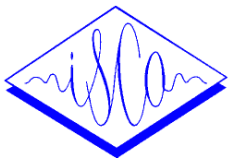


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Voice Parameter Estimation Using Sequential SVD and Wave Shaping Filter Bank

SungHoon HONG*, SangKi KANG* and SouGuil ANN*

*Department of Electronics Engineering
Seoul National University
San 56-1, Shillim-dong, Kwanak-gu
Seoul 151-742, Korea

ABSTRACT

In this paper, we propose new estimation technique of voice source and vocal-tract parameters in voiced speech. First, Kalman filtering technique which is a new approach to time-varying analysis is examined. Second, in order to overcome the problems of Kalman filtering approach, the error covariance matrix in Kalman filter is decomposed by reduced-rank SVD(singular value decomposition). Third, in order to estimate more accurate vocal-tract parameters we introduce robustness concept which assumes residual characteristics as normal mixture density. And, finally we propose wave-shaping filter bank approach as a new voice source modeling method which can represents and controls various source characteristics in frequency domain. The proposed methods give more accurate and more various time-varying voice parameters. These methods can be used to improve the speech quality in speech synthesis and coding techniques.

I. INTRODUCTION

Most of recent speech analysis and synthesis techniques are based source-filter concept which considers speech as the output of vocal tract excited by voice source as Fig. 1. Therefore, it is important to model the characteristics of the source and vocal tract, and to estimate accurate model parameters. The linear predictive coding(LPC) technique has been widely used in the area of speech analysis and synthesis [1]. Although the linear prediction theory combined with the all-pole filter is quite excellent, there exists room for further consideration: 1) it can not effectively estimate the characteristics of time-varying voice. 2) it can not effectively model the charateristics of source. 3) it is difficult to estimate source parameters. Recently, there has been reported a new approach to time-varying analysis based on Kalman filtering [2]. But, it can lead to numerical results that are inaccurate or even meaningless, be unstable, and can not decouple noise effects [3].

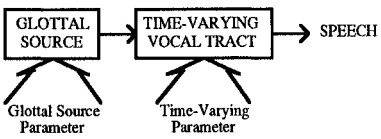


Fig. 1 Speech model using source-filter concept

In existing voice source model, voice source waveforms are represented by sum-of-exponentials or any polynomials in time domain [4]. These simplified models can not represent a variety of source waveforms and complex characteristics in frequency domain. In order to overcome these problems, sum-of-basis-functions model was proposed by Thomson recently [5]. This model represents voice source waveforms as a weighted sum of basis functions. But, Thomson model has a several serious problems: 1) real voice source waveforms can not be effectively modeled by estimated coefficients. 2) reduced-order model to overcome this problem becomes to very simplified model. 3) estimation technique of model parameters is not complete and gives inaccurate solutions.

In this paper, we propose the sequential SVD(singular value decomposition) algorithm in order to overcome the problems of the exsting estimation techniques. This algorithm split error covariance matrix of the Kalman algorithm into signal subspace and noise subspace using reduced-rank SVD [6], and then update vocal tract parameters by applying robustness concept which assumes residual characteristics as normal mixture density. This proposed algorithm gives more accurate time-varying voice parameters.

And, we propose wave shaping filter bank as a new voice source modeling method. The proposed modeling method can represent a variety of voice source waveshapes. Finally, the weights and time-shift parameters of the proposed voice source model are estimeted using ML(maximum likelihood) method [7].

The description of the sequential SVD algorithm in voice source waveform and vocal-tract parameter estimation is presented in Section II. The wave-shaping filter bank as a new voice source modeling method is described in Section III. In Section IV, simulation results are given that compare the existing methods. Summary and conclusions are discussed in Section V.

II. VOICE PARAMETER ESTIMATION USING SEQUENTIAL SVD

In order to overcome the problems of block processing techniques, sequential processing techniques including RLS(recursive least square) and Kalman filtering have been widely used in speech analysis recently [2], [3]. But, the existing sequential processing techniques have difficulties in estimating accurate vocal tract parameters in noise environment and have poor results in stability [3]. In this section, sequential algorithm using SVD and robustness concept is proposed to overcome the above problems.

