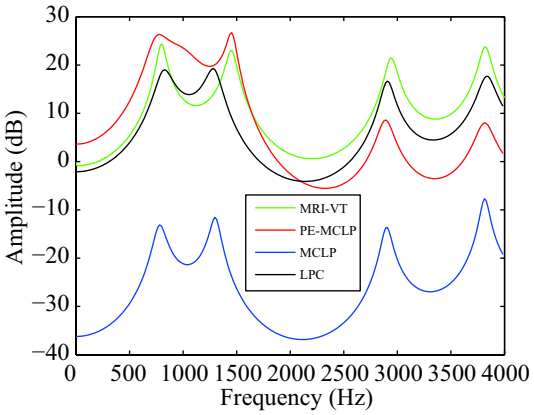

Fig. 5. *Vocal-tract spectra of the Japanese vowel /a/. 'True' spectrum from MRI data (green). Spectra from MCLP based on phase-equalized speech (PE-MCLP) (red), from MCLP (blue), and from conventional LPC (black).*

In MCLP, we define ϕ_j and Φ_j as

$$\phi_j = (x(n_j+p), \cdots, x(n_j+N_j-1))^T \quad (10)$$

$$\Phi_j = \begin{pmatrix} x(n_j+p-1) & \dots & x(n_j) \\ \vdots & & \vdots \\ x(n_j+N_j-2) & \dots & x(n_j+N_j-p-1) \end{pmatrix},$$

where n_j is the initial point of j-th closure phase interval in the analysis window, and N_j is the length of j-th closure interval (See Fig. 1). The LPC coefficients $\hat{a}$ are computed by solving the following simultaneous equations:

$$[{\Phi_1}^T \Phi_1 + \cdots + {\Phi_K}^T \Phi_K]\hat{a} = [{\Phi_1}^T \phi_1 + \cdots + {\Phi_K}^T \phi_K]. \quad (11)$$

C. Algorithms

Figure 4 shows the algorithms of the proposed method. For voiced speech of original speech signals, the phase-equalized speech signals and the pitch mark are computed. Pitch marks are obtained by using the LPC residual signals [12]. MCLP is applied to the multi-closure intervals of phase-equalized speech signals, which are computed with a predetermined open quotient (OQ) and pitch marks.

IV. Experiments

We evaluated the proposed method using synthesized steady-state vowels /a/, /i/, and /u/ obtained by the vocal-tract

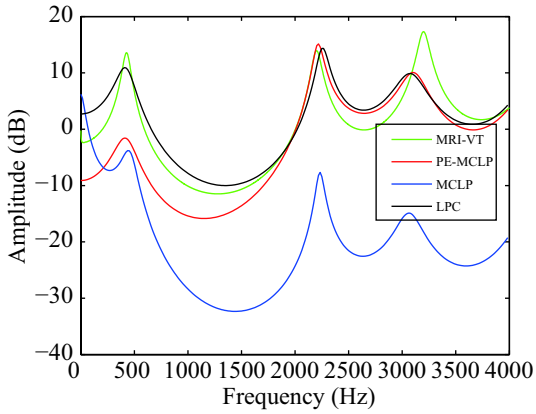

Fig. 6. *Vocal-tract spectra of the Japanese vowel /i/.*

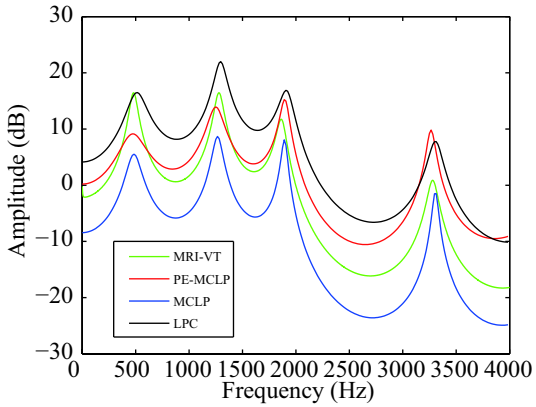

Fig. 7. *Vocal-tract spectra of the Japanese vowel /u/.*

simulator. The speech signals at a sampling rate of 179 kHz were converted to 16 kHz and pre-emphasized with first order differentiation. Eighteen LPC coefficients were obtained. We show results for intervals of steady-state oscillation of vocal folds. The fundamental frequencies were estimated from the instantaneous frequency amplitude spectrum [13] and were 155, 147, and 149 Hz for /a/, /i/, and /u/, respectively. The OQs obtained from the volume velocity were 62, 65, and 64 % for /a/, /i/, and /u/, respectively.

