VOICE SOURCE ESTIMATION USING SEQUENTIAL SVD AND EXTRACTION OF COMPOSITE SOURCE PARAMETERS USING EM ALGORITHM

SungHoon HONG, HongSub CHOI, 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, the influence of voice source estimation and modeling on speech synthesis and coding is examined and then their new estimation and modeling techniques are proposed and verified by computer simulations. It is known that the existing speech synthesizer produced the speech which is dull and inanimated. These problems are arised from the fact that existing estimation and modeling techniques can not give more accurate voice parameters. Therefore, in this paper we propose a new voice source estimation algorithm and modeling techniques which can represent a variety of source characteristics. First, we divide speech samples in one pitch region into four parts having different characteristics. Second, the vocal-tract parameters and voice source waveforms are estimated in each regions differently using sequential SVD. Third, we propose composite source model as a new voice source model which is represented by weighted sum of pre-defined basis functions. And finally, the weights and time-shift parameters of the proposed composite source model are estimeted using EM(estimate maximize) algorithm. Experimental results indicate that the proposed estimation and modeling methods can estimate more accurate voice source waveforms and represent various source characteristics.

1. 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].

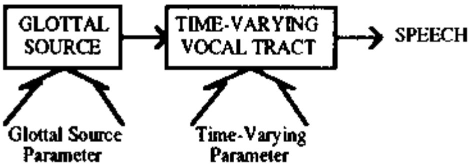

Fig. 1 Speech model using source-filter concept

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.

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 region division concept which differently apply the estimation algorithm into divided sub-regions. This proposed algorithm gives more accurate time-varying voice parameters.

And, we propose the composite source model as a new voice source model. This model represents voice source waveforms as a weighted sum of pre-defined basis functions. The proposed model can represent a variety of voice source waveshapes. Finally, the weights and time-shift parameters of the proposed composite source model are estimeted using EM(estimate maximize) algorithm [7].

2. VOICE SOURCE WAVEFORM 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 region division 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).

$$x_k = \left[x_{1,k}, x_{2,k}, \ldots\ldots, x_{p,k}\right]^T \tag{1}$$

$$a_k = \left[a_{1,k}, a_{2,k}, \ldots\ldots, a_{p,k}\right]^T \tag{2}$$

$$K_k = \left[K_{1,k}, K_{2,k}, \ldots\ldots, K_{p,k}\right]^T \tag{3}$$

$$\varepsilon_k = s_k - x_k^T a_k \tag{4}$$

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.

$$\hat{a}_k = \hat{a}_{k-1} + K_k\{s_k - x_k\hat{a}_{k-1}\} \tag{5}$$

$$K_k = P_{k-1}x_k^T\alpha^{-1} \tag{6}$$

$$P_k = P_{k-1} - K_k x_k^T P_{k-1} \tag{7}$$

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

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].

$$P_{k-1} = U_{k-1}\,\Sigma_{k-1}\,V_{k-1} \tag{8}$$

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.

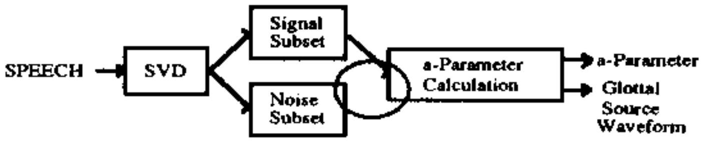

Fig. 2 Speech analysis using reduced-rank SVD

$$P_{k-1} = \left[U(r) \mid U(p-r)\right]\left[\frac{\Sigma(r) \mid 0}{0 \mid \Sigma(p-r)}\right]\left[\frac{V^T(r)}{V^T(p-r)}\right] \qquad (9)$$

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. These problems are arrised from following reasons: 1) Vocal tract has different characteristics in glottal open, close, and transition regions. 2) Interaction effects between voice source and vocal tract give difficulties in estimationg the accurate vocal tract parameters.

