REAL-TIME ROBUST FORMANT TRACKING SYSTEM USING A PHASE EQUALIZATION-BASED AUTOREGRESSIVE EXOGENOUS MODEL

Hiroki Oohashi, Sadao Hiroya and Takemi Mochida

NTT Communication Science Laboratories, NTT Corporation

ABSTRACT

This paper presents a real-time robust formant tracking system for speech signals and electroglottography (EGG) signals using a real-time phase equalization-based autoregressive exogenous model (RT-PEAR). PEAR can estimate formant frequencies robustly even for speech with high fundamental frequencies using phase equalization preprocessing and LPC with an impulse train. To reduce the computational complexity of original PEAR, a novel formulation of LPC with an impulse train is derived. EGG signals were used for stable detection of pitch marks since PEAR requires them. Formant estimation errors for the proposed method were less than 5 % regardless of fundamental frequencies with 12-ms processing delay. This technique will be useful for real-time speech conversion and speech-language therapy.

Index Terms— LPC, formant frequency, phase equalization

1. INTRODUCTION

Linear prediction coding (LPC) is a fundamental technique for the estimation of formant frequencies from speech signals. However, the estimated formant frequencies of voiced speech are affected by harmonics, because the model assumes Gaussian noise as the excitation signals even for voiced speech. To overcome the problem, methods based on the modeling of excitation signals for voiced speech have been proposed. One of the methods is discrete all-pole (DAP) modeling [1], which assumes a periodic impulse excitation in LPC for voiced speech. Another is LPC with a glottal source hidden Markov model (HMM) [2]. These methods are robust to harmonics but have high computational complexity and need around ten iterations. One of the reasons is that the phase characteristics are different between the speech production model (minimum phase) and natural speech. To reduce the computational complexity of robust estimation of formant frequencies using LPC, modifying the phase characteristics of natural speech so that they fit into a simple periodic impulse excitation model would be beneficial.

Hiroya and Mochida have proposed a phase equalization-based autoregressive exogenous model (PEAR) of speech signals and showed that a robust vocal-tract spectrum can be obtained using it [3]. Phase equalization is way to compensate phase characteristics of speech signals using a matched filter [4]. Both the speech spectrum and the subjective quality of the phase-equalized speech are almost equivalent to those of the original speech: The human auditory perception is less sensitive to short-term phase characteristics of speech signals. 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. Due to the phase equalization, an iteration is hardly necessary for PEAR.

A real-time formant tracking system would be important for investigating human speech-production mechanisms [5, 6] and for speech-language therapy. However, there are few studies on real-time robust formant tracking. Thus, PEAR may be effective for these purposes, but further reduction in computational complexity and stable pitch mark extraction are required for real-time PEAR (RT-PEAR).

In this paper, we report a format tracking system using an RT-PEAR and show that robust formant frequencies are obtained with the proposed method in real time.

2. 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 converting the LPC residual signals to a nearly zero phase [4]. In the voiced speech frame, the LPC residual signals $e(t)$ are considered to be the impulse train of the pitch period: $e(t) = s(t) - \sum_{p=1}^{P} a(p)s(t-p)$, where $s(t)$ represents the original speech signals, $a(p)$ denotes the LPC coefficients, and P is the dimension of the LPC coefficients. However, the LPC residual signals for natural speech are not a zero-phase [Fig. 1 (C)]. So 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 $\sum_{\tau=-M/2}^{M/2} h(\tau)e(t-\tau)$. The optimum filter coefficients h are derived by minimizing the mean squared error between them in the frame: $\underset{h}{\operatorname{argmin}} \sum_t (\sum_{\tau=-M/2}^{M/2} h(\tau)e(t-\tau) - \delta(t-t_0))^2$. If the autocorrelation function of e is a delta function for the time delay up to $M+1$, then

$$h(t) = e(t_0 - t) \Big/ \sqrt{\sum_{\tau=-M/2}^{M/2} e(t_0+\tau)^2}. \qquad (1)$$

That is, the LPC residual is converted into a positive impulse train through the FIR filter whose coefficients are the values

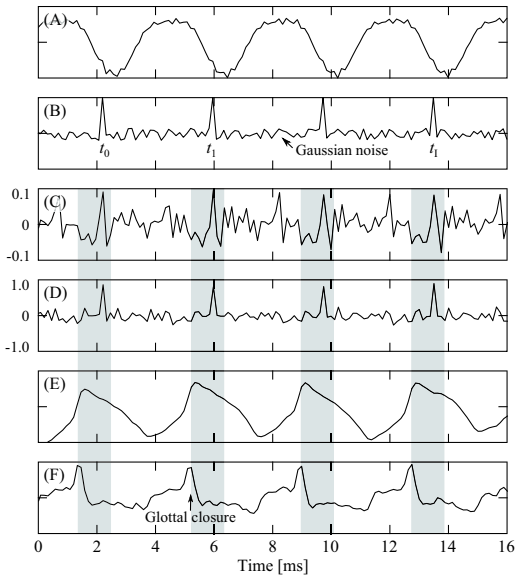

Fig. 1. *Examples of waveforms for the Japanese vowel /i/. (A) Original speech signals; (B) excitation signal model for PEAR; (C) LPC residual signals; (D) phase-equalized LPC residual signals; (E) EGG signals; (F) ΔEGG signals.*

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 by

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

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)].

