AUTOMATIC ESTIMATION OF FORMANT AND VOICE SOURCE PARAMETERS USING A SUBSPACE BASED ALGORITHM

Chang-Sheng Yang and Hideki Kasuya
Faculty of Engineering, Utsunomiya University,
2753 Ishii-machi, Utsunomiya 321, Japan
E-mail: yang@utsunomiya-u.ac.jp

ABSTRACT

An automatic method is proposed to estimate jointly formant and voice source parameters from a speech signal. A Rosenberg-Klatt model is used to approximate a voicing source waveform for voiced speech, whereas a white noise signal is assumed for the unvoiced. The vocal tract characteristic is represented by an IIR filter. The formant and anti-formant values are calculated from the IIR filter coefficients which are estimated by using the subspace-based system identification algorithm, while an exhaustive search procedure is applied to obtain the optimal source parameter values, where an error criterion is introduced in the frequency domain. An experiment has been performed to examine performance of the proposed method with natural speech. The results show that the source parameters such as open and closure instants estimated by the method is in good agreement with those defined on the electro-glottograph signals and the formant values estimated are also accurate.

1. INTRODUCTION

High quality parametric analysis techniques are required in almost all speech research areas, such as synthesis, perception, production, voice quality, voice conversion, speech coding, recognition and speaker identification. For formant type speech synthesis, a two-channel analysis method is developed [1], in which an EGG signal is used to estimate the source parameters and an acoustic signal to obtain the vocal tract (VT) parameters. This method requires an extra EGG signal, which makes it unsuitable to process a large amount of data. An ARMA method with a glottal source (GARMA) model is proposed to jointly estimate the source and VT parameters [2]. An AR with exogenous input (ARX) method based on the extended Kalman filter algorithm has shown good performance for natural speech [3]. Both the GARMA and ARX methods, however, are known to be quite sensitive to the model order selected and do not provide with any reasonable algorithm to automatically determine the order.

Recently, much attention has been paid to the state space method in diverse disciplines [4]-[7]. In this method, a signal is processed into a linear model and its parameters are estimated. The desired information is extracted from the estimated model parameters. Little *a priori* parameterization and no non-linear process are needed in the algorithm. The subspace-based approach is numerically robust for high-resolution model-based signal processing. It is especially preferable to the polynomial approach for a short record of noise-corrupted data [4]. These merits are very useful in overcoming the shortcomings of conventional approaches for analysis of speech signals which vary slowly with time.

This paper presents a novel speech analysis method using a direct subspace-based state-space system identification (4SID) algorithm. In Section 2, parameters of the Rosenberg-Klatt (RK) voicing source model and the IIR filter are described. Except for the pitch which is extracted from the glottal closure instants (GCI), the parameters of the RK model are jointly estimated with the VT parameters by an exhaustive search procedure. To obtain the optimal parameter values, an error criterion is defined in the frequency domain by which the effect of phase distortion can be avoided. The direct 4SID algorithm [5][6] is presented in Section 3. The signal subspace order is determined from a singular value matrix that is calculated during identification. Formants and anti-formants are extracted from the transfer function. Post-processing for accurate estimation of the glottal open instant is described in Section 4, which is followed by experiment in Section 5 and results and discussion in Section 6. A two-channel signal, consisting of the speech signal in one channel and the EGG waveform in the other, is used to examine the effectiveness of the proposed method.

2. ANALYSIS MODEL

2.1 Source-filter Model

Speech production process can be represented by a source-filter model [8], in which a speech signal is regarded as the output of a filter (the vocal tract) excited by a sound source. In this paper, the RK model [9] is used to approximate the glottal volume velocity waveform for voiced speech:

$$g(n)=\begin{cases}2an-3bn^2 & (0<n\le T_0\times OQ)\\ 0 & (T_0\times OQ\le n\le T_0\end{cases} \tag{1}$$

$$a=\frac{27AV}{4OQ^2T_0},\quad b=\frac{27AV}{4OQ^3T_0^2}. \tag{2}$$

Parameters of the model are *T*o, *AV* and *OQ*, which correspond to pitch period, amplitude and open quotient of the waveform, respectively. The parameter *TL* is used to control spectral tilting characteristics (in dB down at 3kHz) of the glottal waveform. The vocal tract IIR filter is represented by a transfer function:

$$H(z)=\frac{K(1+\sum_{i=1}^{m}b_iz^{-i})}{1+\sum_{i=1}^{n}a_iz^{-i}}=\frac{B(z)}{A(z)}. \tag{3}$$

Vocal tract parameters, frequencies and bandwidths of the formant and anti-formant, are calculated from the poles and zeros of $H(z)$, respectively.

