ESTIMATION OF VOICE SOURCE AND VOCAL TRACT PARAMETERS USING COMBINED SUBSPACE-BASED AND AMPLITUDE SPECTRUM-BASED ALGORITHM

Chang-Sheng YANG and Hideki KASUYA
Faculty of Engineering, Utsunomiya University
7-1-2 Yoto, Utsunomiya 321-8585, Japan
E-mail: yang@utsunomiya-u.ac.jp

ABSTRACT

In this paper, a high quality pole-zero speech analysis technique is proposed. The speech production process is represented by a source-filter model. A Rosenberg-Klatt model is used to approximate a voicing source waveform for voiced speech, whereas a white noise is assumed for unvoiced. The vocal tract transfer function is represented by a pole-zero filter. For voiced speech, parameters of the source model are jointly estimated with those of the vocal tract filter. A combined algorithm is developed to estimate the vocal tract parameters, i.e., formants and anti-formants which are calculated from the poles and zeros of the filter. By the algorithm, poles are estimated based on a subspace algorithm, while zeros are estimated from the amplitude spectrum. For unvoiced speech, an AR model is assumed, which can be solved by LPC analysis. An experiment using synthesized nasal sounds shows that the poles and zeros are estimated quite accurately.

1. INTRODUCTION

High quality pole-zero speech analysis techniques are required in almost all speech research areas. Conventional approaches [1]-[3], however, are sensitive to the analysis model order, particularly when zeros are included in the model. We have proposed a novel analysis method to jointly estimate voice source and vocal tract (VT) parameters of vowels, which is based on a Direct Subspace State Space System IDentification (D4SID) algorithm [4]-[7]. The model order can be determined within the algorithm on the basis of an SVD (singular value decomposition). Experimental results showed that even higher formants of the vowels were estimated quite accurately [8].

In this paper, a combined method is developed for high quality estimation of formant (pole) and anti-formant (zero) parameters, by which poles are estimated by the D4SID algorithm, while zeros are estimated from the amplitude spectrum of the speech signal to overcome the difficulty that the subspace-based algorithm produces too many zeros and the problem of accuracy. In section 2, a source-filter model and analysis algorithm are described. A Rosenberg-Klatt (RK) model is used to approximate a voicing source waveform for voiced speech, whereas a white noise signal is assumed for unvoiced. The VT transfer function is represented by an IIR filter. The parameters of the RK model are jointly estimated with the VT parameters by an iterative procedure. A sophisticated method is proposed to estimate gain and spectral tilt of the RK model within the procedure. An error criterion defined in the frequency domain is introduced to evaluate the estimated parameter values. In section 3, an experiment using synthesized nasal sounds is described. Results of the experiment and estimation precision of the proposed method are discussed.

2. METHOD

2.1. Source-Filter Model

The speech production process is represented by a source-filter model [9], in which a speech signal is regarded as the output of a filter (the vocal tract) excited by a sound source. In this paper, a Rosenberg-Klatt (RK) model [10] is used to approximate the glottal volume velocity waveform for voiced speech, whereas a white noise is assumed for unvoiced. Signal of the RK model is given by:

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

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

Parameters of the model are T_0, *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 VT filter is represented by a transfer function:

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

The VT parameters, i.e., frequencies and bandwidths of the formant and anti-formant, are calculated from the poles and zeros of *H*(*z*), respectively.

2.2. Analysis of Voiced Speech

A. Estimation of Voicing Source Parameters

Since the glottal closure instant (GCI) corresponds approximately to the negative peak of the voicing source waveform, the pitch interval T_0 is defined as the interval between the two successive negative peaks of the speech waveform. The current pitch T_0 is detected by finding the negative peak within the range of 0.7-1.3*T*(-1). *T*(-1) is the previous pitch interval. The *AV*, *OQ*, *TL* and VT parameters are all estimated simultaneously by the procedure described below.