3. PROPOSED METHOD

3.1. Original PEAR

The phase-equalized speech signals are considered to be the output of the LPC filter whose input is the impulse train corresponding to pitch mark $t_0, \cdots, t_I$ and the Gaussian noise elsewhere in the frame [Fig. 1 (B)]. Thus, we consider minimizing the following function:

$$\sum_{t \neq t_0,\cdots,t_I} (x_w(t) - \sum_{p=1}^{P} \hat{a}(p)x_w(t-p))^2 + \sum_{t=t_0,\cdots,t_I} (x_w(t) - \sum_{p=1}^{P} \hat{a}(p)x_w(t-p) - G_w(t))^2, \quad (3)$$

where $G(t_i)$ for $i = 0, \cdots, I$ is the impulse amplitude, $*_w$ is the windowed signal and $I+1$ is the number of impulses in the frame. The LPC coefficients $\hat{a}$ are calculated by solving the following simultaneous equation:

$$\begin{pmatrix} R_{xx}(0) & \dots & R_{xx}(P-1) \\ \vdots & \ddots & \vdots \\ R_{xx}(P-1) & \dots & R_{xx}(0) \end{pmatrix} \begin{pmatrix} \hat{a}(1) \\ \vdots \\ \hat{a}(P) \end{pmatrix} = \begin{pmatrix} R_{xx}(1) - \sum_{i=0}^{I} x_w(t_i-1)G_w(t_i) \\ \vdots \\ R_{xx}(P) - \sum_{i=0}^{I} x_w(t_i-P)G_w(t_i) \end{pmatrix}, \quad (4)$$

where R_{xx} is an autocorrelation function of the windowed phase-equalized speech signals x_w:

$$R_{xx}(q) = \sum_{t=0}^{L-1} x_w(t)x_w(t+q), \quad (5)$$

where L is the window length. As Eq. (4) is a Toeplitz matrix, we can use the Levinson algorithm to efficiently solve the equation [7]. The impulse amplitude is obtained so that Eq. (3) is minimized:

$$G(t_i) = x(t_i) - \sum_{p=1}^{P} \hat{a}(p)x(t_i-p)w(t_i-p)/w(t_i), \quad (6)$$

where w is the window function. Therefore, we determine the LPC coefficients and the amplitude iteratively, but we find iteration is hardly necessary [3]. If $G(t) = 0$ for all t, e.g., the unvoiced speech, then Eq. (4) is equivalent to the conventional LPC, i.e., the autocorrelation method for phase-equalized speech signals.

3.2. RT-PEAR

In Eq. (4), calculations of phase-equalized speech signals and their autocorrelation function and the impulse amplitude are required. To reduce the computational complexity, we introduce the following assumptions. By substituting Eqs. (1) and (2) into Eq. (5) under the assumption that the autocorrelation function of e is a delta function for the time delay up to $M+1$, $R_{xx}(q) = \sum_{t=0}^{L-1} s_w(t)s_w(t+q) = R_{ss}(q)$. Moreover, let $w(t_i-p)$ be $w(t_i)$, then Eq. (6) can be approximated as $G(t_i) \simeq \sqrt{\sum_{\tau=-M/2}^{M/2} e(t_i-\tau)^2}$. Therefore,

$$\begin{aligned} V(p) &= \sum_{i=0}^{I} x_w(t_i-p)G_w(t_i) \\ &= \sum_{i=0}^{I} w(t_i-p)w(t_i)G(t_i) \sum_{\tau=-M/2}^{M/2} h(\tau)s(t_i-p-\tau) \\ &\simeq \sum_{i=0}^{I} w(t_i-p)w(t_i) \sum_{\tau=-M/2}^{M/2} e(t_i-\tau)s(t_i-p-\tau). \end{aligned} \quad (7)$$

The LPC coefficients $\hat{a}$ are obtained by solving the following equation:

$$\begin{pmatrix} R_{ss}(0) & \dots & R_{ss}(P-1) \\ \vdots & \ddots & \vdots \\ R_{ss}(P-1) & \dots & R_{ss}(0) \end{pmatrix} \begin{pmatrix} \hat{a}(1) \\ \vdots \\ \hat{a}(P) \end{pmatrix} = \begin{pmatrix} R_{ss}(1) - V(1) \\ \vdots \\ R_{ss}(P) - V(P) \end{pmatrix}. \quad (8)$$

Note that phase-equalized speech signals and their autocorrelation functions and G are not included in Eq. (8). The left-hand side matrix has already been decomposed in the Levinson-Durbin algorithm [7] for conventional LPC. Since G does not exist, an iteration is not necessary. Thus, computational complexity in RT-PEAR is smaller than that in the original PEAR.