Let us assume that the received signal, x_k, is an p-dimensional vector given by (1). The Kalman coefficients vector, a_k is an p-dimensional vector denoted by (2). The Kalman gain vector, K_k, denoted by (3). The estimation error, e_k can then be expressed as (4).

$$x_k = [x_{1,k}, x_{2,k}, \ldots\ldots, x_{p,k}]^T \qquad (1)$$

$$a_k = [a_{1,k}, a_{2,k}, \ldots\ldots, a_{p,k}]^T \quad (2)$$

$$K_k = [K_{1,k}, K_{2,k}, \ldots\ldots, K_{p,k}]^T \quad (3)$$

$$\varepsilon_k = s_k - x_k^T a_k \quad (4)$$

In the above equations, k is time-index and s_k is the k th signal sample. Then the Kalman filter equations become (5), (6) and (7). P_k means the error covariance matrix in Kalman filter.

$$\hat{a}_k = \hat{a}_{k-1} + K_k[s_k - x_k \hat{a}_{k-1}] \quad (5)$$

$$K_k = P_{k-1} x_k^T \alpha^{-1} \quad (6)$$

$$P_k = P_{k-1} - K_k x_k^T P_{k-1} \quad (7)$$

where $\alpha = [x_k^T P_{k-1} x_k^T + R_{k\varepsilon}]$

But, this Kalman filter equations can lead to numerical results that are inaccurate or even meaningless, be unstable, and can not effectively decouple noise effects [3]. In order to overcome these problems, we reconstruct the error covariance matrix using reduced-rank SVD technique [6].

The previous error covariance, P_{k-1}, can be represented by multiply of orthogonal matrices, U, V, and diagonal matrix using QR decomposition and Givens rotations [6].

$$P_{k-1} = U_{k-1} \Sigma_{k-1} V_{k-1} \quad (8)$$

This SVD of p-order error covariance matrix, P_{k-1}, can be splitted into r-order dominant subspace and p-r order subdominant subspaces as shown in (9) and Fig. 2.

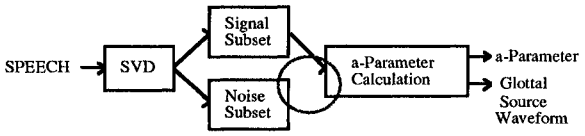


Fig. 2 Speech analysis using reduced-rank SVD

$$P_{k-1} = [U(r) | U(p-r)] \begin{bmatrix} \Sigma(r) & 0 \\ 0 & \Sigma(p-r) \end{bmatrix} \begin{bmatrix} V^T(r) \\ V^T(p-r) \end{bmatrix} \quad (9)$$

And, the error covariance matrix can be reconstructed by only dominant subspaces. This reconstructed error covariance matrix can lead to optimal solution in least square sense. In the equation (9), the reduced-order r can be determined by mean-squared error [6]. Thus, the proposed approach can guarantee positive eigenvalue and improved stability [3], [6].

But, the proposed sequential algorithm using reduced-rank SVD without any decoupling algorithm can not effectively estimate the accurate vocal-tract parameters and voice source waveforms. In general, closed-phase concept which estimates vocal-tract parameters in only glottal closure region is used for decoupling voice source effects. But, this method has a problem that accurate selection of glottal closure region is very difficult and more difficult specially in female or child voices and transition regions. In order to overcome these problems, we introduce robustness concept which assumes residual characteristics as normal mixture density.

The observed speech $s(t)$ is assumed to be of the form

$$s(t) = \hat{s}(t) + u(t) \quad (10)$$

where $u(t)$ is a zero-mean white Gaussian noise with variance . Strictly speaking, $u(t)$ is a non-Gaussian with a mixture distribution for voiced speech, such that a small share of the glottal excitation is from a normal distribution with a much larger variance. Then the driving source can be represented as follows:

$$u(t) = (1 - q_t) u_0(t) + q_t u_1(t) \quad (11)$$

where $u_0(t)$, $u_1(t)$ are independent zero-mean white Gaussian processes with variances σ_0^2 and $\sigma_1^2 (>> \sigma_0^2)$, respectively. $\{q_t\}$ is a sequence of independent random variables with probability $P\{q_t=1\} = \lambda$ and $P\{q_t=0\} = 1-\lambda$ and are independent of $u_0(t)$ and $u_1(t)$. Eq. (3) is a good approximation to voiced speech [2]. In this model, $u_0(t)$, $u_1(t)$ are pre-defined distributions, and q_t can be estimated using ML(maximum likelihood) method [9].

From this assumed model, we can modify equation (5), the coefficient update formula in Kalman filtering approach as following equation.

$$\hat{a}_k = \hat{a}_{k-1} + K_k \psi(\varepsilon_k) \quad (12)$$

$$\varepsilon_k = s_k - x_k \hat{a}_{k-1} \quad (13)$$

In the above equations, q_t is estimated from the vector set of (13), and from this estimated q_t, we can determine the weighting function. In this paper, we intoduce Huber function as a weighting function [10]. This proposed algorithm gives more accurate time-varying voice parameters. Detailed experiment results will be shown in Section IV.