2.2 Estimation of Model Parameters

Since the GCI corresponds to the negative peak of the voicing source waveform, *T*o is defined as the interval between the two successive negative peaks. In this paper, we use a simplified method to detect pitches from the speech waveform. Using *T*o of the previous pitch *T*(-1), a maximum peak is found within the range of 0.7-1.3*T*(-1). Then a negative peak before that point is picked up to obtain *T*o of the current pitch.

The *AV*, *OQ*, *TL* and vocal tract parameters are all estimated simultaneously by the procedure described below. The duration of glottal open (*GO*, *OQ*=*GO*/*T*o) is varied in steps of one point within the range of 0.35-0.7*T*o to find the optimal source and VT parameters. In this process, *AV* is fixed at 50. For each *OQ*, *TL* is looped from 0 to 5 in steps of 1. One pitch of the voicing source waveform $u(n)$ is then synthesized. From the speech signal $s(n)$, one pitch of the waveform is so extracted that its GCI is aligned with $u(n)$ at its negative peak. Using $u(n)$ and $s(n)$, the transfer function of (3) is estimated using the direct 4SID method described in the next section. Formants of the VT parameters are selected so that unstable poles are dropped out from $H(z)$ (bandwidth too narrow or minus). The speech waveform of one pitch is synthesized and compared to the original $s(n)$ as follows:

$$E = \sum_{n=0}^{N-1} \{ s(n) - \beta \hat{s}(n) \}^2 . \tag{4}$$

AV is estimated from (4) in a least square error sense. Consider the partial differential of (4) with respect to β. By setting it to zero, β is obtained by (5), so that *AV*=50 β.

$$\beta = \frac{\sum_{n=0}^{N-1} s(n)\hat{s}(n)}{\sum_{n=0}^{N-1} \hat{s}^2(n)} . \tag{5}$$

Estimation error is evaluated in the frequency domain using the criterion:

$$J_E = \frac{1}{2\pi} \int_{-\pi}^{\pi} w(\omega) \times \left\{ \log \frac{|s(e^{j\omega})|^2}{|\beta \hat{s}(e^{j\omega})|^2} \right\} d\omega, \tag{6}$$

where $w(\omega)$ is a weighting function which is designed as:

$w(\omega)$ = 1, for 0-3.5kHz,

= (J-i) / J, for 3.5-5kHz, i = 1, 2, …, J,

where J is a constant. This is because errors in the low frequency band are much more significant than in the high frequency band. The parameter values which minimize the criterion (6) are regarded as optimal in the current pitch.

3. DIRECT 4SID ALGORITHM

The formants and anti-formants are calculated from the transfer function $H(z)$ which is identified using a direct 4SID algorithm. A linear time-invariant state-space model is defined as:

$$\begin{aligned} \mathbf{x}(n+1) &= A\mathbf{x}(n) + Bu(n) \\ y(n) &= C\mathbf{x}(n) + Du(n) + w(n), \end{aligned} \tag{7}$$

where $u(n)$ is the input, $y(n)$ the output, $w(n)$ zero mean and limited variance white noise, and $\mathbf{x}(n)$ a p-dimensional state vector. The unknown system matrices A, B, C and D have appropriate dimensions. The transfer function of the system is obtained by

$$H(z) = C(zI - A)^{-1} B + D. \tag{8}$$

The system description (7) can be represented as:

$$Y = \Gamma_i X + \Phi_i U + W , \tag{9}$$

where U, W and Y are Hankel matrices of input, noise and output signals, with

$$\Gamma_i = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{i-1} \end{bmatrix}, \quad \Phi_i = \begin{bmatrix} D & 0 & \cdots & 0 \\ CB & D & \ddots & 0 \\ \vdots & & \ddots & \\ CA^{i-2}B & \cdots & CB & D \end{bmatrix}. \tag{10}$$