In order to overcome these problems, we divide speech samples in one pitch region into four parts (glottal closure region, glottal open region, glottal open-to-close transition region, glottal close-to-open transition region). In order to divide into four regions, we introduce ML(maximum likelihood) method and Hilbert transform technique proposed by Cheng and O'Shaughnessy [8]. Thus, The proposed speech analysis is based on the following assumptions: 1) The vocal tract is significantly excited only during short intervals around glottal closure and opening. 2) The interaction between the glottal source and the vocal tract can be effectively modeled by varying the vocal tract parameters at glottal closure and opening. In practice, existing sequential processing techniques have difficulties in decoupling the characteristics of voice source. This problem is arised from the fact that the existing algorithms did not effectively introduce the characteristics of speech. Divided regions have different characteristics in each: 1) In glottal closure region, there are no source effects and vocal tract parameters are estimated very easily. 2) In glottal open region, the characteristics of vocal tract have some variations in front of this region and become a fixed values. and 3) In glottal transition regions, the variations of vocal tract characteristics are very serious and the estimation of characteristics has difficulties.

These characteristics of each speech regions can be used to voice source waveform estimation using the proposed sequential algorithm using SVD. In glottal closure region, vocal tract parameters are proceeded to improve in sample-by-sample, and compute the estimation error. In transition regions, vocal tract parameters are not improved but only estimation errors are computed. And, in glottal open region, the regions that the estimation errors have almost zero characteristics are selected and the estimation errors are computed using fixed vocal tract parameters in the selected region. Thus, the proposed algorithm for voice source waveform and vocal tract parameters estimation is as Table 1.

Table 1. The proposed algorithm for estimation of voice source waveforms and vocal tract parameters

```
Step 1) $k=1$, $m=1$, $P(0)=I$, $a(0)=0$.
Step 2) Compute and save Kalman coefficients $a(k)$ using (5),(6), and (7).
        If GCI(glottal closure instant), go to Step 3.
        Else, go to Step 2).
Step 3) Divide speech samples into 4-regions using the techniques in this section.
Step 4) $k=m$
Step 5) Compute error signal $e(k)$ using the saved Kalman coefficients.
Step 6) $k=k+1$
```

```
        If transition region, $n=k$ and go to Step 7.
        Else, go to Step 5.
Step 7) Compute error signal $e(k)$ using fixed $a(n)$ parameter, and go to Step 8.
Step 8) If glottal open region, go to Step 9.
        Else, $k=k+1$ and go to Step 7.
Step 9) $r=15$ and Compute $a(k)$ parameter.
Step 10) If transition region, go to Step 11.
         Else, $k=k+1$ and go to Step 9.
Step 11) Select optimum $a(l)$ parameter among $a(k-9),a(k-8),......,a(k)$.
Step 12) $k=n+10$
         Compute error signal $e(k)$ using fixed $a(l)$.
Step 13) If transition region, $n=k+1$ and go to Step 14.
         Else, $k=k+1$ and go to Step 12.
Step 14) Compute error signal $e(k)$ using fixed $a(n)$.
         If the end of transiton region, $m=k$ and go to Step 2.
         Else, $k=k+1$ and go to Step 14.
```

3. EXTRACTON OF COMPOSITE SOURCE MODEL PARAMETERS

In existing voice source source models, source waveforms are represented by simplified sum-of-exponentials types or polynomial types in time domain [4]. These existing simplified models can not represent a variety of voice source characteristics and have the more difficulties in representing in frequency domain. In order to overcome these problems, new model that represent voice source waveforms as weighted sum of basis functions is proposed by Thomson [5]. But, Thomson model has several problems: 1) real voice source waveforms are not effectively modeled by estimated coefficients. 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 composite source model as a new voice source model. Composite source model represents voice source waveforms as a weighted sum of pre-defined basis functions. Voice source waveforms, $y(t)$, in one pitch region excluding glottal closure region can be represented by composite source model as (10).

$$y(t)=\sum_{k=1}^{p}\alpha_k s_k\left(t-\tau_k\right)+n(t) \tag{10}$$

The proposed composite source model in (10) is similiar to Thomson model. But, the proposed model has pre-defined basis functions, time-shift value to represent a variety of waveshapes, and be modeled in only glottal open and transition regions to represent accurate waveshapes.

In the proposed composite source model, ability of representation of frequency characteristics is determined by selection of basis functions. In this paper, we select rectangular, trapezoidal, and beta functions as basis function. Beta function is as following equation and a,b,c,t_{max} are contants.

$$s(t)=a\left(\frac{t}{t_{\max}}\right)^{b}\left(1-\frac{t}{t_{\max}}\right)^{c}, \qquad 0<t<t_{\max} \tag{11}$$

In composite source model that represent voice source waveforms, selection of basis functions is important issue. As a result of the experiment, trapezoidal basis function gives better result compared to rectangular function. But, trapezoidal functions have unwanted high frequency distortion. In order to overcome this problem, we introduce beta function that have improved high-frequency characteristics.