III. VOICE SOURCE MODELING USING WAVE SHAPING FILTER BANK

In existing voice source models, source waveforms are represented by simplified sum-of-exponentials or polynomial types in time domain [4]. These existing simplified models can not represent a variety of voice source characteristics and have more difficulties in representing any characteristics in frequency domain. In order to overcome these problems, new model that represent voice source waveforms as weighted sum of basis functions is proposed recently [5]. But, This model has several problems: 1) real voice source waveforms are not effectively modeled by estimated coefficents. 2) reduced-order model to overcome the problem of 1) becomes to very simplified model. 3) estimation technique of model parameters is not complete. 4) this model can not effectively represent the frequency characteristics of voice source. In order to overcome the above problems, we propose wave-shaping filter bank as a new voice source modeling method.

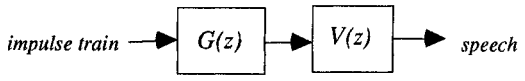


Fig. 3 Speech generation model

In Fig. 3, voice source signals are assumed to be the output of all-zero filter $G(z)$ driven by excitation signals, and excitation signals are assumed to be pulses that are spread in the noise environment. The coefficients of all-zero filter $G(z)$ can be obtained by truncating residual waveforms of vocal-tract filter $V(z)$ using sequential SVD [11]. We called this all-zero filter as wave-shaping filter. But, this wave-shaping filter has a

excessive number of filter coefficients. In order to avoid this problem, we split estimated impulse response of the wave-shaping filter into sub-banded versions using filter bank concept [12].

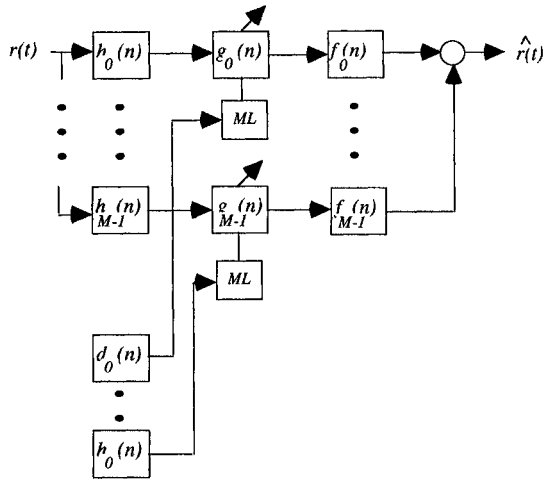


Fig. 4 Detailed description of wave-shaping filter bank

Detailed description description of proposed voice source modeling method is shown in Fig. 4. The impulse response of filter in Fig. 4 $h_k(n)$, k=0,1,...,M-1 represent sub-band filter for splitting estimated residual *r(t)*. The filter impulse responses $g_k(n)$, k=0,1,...,M-1 repres-nent wave-shaping filter bank with 2^k fold decimator (when k=0, 2-fold decimator). The filter impulse response $f_k(n)$, k=0,1,...,M-1 represent 2^k-fold interpolation filter with 2^k fold interpolator (when k=0, 2-fold decimator). The filter imkulse response $d_k(n)$, k=0,1,...,M-1, represent saved patterns of sub-banded voice source characteristics.

Estimated voice source waveforms are splitted into each sub-banded versions. These versions are compared with saved patterns, and appropriate saved patterns are selected and its weight and time-shift parameters are estimated using ML(maximum likelihood) method. The ML approach requires the solution to

$$\min_{\alpha,\tau} \int_T |\hat{r}_k^{(n)}(t) - \alpha d_k(t-\tau)|^2 dt \rightarrow \hat{\alpha}_k^{(n+1)}, \hat{\tau}_k^{(n+1)} \tag{14}$$

where *(n)* is a iteration number. Assuming that the observation interval *T* is long compared to the duration of the signal and the maximum expected delay, the two-parameter maximization required in (15) can be carried out in two steps as follows:

$$\max_{\tau} |g_k^{(n)}(\tau)| \rightarrow \tau_k^{(n+1)} \tag{15}$$

$$\alpha_k^{(n+1)} = \frac{g_k^{(n)}(\hat{\tau}_k^{(n+1)})}{E} \tag{16}$$

$$E = \int_T |d_k(t)|^2 dt \tag{17}$$

$$g_k^{(n)}(\tau) = \int_T \hat{r}_k^{(n)}(t).d_k^*(t-\tau)dt \tag{18}$$

We note that $g_k^{(n)}(\tau)$ can be generated by passing $r_k^{(n)}(t)$ through a filter matched to $d_k(t)$. The algorithm is illustrated in Fig. 5.