The basic algorithm of the direct 4SID is: partition the input-output data U and Y into two orthogonal parts via an RQ factorization; compute a singular value decomposition (SVD) of the projection of Y on the orthogonal space of U; determine the order of the signal subspace; estimate matrices of (10). Matrices A, B, C and D are calculated from (10). The steps of the direct 4SID algorithm is described as follows:

Step 1. Make Hankel matrices from input-output signals:

$$U = \begin{bmatrix} u(1) & u(2) & \cdots & u(N) \\ u(2) & u(3) & \cdots & u(N+1) \\ \vdots & \vdots & \ddots & \vdots \\ u(m) & u(m+1) & \cdots & u(N+m-1) \end{bmatrix}, \tag{11}$$

$$Y = \begin{bmatrix} y(1) & y(2) & \cdots & y(N) \\ y(2) & y(3) & \cdots & y(N+1) \\ \vdots & \vdots & \ddots & \vdots \\ y(m) & y(m+1) & \cdots & y(N+m-1) \end{bmatrix}. \tag{12}$$

Step 2. By an RQ factorization, matrix pair (U,Y) is expressed as

$$\begin{bmatrix} U \\ Y \end{bmatrix} = \begin{bmatrix} R_{11} & 0 \\ R_{21} & R_{22} \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}, \tag{13}$$

where R_{11} and R_{21} are $m \times m$, R_{22} is an $m \times (m-p)$ matrix, Q_1 is an orthogonal complementary space of Q_2.

The projection of Y on the orthogonal space of U is obtained:

$$Y\Pi_{U^T}^{\perp} = (R_{21}Q_1 + R_{22}Q_2)\Pi_{U^T}^{\perp} = R_{22}Q_2 , \tag{14}$$

where $\Pi_{U^T}^{\perp} = I - U^T(\mathrm{UU^T})^{-1}\mathrm{U}$ is the orthogonal projection.

Step 3. The signal subspace is extracted from the SVD of (14).

$$R_{22}Q_2 = USV^T = [U_1 \quad U_2]\begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}\begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}, \qquad (15)$$

where U_1 is $m \times p$, U_2 is an $m \times (m-p)$ matrix, and p is the order of the signal subspace.

For speech signals, it is difficult to determine the order of the subspace directly according to singular values (diagonal elements of S), because noise components have relatively large variances. An evaluation function is calculated with the method similar to [10] :

$$J(i) = \lambda_i (N^{1/N})^i, \quad i = 1,2,\cdots,m-2. \qquad (16)$$

The order p is the value of i when $J(i) - J(i-1)$ is minimum. The matrices of (10) are estimated by

$$\Gamma_i T = U_1, \qquad (17)$$

where T is a nonsingular matrix, and

$$U_2^T R_{21} R_{11}^{-1} = U_2^T \Phi_i. \qquad (18)$$

Equation (18) is obtained by applying the orthogonal property of U and W to (9).

$$U_2^T Y U^+ = U_2^T (U_1 T^{-1} X + \Phi_i U + W) U^+ \qquad (19)$$
$$(U = R_{11}Q_1, \quad U_2 = U_1^{\perp}),$$

where the superscript "+" denotes the Moore-Penrose's generalized inverse.

step.4 The system matrices A and C are calculated as follows:

$$\hat{A} = U_1(1:m-1,1:p)^+ U_1(2:m,1:p), \qquad (20)$$

$$\hat{C} = U_1(1:1,1:p). \qquad (21)$$

The matrices B and D are obtained by solving the equation :

$$X_y = X_u \hat{\theta}, \qquad (22)$$

where

$$\hat{\theta} = \begin{bmatrix} D \\ B \end{bmatrix}, \qquad (23)$$

$$X_y = \begin{bmatrix} \mathrm{E}(1:m-p,1:1) \\ \mathrm{E}(1:m-p,2:2) \\ \vdots \\ \mathrm{E}(1:m-p,m:m) \end{bmatrix}, \qquad (24)$$

$$\mathrm{E} = (U_2)^T R_{21} R_{11}^{-1}, \qquad (25)$$

and

$$X_u = \begin{bmatrix} U_2(1,:)^{\mathrm{T}} & \cdots & U_2(m,:)^{\mathrm{T}} \\ U_2(2,:)^{\mathrm{T}} & 0 & 0 \\ \vdots & 0 & \vdots \\ U_2(m,:)^{\mathrm{T}} & \cdots & 0 \end{bmatrix} \cdot \begin{bmatrix} I & 0 \\ 0 & U_1(1:m-1,1:p) \end{bmatrix}. \qquad (26)$$