The duration of glottal open phase (*GO, OQ=GO/T*$_0$) is varied in steps of one point within the range of 0.35-0.7T_0 to find the optimal source and VT parameters. For each *GO*, *AV* and *TL* are fixed at first, i.e., *AV* = 50 and *TL* = 0. This will not affect the estimated result of poles and zeros. Using the values of T_0, *AV*, *OQ* and *TL*, the voicing source waveform *u*(*n*) is synthesized. The waveform of speech signal *s(n)* 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 eq. (3) is estimated using the algorithm described in the next subsection. The parameter *TL* is estimated by comparing the spectrum tilt of the original signal with that of the synthesized signal (see Figure 1). *AV* is estimated in a least square error sense. One pitch of estimated signal *s'*(*n*) (*AV* = fixed) is synthesized and compared with the original *s*(*n*) as follows:

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

β is obtained by

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

so that $AV = 50\,\beta$.

An estimation error of the synthesized speech is evaluated in the frequency domain using the criterion:

$$J_E = \frac{1}{\pi} \int_0^{\pi} \left\{ w(\omega) \times \log \frac{|s(e^{j\omega})|^2}{|\beta s'(e^{j\omega})|^2} \right\}^2 d\omega, \qquad (6)$$

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

$w(\omega) = 1$, $\qquad \omega = 0 \sim \pi/2$,

$w(\omega) = (\pi - \omega) / (\pi/2)$, $\qquad \omega > \pi/2$.

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.

B. Estimation of VT Parameters

The linear time-invariant system with a transfer function given in eq.(3) can be alternatively described by the state space equation:

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

where *u*(*n*) is the input, *y*(*n*) the output, *w*(*n*) white noise of zero mean and limited variance, and **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:

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

B.1 Estimation of Poles

The system description (7) can be represented as:

$$Y = \Gamma_i X + \Phi_i U + W, \qquad (9)$$

where *U, W* and *Y* are Hankel matrices of input, noise and output signals, *X* is the state trajectory matrix, and

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

The basic idea of the D4SID based method is: partition the input-output data *U* and *Y* into two orthogonal parts via an RQ factorization; remove the part related to the input *U* from *Y* by the orthogonal space of *U*; perform an SVD on the left part of *Y* to separate signal subspace from noise; extract the signal subspace to estimate the system matrices. An efficient procedure to estimate the system matrix *A* by the D4SID algorithm is described as follows.

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

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

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

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

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

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

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

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

Where the left of (14) is the orthogonal projection

$$\Pi_{U^T}^{\perp} = I - U^T (\mathrm{UU}^{\mathrm{T}})^{-1} \mathrm{U}. \qquad (15)$$

This operation in effect removes the part of the output *Y* that is related to the input *U*. Comparing eqs. (14) with (9), it becomes

$$Y\Pi_{U^T}^{\perp} = \Gamma_i X \Pi_{U^T}^{\perp} + W = R_{22}Q_2 \ . \qquad (16)$$

Step 3. The signal subspace is extracted from the SVD of eq. (16)

$$R_{22}Q_2 = USV^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}, \qquad (17)$$

where U_1 is $m \times p$, U_2 is an $m \times (m\text{-}p)$ matrix, and p is the order of the signal subspace which is determined by finding the significant singular values.

For speech signals, it is difficult to determine the order directly from the singular values (diagonal elements of S), because noise components have relatively large variances. An evaluation function is calculated with the method similar to [11]:

$$J(i) = \lambda_i (N^{1/N})^i, \quad i = 1,2,\cdots,m-1. \tag{18}$$

The order p is the value of $J(i)$ when $J(i)$ - $J(i\text{-}1)$ is minimum. Since the left singular vectors of R_{22} coincide with those of eq. (17), the SVD is performed as

$$R_{22} = U_1 S_1 V_1^T Q_2^T + U_2 S_2 V_2^T Q_2^T. \tag{19}$$

An estimate of Γ_i of eq. (10) is

$$\Gamma_i = U_1. \tag{20}$$

Note that Γ_i satisfies

$$\Gamma_{i1} A = \Gamma_{i2}, \tag{21}$$

where

$$\Gamma_{i1} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{i-2} \end{bmatrix}, \quad \Gamma_{i2} = \begin{bmatrix} CA \\ CA^2 \\ \vdots \\ CA^{i-1} \end{bmatrix}. \tag{22}$$

Then the state transition matrix A is calculated by:

$$A = \Gamma_{i1}^{+} \Gamma_{i2}, \tag{23}$$

where the superscript "+" denotes the Moore-Penrose's generalized inverse. Obviously, poles of eq. (8) are the eigenvalues of A.

