Proc. Interspeech, pp. 42-45, 2010.

Phase Equalization-Based Autoregressive Model of Speech Signals

Sadao Hiroya, Takemi Mochida

NTT Communication Science Laboratories, NTT Corporation, Japan
`hiroya@idea.brl.ntt.co.jp, mochida@idea.brl.ntt.co.jp`

Abstract

This paper presents a novel method for estimating a vocal-tract spectrum from speech signals, based on a modeling of excitation signals of voiced speech. A formulation of linear prediction coding with impulse train is derived and applied to the phase-equalized speech signals, which are converted from the original speech signals by phase equalization. Preliminary results show that the proposed method improves the robustness of the estimation of a vocal-tract spectrum and the quality of re-synthesized speech compared with the conventional method. This technique will be useful for speech coding, speech synthesis, and real-time speech conversion.

Index Terms: LPC, vocal-tract spectrum, phase equalization

1. Introduction

Linear prediction coding (LPC) (or the autoregressive model: AR) is a fundamental technique for the estimation of a vocal-tract spectrum from speech signals and has been widely used for speech synthesis and speech coding. However, the estimated vocal-tract spectrum of voiced speech is affected by harmonics, because the model assumes Gaussian noise as the excitation signals even for voiced speech. This causes degradation of the re-synthesized speech quality. To overcome the problem, methods based on an adjustment of the analysis window or on modeling of excitation signals for voiced speech have been proposed. The former refers to a pitch synchronous analysis [1] and a glottal closure interval analysis [2], but the problems still remain: the analysis window is not long enough and estimating glottal closure intervals is difficult. One of the latter is discrete all-pole (DAP) modeling [3] which assumes a periodic impulse excitation in LPC for voiced speech, but the assumption is not satisfied for natural speech. Another is LPC with a glottal source hidden Markov model (HMM) [4]. The method is robust but has high computational complexity because the glottal source HMM requires the estimation of many parameters (e.g. the mean and the covariance of fifteen HMM states for each vowel) in order to represent phase characteristics of speech signals. To reduce the computational complexity of robust estimation of vocal-tract spectrum using LPC, modifying phase characteristics of natural speech to be fitted into a simple periodic impulse excitation model would be beneficial.

Moriya and Honda have proposed a method for compensating phase characteristics of speech signals using a matched filter, called phase equalization [5]. They found both the speech spectrum and the quality of the phase-equalized speech is almost equivalent to those of the original speech, respectively: humans cannot distinguish changes in short-time phase characteristics of speech signals. Their method is convenient compared with a phase compensation using an all-pass filter [6]. 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. Thus, it is expected that a formulation of LPC with an impulse train would be easy, and then the computational complexity would be reduced.

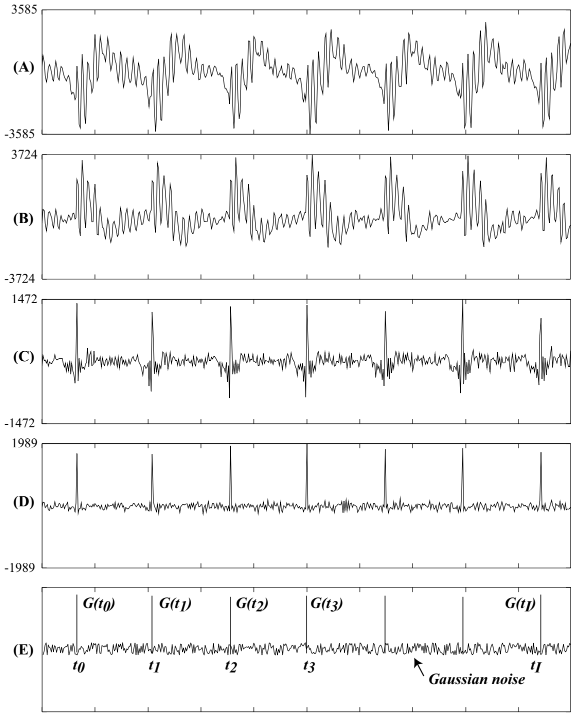


Figure 1: *Examples of the English vowel /i/. (A) Original speech signals; (B) phase-equalized speech signals; (C) LPC residual signals; (D) phase-equalized LPC residual signals; (E) excitation signal model for PEAR.*

In this paper, we propose a phase equalization-based autoregressive (PEAR) model of speech signals and show that a robust vocal-tract spectrum is obtained from the phase-equalized speech signals using PEAR.