To make it easy to understand, the equations are represented by notations used in the software MATLAB.

4. POST-PROCESSING

In order to further improve accuracy of OQ estimated by the procedure described in Section 2, a post-processing algorithm is introduced. The algorithm is as follows. The optimal vocal tract parameters are first used to calculate a zero input response (ZIR). Initial values for the ZIR are selected from the GCI of the previous pitch. Then ZIR is subtracted from the waveform of the current pitch, yielding the deviation from the ZIR (DZIR). The maximum peak is located on the DZIR and the GOI is defined as the zero-crossing point just before that peak.

5. EXPERIMENT

An experiment was performed on a continuous speech in Japanese /*aoiue*/ ("blue top" in English) phonated by a male adult in a soundproof room. The sampling frequency was 10 kHz. An EGG signal was recorded simultaneously with the speech signal. Differentiated EGG (DEGG) was used to evaluate estimation accuracy of pitch and *OQ*. Before the construction of Hankel matrices U and Y, $u(n)$ and $y(n)$ were pre-emphasized by

$$P(z) = 1 - 0.98z^{-1}. \qquad (27)$$

The number m of rows for U and Y, was 20. The data length N was equal to one pitch.

6. RESULTS AND DISCUSSION

The true pitch was defined as the interval between two GCI's extracted from the DEGG signal. The pitch intervals estimated from the speech signal are shown in Figure 1 together with those from DEGG. It can be seen that the both results are in good agreement. We therefore conclude that this method is very suitable in detecting cycle-to-cycle pitches directly from the speech signal.

The GO values estimated by this method are illustrated in Figure 2 together with those extracted from the DEGG signal. It is obvious that the method works considerably well. We found that the post-processing described in Section 4 was significantly important in detecting the GOI. Our sophisticated method provides a technique to detect the GOI and GCI directly from the speech signal without relying on the EGG signal. The re-synthesized speech waveform showed that the amplitude parameter *AV* estimated by (5) was also adequate.

Figure 3 shows trajectories of the formant frequencies obtained by the method. Even the higher formant frequencies are estimated well. Comparing with the results obtained by the ARX model [3], the proposed method has shown to be superior in accuracy of estimated formant bandwidths and higher formant frequencies [11]. The spectral error criterion (6) played an important role in selecting the best parameter values. Perceived quality of re-synthesized speech with the estimated parameter values was remarkably close to the natural speech.

7. SUMMARY

A novel speech analysis technique using a direct 4SID algorithm was described. The voicing source and vocal tract parameters were estimated jointly. Error minimization was performed in the frequency domain. An idea was presented to precisely detect the glottal open and closure instants directly from speech signals without an extra channel of EGG. The system order could be selected in the estimation process. Experimental results showed that not only the vocal tract parameters including the higher formant frequencies but also the source parameters were estimated quite well. This ability can be expected to provide an attractive tool for speech analysis.

Handling of zeros (anti-formants) in the estimation procedure is an on-going discussion for analysis of nasalized vowels and consonants.

Acknowledgement

This work was partly supported by Grant-in-Aid for Scientific Research from the Ministry of education, Science and Culture of Japan, No. 08650420.

8. REFERENCES

[1] Pinto, N. B., Childers, D. G. and Lalwani, A. L. "Formant speech synthesis: improving production quality". *IEEE Trans. Acoustics, Speech, and Signal Processing*, 37(12): 1870-1887, 1989.

[2] Fujisaki, H and Ljungqvist, M. "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform". *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Texas, April 1987.