B.2 Estimation of Zeros

The subspace-based algorithm, in definition, produces the same number of zeros as poles as shown in the equation below.

$$H(z) = K \frac{(1 - z_1 z^{-1})(1 - z_2 z^{-1}) \cdots (1 - z_n z^{-1})}{(1 - p_1 z^{-1})(1 - p_2 z^{-1}) \cdots (1 - p_n z^{-1})} \tag{24}$$

Therefore, for speech signals, physically spurious zeros may be included in the estimated zeros. It is often difficult to select appropriate zeros. Moreover, zeros estimated by the subspace algorithm are not always accurate. To overcome these difficulties, zero candidates are calculated from the amplitude spectrum. Then appropriate zeros are selected by an analysis-by-synthesis procedure according to the error criterion shown in eq. (6). The case of no zero is also considered in this procedure.

2.3. Analysis of Unvoiced Speech

For unvoiced speech, an AR model which is excited by a white noise is assumed. The transfer function (3) becomes

$$H(z) = \frac{K}{A(z)}. \tag{25}$$

In this study, the autocorrelation LPC analysis is used with a 15 ms Hamming window and frame shift of 5ms. Gain K is calculated from the residual signal every 5 ms.

$A(z)$ of eq. (25) can also be estimated by a subspace based method [12]. Using the output $s(n)$, a relatively large size covariance matrix is constructed. Then an SVD is performed on it. Γ_i of eq. (10) is estimated according to the p largest singular values. So that poles of $A(z)$ is estimated. The problem is that the covariance matrix of $s(n)$ usually has nearly full rank when the spectrum is flat. It is difficult to truncate an order p signal subspace from that. It is also indicated by Viberg [6] that subspace based methods are sensitive to the excitation of the system. In this case, the traditional prediction error methods are preferable.

3. EXPERIMANT AND DISCUSSION

An experiment was performed on a nasal sound segment including 40 pitches which is synthesized by the voice source parameters obtained from a natural sound with constant VT parameters of /m/. Zeros were estimated from the amplitude spectrum using a 10th order FFT. The sampling frequency was 10 kHz. The number m of rows for U and Y, was 25. The data length N was equal to one pitch.

Figures 2 and 3 illustrate the waveform and spectrum of the result, respectively. We can see that not only the spectrum but waveform of the estimated signal is very close to the original one. The true values and the estimated values (averaged over 40 periods) of the formants and anti-formants are shown in Table 1. The results show that errors are all less than 15 Hz for the formant frequencies (F1-F5), and less than 19 Hz for the bandwidths. For the anti-formants, errors are less than 15Hz for the frequencies (AF1-AF3), and less than 32 Hz for the bandwidths. Standard deviations of the results are all less than 9.5 Hz for the frequencies. This implies that the proposed method is very stable.

	FREQUENCY			×BANDWIDTH		
	True	estd.	(σ)	True	estd.	(σ)
F1	230	231	(2.9)	100	101	(3.7)
F2	860	856	(3.9)	200	207	(15.6)
F3	1300	1315	(7.9)	250	269	(13.6)
F4	2320	2324	(9.5)	50	47	(27.7)
F5	3320	3319	(5.7)	300	295	(19.1)
AF1	500	515	(2.4)	100	112	(5.0)
AF2	1100	1110	(5.0)	200	232	(22.2)
AF3	3000	2998	(2.7)	100	98	(7.0)

Table 1: Formants and anti-formants of /m/ and the estimated values which are averaged over 40 periods (in Hz). 'estd' means the estimated values. ' σ ' means the standard deviations.