3.3. Pitch mark extraction

A pitch mark refers to the closure timing of the glottis. In [3], the positions of pitch marks $t_0, \cdots, t_I$ in the frame are detected on the basis of the LPC residual signals as in [4]. However, pitch-mark extraction had a problem for speech with high fundamental frequency (F0). Thus, we combine a derivative of electroglottography (EGG) signals. Concretely, pitch-mark detection in [4] was conducted for around few samples from the closure timing of the glottis obtained by ΔEGG signals (gray in Fig. 1).

3.4. TANDEM window

Even when RT-PEAR is applied to estimate a vocal-tract spectrum, the obtained spectrum is not temporally stable. Kawahara has found that the temporally stable power spectrum of a periodic signal can be calculated as the average of two power spectra by using a pair of time windows temporally separated for half of the fundamental period, called a TANDEM window [8]. According to the Wiener-Khinchin theorem, the power spectrum is the Fourier transform of the corresponding autocorrelation function. Thus, we can apply the TANDEM window with RT-PEAR as follows: We use the average of two autocorrelation functions and the average of two terms of V in Eq. (8) for the temporally separated windows.

3.5. Algorithms

Conventional LPC calculates the autocorrelation function R_{ss}, the LPC coefficients, and the LPC residual signals from speech signals. Then, for voiced speech, the pitch marks are obtained by using the LPC residual signals and EGG signals. For the pitch marks and the LPC residual signals, we determine the LPC (RT-PEAR) coefficients by Eq. (8).

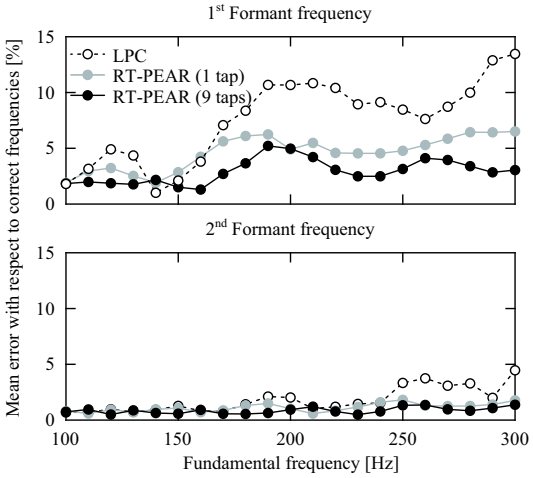

Fig. 2. *Mean percent errors in F1 and F2 for the five Japanese vowels.*

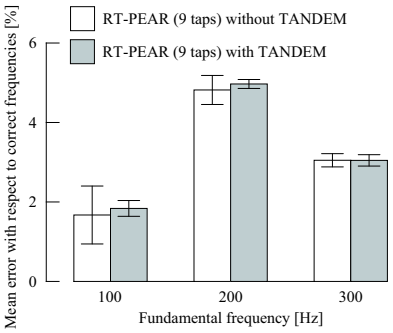

Fig. 3. *Mean percent errors and their standard deviations in F1 for F0s.*

4. EXPERIMENTS

We constructed the RT-PEAR system with a Renesas SH7785 and evaluated the proposed method using synthesized and natural vowels. This microprocessor used an SH-4A CPU core with a maximum operating frequency of 600 MHz and realizes a processing performance of 1080 MIPS. The speech signals were recorded at a sampling rate of 8 kHz and pre-emphasized with first-order differentiation. Eight LPC coefficients were obtained with a 4-ms frame shift using a 16-ms Blackman window. No lag window was used.

4.1. Results for synthesized vowels

The five Japanese vowels /a,i,u,e,o/ were examined. These steady-state vowels were synthesized from the first four formant frequencies, their bandwidths, and F0 using the Klatt formant synthesizer [9]. Duration was 2 sec. In this system, EGG signals are also required. Thus, we created quasi-EGG signals spaced at pitch period and used them.