[3] Ding, W., Kasuya, H. "A novel approach to the estimation of voice source and vocal tract parameters from speech signals". *International Conference on Spoken Language Processing*. Philadelphia, Oct. 1996.

[4] Rao, B. D. and Arun K. S. "Model based processing of signals: a state space approach". *Proceeding of the IEEE*, 80(2):283-309, 1992.

[5] Verhaegen, M. and Dewilde, P. "Subspace model identification, Part 1. The output-error state-space model identification class of algorithms". *International Journal of Control*, 56(5):1187-1210, 1992.

[6] Viberg, M. "Subspace-based methods for the identification of linear time-invariant systems". *Automatica*, 31(12):1835-1851, 1995.

[7] Overschee, P.V. and Moor, B. D. "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems". *Automatica*, 30(1):75-93, 1994.

[8] Fant, G. *Acoustic theory of speech production*, Mouton, The Hague, The Netherlands, 1960.

[9] Klatt, D. and Klatt, L. "Analysis, synthesis and perception of voice quality variations among female and male talkers". *Journal of the Acoustical Society of America*, 87(2):820-857, 1990.

[10] Liang, G., Wilkers, D.M. and Cadzow, J.A. "ARMA model order estimation based on the eigenvalues of the covariance matrix". *IEEE Trans. Signal Processing*, 41(10):3003-3009,1993.

[11] Yang, C.-S. and Kasuya, H. "Estimation of source and vocal tract parameters by using subspace method". *Proceeding of Autumn meeting of the Acoustic Society of Japan*, 3-213, 1997.

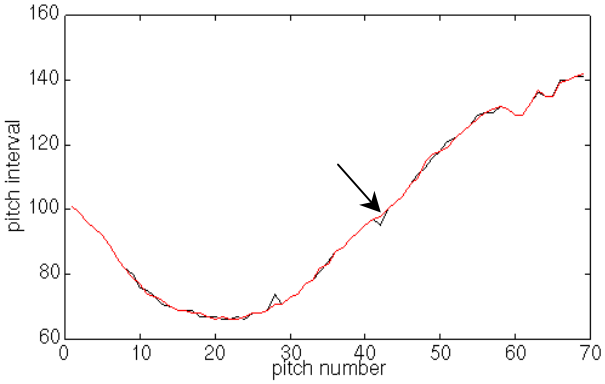

Figure 1. Pitch intervals estimated from glottal close instants. The arrow indicates those of DEGG.

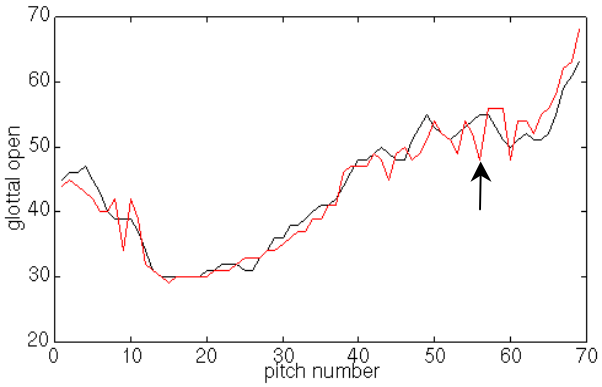

Figure 2. Values of the glottal open phase. The estimated values are smoothed by a three point median. The arrow indicates those of DEGG.

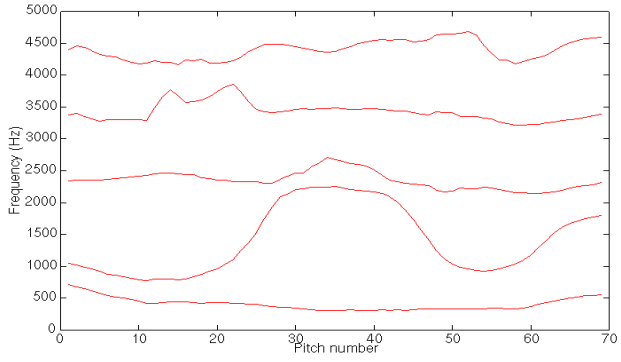

Figure 3. Trajectories of the formant frequencies estimated from the utterance of Japanese vowels /aoiue/.