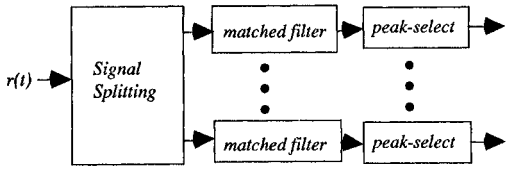


Fig. 5 ML processing for extraction of voice source parameters

Thus, pattern number of d_k, weighting parameter α, and time-shifting parameter τ are parameters of this proposed voice source modeling methods. If compared pattern data set is composed in accurate, we can reconstruct voice source waveforms from these parameters. Detailed experimental results are shown in Section IV.

IV. EXPERIMENT AND DISCUSSION

We have examined data for voiced sounds from two talkers, one male and one female, using proposed analysis procedure. Voice sounds are recorded by a recording system through the microphone and then sent into a computer after being A/D converted with sampling frequency of 10kHz and quantization of 16 bits. Fig. 6 shows the results of the experiments for voice source waveforms estimation in Korean vowel /a/ pronounced by male talker. The proposed method gives better results compared to conventional Kalman filtering algorithm. Fig. 7 shows the results of the proposed voice source model fitting in time-domain compared with existing LF model. Fig. 8 shows the results of model fitting in frequency domain. The proposed method gives better results compared to one of existing LF(Liljencrants Fant) voice source model [4]. And, Fig. 9 shows the results of synthetic speech waveforms. But, performance of the proposed modeling methods is dependant on how to the saved patterns. And, optimum filter bank design which guarantees perfect reconstruction is is becoming important issue. We are proceeding any experiments in order to overcome these problems.

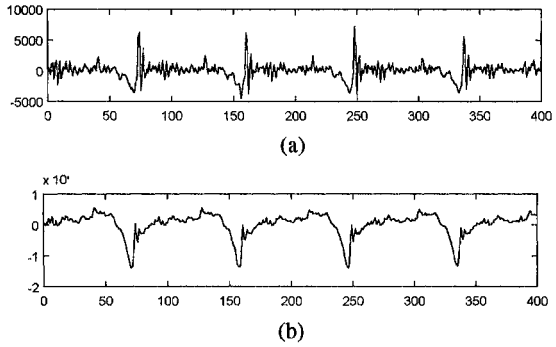
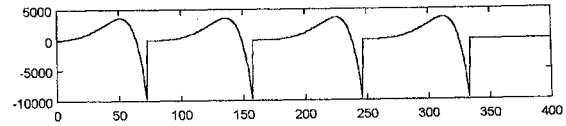
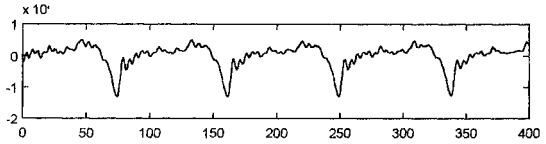


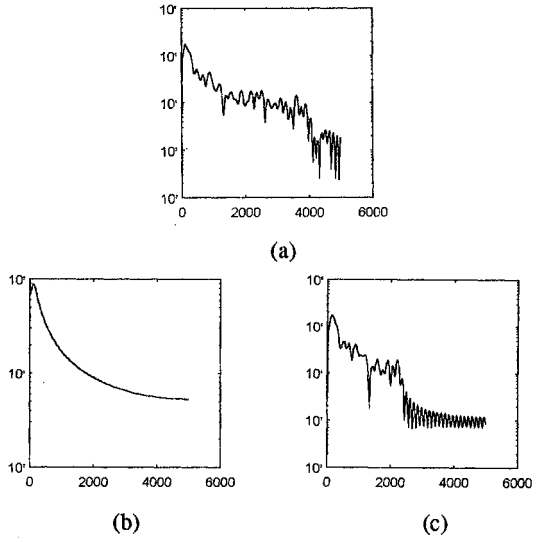
Fig. 6 The results of voice source waveform estimation
(a) Convential Kalman filtering (b) Proposed method





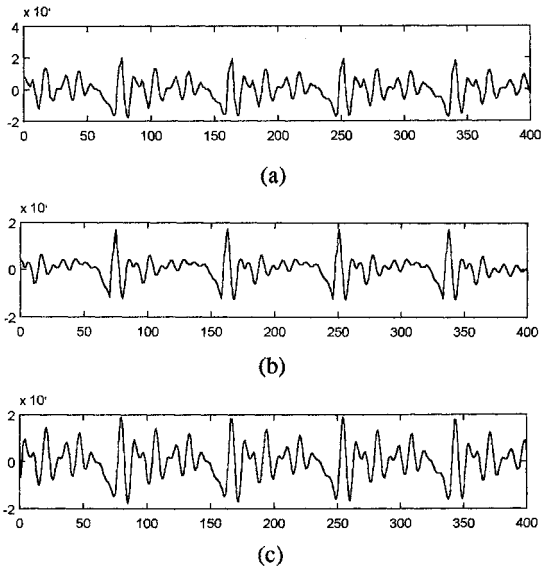
(b)