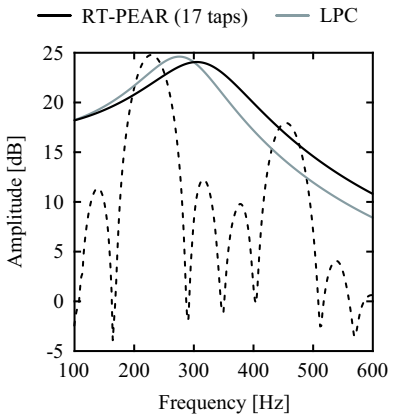

Fig. 4. *Vocal-tract spectrum of Japanese vowel /i/ using LPC (thin line) and RT-PEAR (17 taps; thick line). Speech spectrum (dashed line).*

Figure 2 shows the mean percent errors in F1 and F2 for the five vowels for F0 using LPC and PEAR (1 tap and 9 taps) with a TANDEM window. For LPC, the errors in F1 increased for with the value of F0. On the other hand, the results of PEAR (9 taps) showed the error was less than 5 % regardless of F0. PEAR (1 tap) means that phase equalization was not conducted for LPC residual signals and the errors of PEAR (1 tap) were larger than those of PEAR (9 taps), indicating phase equalization is effective for reducing the errors. The errors in F2 had small differences between methods.

Figure 3 shows their standard deviations with/without a TANDEM window. The standard deviations with a TANDEM window were smaller than those without it, in particular for low F0s. This indicated that the TANDEM window improves temporal stability in formant frequencies for low F0s.

4.2. Results for natural speech

Figure 4 shows the vocal-tract spectrum of the Japanese vowel /i/ for the conventional LPC and RT-PEAR with a TANDEM window for a female speaker. The average fundamental frequency was 229 Hz. We can see that the first peak of the vocal-tract spectrum (F1) of LPC was closer to first harmonics than that of RT-PEAR, indicating that RT-PEAR was less biased toward harmonics than LPC. This figure also shows that the envelope estimated by RT-PEAR fitted closer to the harmonics peaks, as in another robust estimation method [1].

5. DISCUSSION

Computational complexity is evaluated in terms of the number of products in algorithms (Table 1). The number of quotients is negligibly small. S means a frame shift size. For $I = 3$ and $M = 4$, the computational complexities of PEAR, RT-PEAR and TANDEM RT-PEAR are 2.8, 1.4, and 2.6 times as large as that of LPC, respectively. Thus, the computational complexity of RT-PEAR is half of that of original PEAR.

Table 1. Computational complexity

	Number of products
LPC	$O(LP + P^2)$
PEAR (LPC+Eq.(4))	$O(2LP + 3P^2 + SP + IM + LM + PI)$
RT-PEAR (LPC+Eq.(8))	$O(LP + 2P^2 + SP + PIM + 2PI + PM)$
TANDEM RT-PEAR	$O(2LP + 2P^2 + SP + 2PIM + 4PI + 2PM)$

Because of the low computational complexity using RT-PEAR, the processing delay using it was 12 ms. The delay is small enough for transformed auditory feedback experiments [5, 6].

6. CONCLUSIONS

We presented a real-time (12-ms delay) robust formant tracking system using RT-PEAR and showed that RT-PEAR with more than one tap is superior to conventional LPC in terms of robust estimation of formant frequencies to F0. The optimal number of taps and precise detection of pitch marks without EGG are an issue for the future.

7. ACKNOWLEDGEMENTS

The authors thank H. Uchida of Univ. of Tokyo for help with programs.

8. REFERENCES

[1] El-Jaroudi, A. and Makhoul, J., "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, pp. 411–423, 1991.

[2] Sasou, A. and Tanaka, K., "Glottal excitation modeling using HMM with application to robust analysis of speech," in *ICSLP*, 2000, pp. 704–707.

[3] Hiroya, S. and Mochida, T., "Phase equalization-based autoregressive model of speech signals," in *Interspeech*, 2010, pp. 42–45.

[4] Honda, M., "Speech coding using waveform matching based on LPC residual phase equalization," in *ICASSP*, 1990, pp. 213–216.

[5] Purcell, D.W. and Munhall, K.G., "Compensation following real-time manipulation of formants in isolated vowels," *J. Acoust. Soc. Am.*, pp. 2288–2297, 2006.

[6] Villacorta, V.M., Perkell, J.S., and Guenther, F.H., "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," *J. Acoust. Soc. Am.*, pp. 2306–2319, 2007.

[7] Golub, G.H. and van Loan, C.F., *Matrix computations (3rd ed.)*, The Johns Hopkins University Press, 1996.

[8] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., and Banno, H., "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum," in *ICASSP*, 2008, pp. 3933–3936.

[9] Klatt, D.H., "Software for cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, pp. 971–995, 1980.