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 applying an adaptively matched filter and converting the LPC residual signals to a nearly zero phase [5]. In the voiced speech frame, the LPC residual signals $e(t)$ are considered to be the impulse train of the pitch period:

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

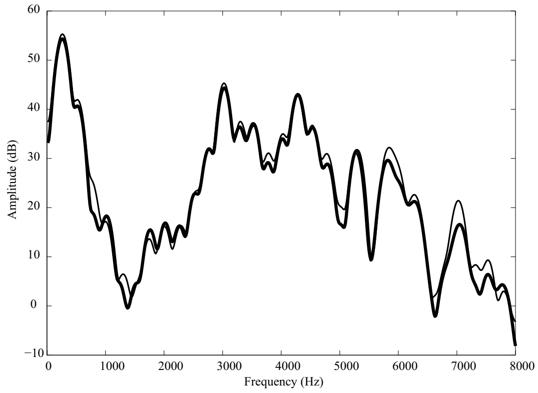


Figure 2: *Speech spectrum of the original speech signals (thin line) and the phase-equalized speech signals (thick line) of the English vowel /i/.*

where $s(t)$ is the original speech signals, $a(i)$ is 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 $g(t)$ 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:

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

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

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

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. To reduce abrupt change in the equalizing filter response, the FIR filter coefficients are low-pass filtered [7]. For the obtained h, the phase-equalized speech signals x are computed by

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

Here, the number of filter taps $M+1$ matches the pitch period in the frame. The positions of pitch mark $t_0, t_1, \cdots$ in the frame are detected on the basis of the LPC residual signals as in [7].

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)]. Then, the phase-equalized speech signals are approximately the minimum-phase sequence [Fig. 1 (B)]. Figure 2 shows an example of speech spectra of the original speech signals and phase-equalized speech signals. We can see that the spectrum of the phase-equalized speech signals is almost the same as that of the original ones.

3. Proposed method

Phase equalization has been used to optimize the excitation signals of voiced speech for low-bit rate speech coding [5, 7, 8, 9], but not to estimate the vocal-tract spectrum of voiced speech. In this section, we describe our method for estimating a vocal-tract spectrum from the phase-equalized speech signals, based on the modeling of excitation signals of voiced speech.

3.1. 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 (E)]. Thus, we consider minimizing the following function:

$$\sum_{t\neq t_0,\cdots,t_I} \sigma^{-1} f(t)^2 + \sum_{t=t_0,\cdots,t_I} \sigma^{-1}(f(t)-G(t))^2, \tag{6}$$

where $G(t)$ for $t = t_0, \cdots, t_I$ is the impulse amplitude, σ is the covariance, and $I+1$ is the number of impulses in the frame. The phase-equalized LPC residual signals f are calculated from the phase-equalized speech signals x like in Eq. (1):

$$f(t) = x(t) - \sum_{i=1}^{p} \hat{a}(i)x(t-i). \tag{7}$$

The LPC coefficients $\hat{a}$ are calculated by solving the following simultaneous equations:

$$\begin{pmatrix} R(0) & \ldots & R(p-1) \\ R(1) & \ldots & R(p-2) \\ \vdots & \ddots & \vdots \\ R(p-1) & \ldots & R(0) \end{pmatrix} \begin{pmatrix} \hat{a}(1) \\ \hat{a}(2) \\ \vdots \\ \hat{a}(p) \end{pmatrix} = \begin{pmatrix} R(1) - \sum_{i=0}^{I} x(t_i-1)G(t_i) \\ R(2) - \sum_{i=0}^{I} x(t_i-2)G(t_i) \\ \vdots \\ R(p) - \sum_{i=0}^{I} x(t_i-p)G(t_i) \end{pmatrix}, \tag{8}$$