Fig. 7 The results of voice source model fitting in time-domain
(a) LF model fitting (b) Proposed method



(a) (b) (c)

Fig. 8 The results of source model fitting in frequency-domain
(a) Original waveform (b) LF model (c) Proposed method



(a) (b) (c)

Fig. 9 Synthetic speech
(a) Original speech (b) LF model fitting (c) Proposed method

V. CONCLUSION

In this paper, we propose the sequential SVD(singular value decomposition) algorithm in order to overcome the problem of Kalman filter. This algorithm split error covariance matrix of the Kalman algorithm into signal subspace and noise subspace using reduced-rank SVD, and then update vocal-tract parameters by applying robustness concept which assumes residual characteristics as normal mixture density. This proposed algorithm gives more accurate time-varying vioce parameters.

And, we propose the wave-shaping filter bank as a new voice source modeling method which can represent various source characteristics in order to overcome the problem of existing voice source modeling method. In this model, source signals are assumed to be the output of any filter driven by excitation signals, and excitation signals are assumed to be pulses that are spread in the noisy environment. In order to obtain the source signals modeled here, we apply wave shaping filter estimation, where filter coefficients can be obtained by truncating residual waveforms of vocal-tract filter using sequential SVD. But, this wave-shaping filter has a excessive number of filter coefficients. In order to avoid this problem, we split estimated impulse response of the wave-shaping filter into sub-banded versions using filter bank concept. This wave-shaping filter bank shows that the filtered signal can be seen that the peaks are remarkably impulse-like and the resolution is significantly increased.

We have examined data for voiced sounds from four talkers, three male and one female, using proposed analysis procedure. Experimental results indicate that the proposed method gives more accurate source signals, formants, and bandwidths. Specially, in transiton region or female voices which have difficulties in estimating voice parameters, the proposed algorithms show the better results. And, the proposed wave shaping filter bank as a new voice source modeling method can represent the various voice source characteristics.

REFERENCES

1. J.D.Markel and A.H.Gray, Jr., *Linear Prediction of Speech,* Springer-Velag, 1976
2. S.Crisafulli, J.D.Mills and R.R.Bitmead, "Kalman Filtering Techniques in Speech Coding", *Proceedings of ICASSP*, Vol. 1, pp 77-80, 1992
3. A.A.Giodano and F.M.Hsu, *Least Square Estimation with Application to Digital Signal Processing,* A Wiley-Interscience Publication, 1985
4. H.Fujisaki and M.Ljungqvist, "Proposal and Evaluation of Models for the Glottal Source Waveform", *Proceedings of ICASSP*, 31.2.1, pp 1605-1608, 1986
5. M.M.Thomson, "A New Method for Determining the Vocal Tract Transfer Function and Its Excitation from Voiced Speech", Proceedings of ICASSP, vol. 2, pp 37-40, 1992
6. R.Vaccaro (editor), *SVD and Signal Processing II: Algorithms, Analysis and Applications,* Elsevier Science Publishers B. V., 1991
7. M.Feder and E.Weinstein, "Parameter Estimation of Superimposed Signals Using the EM algorithm", *IEEE Trans. on ASSP*, vol. 36, No. 4, April, 1988
8. Y.M.Cheng and D.O'Shaughnessy, "Automatic and Reliable Estimation of Glottal Closure Instant and Period", *IEEE Trans. on ASSP*, vol. 37, No. 12, December, 1989
9. B.G.Lee, K.Y.Lee, S.Ann, and I.Song, " A Sequential Algorithm for Robust Parameter Estimation and Enhancement of Noisy Speech", Proceedings of ISCAS, vol. 1, pp 243-246, 1993
10. C.J.Masreliez and R.D.Martin, "Robust Bayesian Estimation for the Linear Model and Robustfying the Kalman Filter", IEEE Trans. on Automatic Contril, vol. AC-22, No. 3, June, 1977
11. D.H.Yom and S.Ann, "Spiking Filtering of All-Pass Filter Impulse Responses", Proc. IEEE, vol. 75, Dec., 1987
12. P.P.Vaidyanathan, Multirate Systems and Filter Banks, PTR Prentice-Hall, Inc., 1993