Figure 5 shows the 'true' vocal-tract spectrum computed with MRI data (MRI-VT) and the spectrum estimated using MCLP of the phase-equalized speech (PE-MCLP), MCLP of original speech (MCLP), and the covariance method of the original speech (LPC) of the Japanese vowel /a/. Analysis window length was a two-pitch period for PE-MCLP and MCLP, and 32 msec for LPC. We can see the second formant frequency (F2) of MCLP was affected by the subglottal resonance, but that of PE-MCLP was improved compared with the other methods. Figures 6 and 7 show the results for /i/ and /u/, respectively, which indicate there is not a large difference between PE-MCLP and other methods. These results indicate that the proposed method is beneficial for removing subglottal resonance around the frequency of 1 kHz.

V. Discussion

We applied MCLP to phase-equalized speech signals in order to reduce the effect of subglottal resonance. The results showed that the proposed method improved the estimation accuracy of vocal-tract spectrum compared to conventional MCLP: The phase equalization would be beneficial for reducing the effect of subglottal resonance.

Estimation of the vocal-tract spectrum based on phase equalization does not work well for a standard autocorrelation or covariance method. Therefore, a combination of phase equalization and source modeling, such as LPC with a modeling of excitation signals or multi-closure-interval LP, is important for the phase equalization-based analysis.

In this experiment, we predetermined the OQ. However, the finding that F4 is modulated depending on the open/closed intervals of speech signals [5] would be useful for determining the OQ from speech signals. Moreover, the F1 bandwidth of PE-MCLP in Fig. 5 was wide, which is most likely because the data length was insufficient for PE-MCLP. Further studies are needed.

VI. Conclusions

We presented a novel vocal-tract spectrum estimation method and showed that it is superior to conventional MCLP in terms of reducing the effect of subglottal resonance.

Acknowledgement

The authors thank Dr. H. Gomi for many useful and helpful discussions.

References

[1] Mathews, M.V., Miller, J.E., and David, E.E., "Pitch synchronous analysis of voiced sounds," J. Acoust. Soc. Am., 179, 1961.
[2] Lu, J., Murakami, H., and Kasuya, H., "Estimation of vocal tract transfer functions using multi-closure intervals linear prediction," IEICE Trans. Fundamental., 1011–1014, 1990.
[3] El-Jaroudi, A. and Makhoul, J., "Discrete all-pole modeling," IEEE Trans. Signal Processing, 411–423, 1991.
[4] Miki, N., Takemura, K., and Nagai, N., "A short-time speech analysis method with mapping using the Fejér kernel," IEICE Trans. Fundamental., 792–799, 1994.
[5] Kitamura, T., Takemoto, H., Adachi, S., Mokhtari, P., and Honda, K., "Cyclicity of laryngeal cavity resonance due to vocal fold vibration," J. Acoust. Soc. Am., 2239–2249, 2006.
[6] Wang, S., Lulich, S.M., and Alwan, A., "A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation," Proc. Interspeech, 1717–1720, 2008.
[7] Hiroya, S. and Mochida, T., "Phase equalization-based autoregressive model of speech signals," Proc. Interspeech, 42–45, 2010.
[8] Moriya, T. and Honda, M., "Speech coder using phase equalization and vector quantization," Proc. ICASSP, 1701–1704, 1986.
[9] Hayashi, K. and Miki, N., "Approximation method for time-domain simulation of the lossy vocal tract and evaluation of frequency-dependent losses during glottal source flow," Acoust. Sci. & Tech., 130–138, 2008.
[10] Rabiner, L.R. and Schafer, R.W., "Digital processing of speech signals," 1978.
[11] Ishizaka, K. and Flanagan, J.L., "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell Syst. Tech. J., 1233–1268, 1972.
[12] Miyoshi, Y., Yamato, K., Mizoguchi, R., Yanagida, M., and Kakusho, O., "Analysis of speech signals of short pitch period by a sample-selective linear prediction," IEEE Trans. Signal Processing, 1233–1240, 1987.
[13] Arifiant, D., Tanaka, T., Masuko, T., and Kobayashi, T., "Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency," IEICE Trans. Inf. & Syst., 2812–2820, 2004.