where R is an autocorrelation function of the windowed phase-equalized speech signals x:

$$R(q) = \sum_{t=0}^{L-1} x(t)x(t+q), \tag{9}$$

where L is the window length. As Eq. (8) is a Toeplitz matrix, we can use the Levinson algorithm to efficiently solve the equations [10]. For given LPC coefficients, the optimum impulse amplitude is obtained from the phase-equalized LPC residual signals like in Eq. (7): $G(t) = f(t)$ for $t = t_0, \cdots, t_I$. Therefore, we first calculate the phase-equalized LPC residual signals

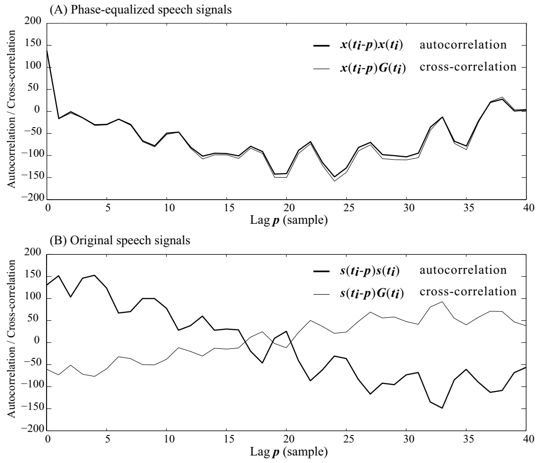


Figure 3: *Examples of the English vowel /i/. (A) Autocorrelation function for the phase-equalized speech signals (thick line) and cross-correlation function between the impulse train and the phase-equalized speech signals (thin line) for the lag. (B) Autocorrelation function for the original speech signals (thick line) and cross-correlation function between the impulse train and the original speech signals (thin line) for the lag.*

as the initial impulse amplitude and then determine the LPC coefficients and the amplitude iteratively. If $G(t) = 0$ for all t, e.g. the unvoiced speech, then Eq. (8) is equivalent to the conventional LPC: the autocorrelation method. By substituting Eqs. (4) and (5) into Eq. (9) under the assumption that the autocorrelation function of e is a delta function for the time delay up to $M+1$, R is an autocorrelation function of the windowed original speech signals s: $R(q) = \sum_{t=0}^{L-1} s(t)s(t+q)$. This supports the result of Fig. 2 by the Wiener-Khinchin theorem.

The differences between PEAR and conventional LPC are the cross-correlation terms of the Eq. (8): $\sum_{i=0}^{I} x(t_i - q)G(t_i)$ ($q = 1, \cdots, p$). We compare the function with the autocorrelation one at the pitch mark: $\sum_{i=0}^{I} x(t_i - q)x(t_i)$ ($q = 1, \cdots, p$). Figure 3(A) shows the cross-correlation function between the impulse train G and the phase-equalized speech signals x, and the autocorrelation function of the phase-equalized speech signals x at the pitch mark. The cross-correlation function is similar to the autocorrelation function. This indicates that PEAR can efficiently reduce the effect of impulse at the pitch mark (i.e. the effect of the harmonics in a vocal-tract spectrum) from the phase-equalized speech signals. Moreover, notice that, though PEAR worked well for the phase-equalized speech, it did not work well for the original speech input because LPC residual signals of the original speech are not considered to be the impulse train. Figure 3(B) shows the cross-correlation function between the impulse train G and the original speech signals s, and the autocorrelation function of the original speech signals s at the pitch mark. The cross-correlation function is different from the autocorrelation function of the original speech signals.

3.2. Algorithms

Figure 4 shows the algorithms of the proposed method. The LPC residual signals are calculated by using the conventional LPC inverse filtering technique. Then, for voiced speech, the phase-equalized speech signals and the pitch mark are obtained by using the LPC residual signals. LPC inverse filtering from the phase-equalized speech signals gives the initial impulse amplitude.