4. SUMMARY

A novel pole-zero speech analysis technique to jointly estimate the voicing source and vocal tract parameters from speech signal is described. Using the property that the D4SID method is not sensitive to values of *TL* and *AV*, parameters of the RK model are estimated by an iterative procedure. A combined algorithm is proposed to estimate poles and zeros of the vocal tract filter. The D4SID method provides high accuracy for pole estimation, while the amplitude spectrum based method is reasonable for zeros. Experimental result of synthesized nasal sounds showed that the voicing source and vocal tract parameters were estimated quite accurately and stable. This performance indicates that the proposed method is expected to provide a high quality analysis technique for speech synthesis, voice conversion, speech coding and speech and speaker recognition.

5. REFERENCES

1. Morikawa, H. and Fujisaki, H "Adaptive analysis of speech based on a pole-zero representation," *IEEE Trans. ASSP* 30: 77-88, 1982.

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*, 637-640, Texas, April 1987.

3. Ding, W. and 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. Yang, C.-S. and Kasuya, H. "Automatic estimation of formant and voice source parameters using subspace based algorithm," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 914-917, Seattle, May 1998.

9. Fant, G. *Acoustic theory of speech production*, Mouton, The Hague, The Netherlands, 1960.

10. 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.

11. 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.

12. Arun, K. S. "Principal components algorithm for ARMA spectrum estimation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, 37(4): 566-571, 1989.

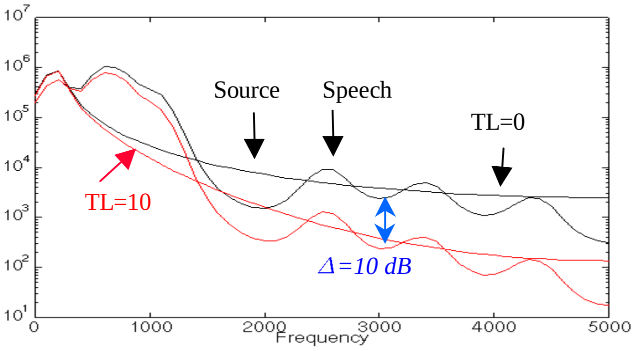

Figure 1: *TL* is estimated by comparing the spectrum of the original signal at 3 kHz with that of the synthesized by *TL*=0.

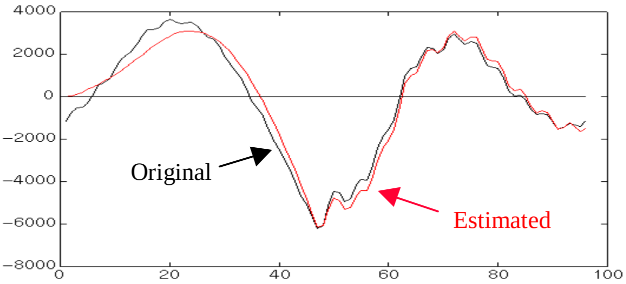

Figure 2: The original and estimated waveform of /m/.

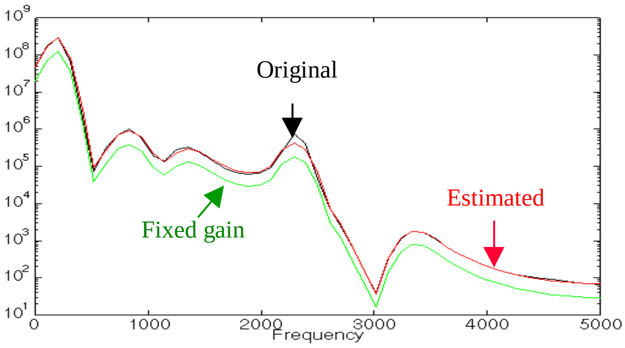

Figure 3: The original and estimated spectrum of /m/.