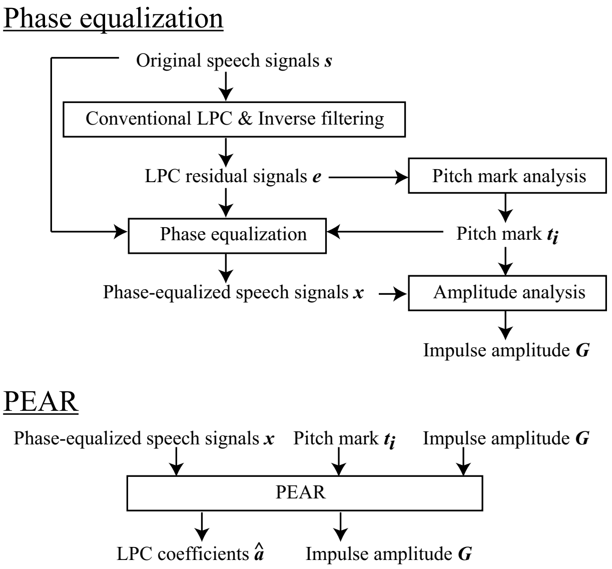


Figure 4: *Algorithms for phase equalization and PEAR.*

For the pitch mark and the initial impulse amplitude, we determine the LPC (PEAR) coefficients. And then the LPC coefficients and the impulse amplitude are determined iteratively.

3.3. TANDEM window

Even when PEAR is applied to estimate the 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 [11]. 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 the PEAR as follows: We use the average of two autocorrelation function and the average of two terms of $x \times G$ in Eq. (8) for the temporally separated windows. Theoretically, the TANDEM window works well when the pitch period is constant in the frame, but the assumption is not satisfied for natural speech. Therefore, to cope with a variable pitch period $T_0, \cdots, T_I$ in the frame, we propose the following shift value:

$$\sum_{i=0}^{I} w_i/T_i \Big/ 2\sum_{i=0}^{I} w_i/T_i^2, \qquad (10)$$

where w is a Gaussian weight function. The analysis window is a Blackman window with a 3.5-fold pitch period [11].

4. Experiments

We evaluated the proposed method using natural speech. The speech signals were recorded at a sampling rate of 16 kHz. Twenty LPC coefficients were obtained with a 4-ms frame shift. No pre-emphasis was used.

4.1. Effect of PEAR

Figure 5 shows the vocal-tract spectrum of the English vowel /i/ for the conventional LPC and PEAR (without a TANDEM win-

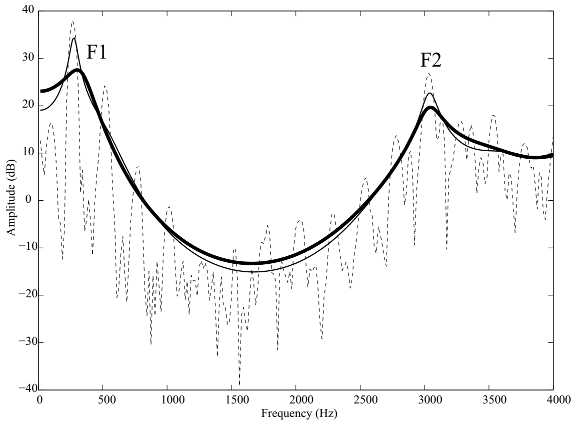
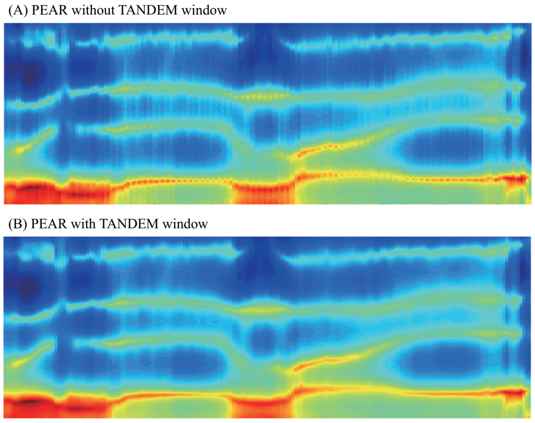


Figure 5: *Vocal-tract spectrum of the English vowel /i/ using conventional LPC (thin line) and PEAR (thick line). Speech spectrum (dashed line).*


(A) PEAR without TANDEM window

(B) PEAR with TANDEM window

Figure 6: *Example of the LPC spectrum sequence of PEAR (A) without and (B) with TANDEM window.*

dow) with a 30-ms Hamming window. A lag window (100 Hz) for the conventional LPC was used. The number of iterations for PEAR was five. The average fundamental frequency was 258 Hz. We can see the vocal-tract spectrum of the conventional LPC was largely affected by the harmonics for the first formant frequency (F1) compared with that of PEAR.

We compared the re-synthesized speech signals of PEAR with those of the conventional LPC. For this purpose, the impulse amplitude was determined to minimize frequency-weighted mean square error between the phase-equalized speech signals and the synthesized speech signals [7]. The signal-to-noise ratios (S/Ns) were 9.26 dB for PEAR and 8.55 dB for the conventional LPC. Moreover, an informal listening test showed the speech re-synthesized using PEAR is superior to that re-synthesized using the conventional LPC.

4.2. Effect of TANDEM window

Figure 6 shows the LPC spectrum sequence of PEAR with and without a TANDEM window of the Japanese word /udemae/. With the TANDEM window, the temporal continuity of the vocal-tract spectrum was slightly improved, and thus the quality of the re-synthesized speech and S/N (9.72 dB) was also improved.

5. Discussion

As we noted in Sec. 3.1, PEAR requires the iterative estimation of the LPC coefficients and the impulse amplitude. However, we found that the initial impulse amplitude gives a good result for estimating the LPC coefficients, so the proposed algorithms may have less computational complexity than other algorithms.

A real-time speech conversion technique is important for investigating human speech production mechanisms. For example, F1 perturbation studies of vowels using the technique have revealed that auditory feedback essentially plays a role in human speech production [11,12]. However, technique utilized by the conventional LPC is sometimes problematic: F1 perturbation for female speaker makes noise in vowel signals due to incorrect vocal-tract estimation. Thus, the proposed method would be useful for solving the problem.

6. Conclusions

We presented a novel vocal-tract spectrum estimation method and showed that it is superior to conventional LPC in terms of robust estimation of vocal-tract spectrum and the quality of re-synthesized speech.

7. Acknowledgements

The authors thank Drs. H. Gomi, H. Kameoka, and Y. Kamamoto for many useful and helpful discussions.

8. 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] Sasou, A. and Tanaka, K., “Glottal excitation modeling using HMM with application to robust analysis of speech,” Proc. ICSLP, 704–707, 2000.

[5] Moriya, T. and Honda, M., “Speech coder using phase equalization and vector quantization,” Proc. ICASSP, 1701–1704, 1986.

[6] Funaki, K., Miyanaga, Y., and Tochinai, K., “Recursive ARMAX speech analysis based on a glottal source model with phase compensation,” Signal Processing, 279–295, 1999.

[7] Honda, M., “Speech coding using waveform matching based on LPC residual phase equalization,” Proc. ICASSP, 213–216, 1990.

[8] Stachurski, J. and McCree, A., “A 4kb/s hybrid MELP/CELP coder with alignment phase encoding and zero-phase equalization,” Proc. ICASSP, 1379–1382, 2000.

[9] Hiwasaki, Y., Mano, K., and Kaneko, T., “An LPC vocorder based on phase-equalized pitch waveform,” Speech Communication, 277–290, 2003.

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

[11] 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,” Proc. ICASSP., 3933–3936, 2008.

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

[13] 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., 2306–2319, 2007.