Parameters of the composite source model are weights and time-shfts of basis functions. The parameters can be estimated by EM(estimate maximize) algorithm [7]. To obtain the weights and time-shifts of basis functions, one must solve (12).

$$\min_{\substack{\tau_{1,2,\ldots,p}\\ \alpha_{1,2,\ldots,p}}} \left[\int_T |y(t) - \sum_{k=1}^{p} \alpha_k s_k(t-\tau_k)|^2 \, dt \right] \tag{12}$$

This is a complicated multiparameter optimization problem. Of course, brute force can always be used to solve the problem, evaluating the objective function on a coarse grid to locate roughly the global minimum and then applying the Gaussian method, the Newton-Raphson, or some other gradient search iterative algorithm. However, when applied to the problem at hand, these methods tend to be computationally complex and time consuming. EM(estimate maximize) algorithm can simplify complicated multiparameter optimization problem as following 2-step algorithm. In (13) and (14), *(n)* is a iteration number.

Estimate Step: *For* $k = 1,2,\ldots\ldots,p$

$$\hat{x}_k^{(n)}(t) = \hat{\alpha}_k^{(n)} s_k(t - \hat{\tau}_k^{(n)}) + \beta_k [y(t) - \sum_{l=1}^{p} \hat{\alpha}_l^{(n)} s_l(t - \hat{\tau}_l^{(n)})] \tag{13}$$

Maximize Step: *For* $k = 1,2,\ldots\ldots,p$

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

But, this algorithm can not be directly used to estimation of parameters of composite source model. This problem is arised from the fact that estimated parameters can converge not to global optimum but to local optimum. In order to avoid the problem of local optimum, we introduce region selection concept: the deterministic observation signals are divided into sub-regions that local optimums are not existed, and then the parameters of basis functions are estimated in each region. In this paper, integrated voice source waveforms are divided by two regions: the region having positive slopes and the region having negative slopes. The experiment gives better results for modeling voice source waveforms.

4. EXPERIMENT AND DISCUSSION

We have examined data for voiced sounds from two talkers, one male and one female, using proposed analysis procedure. Voice signals 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. 3 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. 4 shows the results of the proposed composite source model fitting in time-domain. Fig. 5 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].

5. CONCLUSION

In this paper we propose a new voice source estimation algorithm and modeling techniques which can represent a variety of source characteristics. First, we divide speech samples in one pitch region into four parts having different characteristics. Second, the vocal-tract parameters and voice source waveforms are estimated in each regions differently using sequential SVD. Third, we propose composite source model as a new voice source model which is represented by weighted sum of pre-defined basis functions. And finally, the weights and time-shift parameters

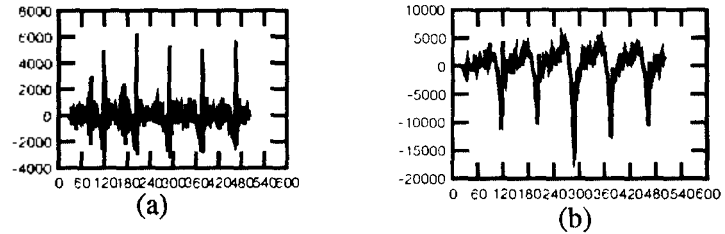

Fig. 3 Experiment of voice source waveform estimation
(a) Kalman filtering (b) proposed method

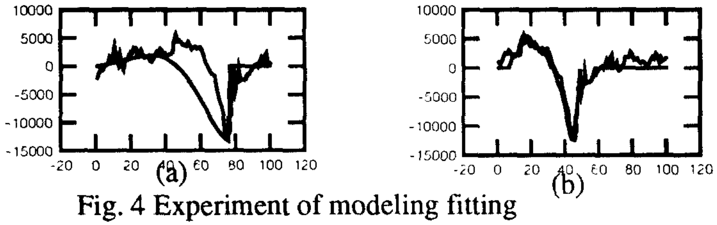

Fig. 4 Experiment of modeling fitting
(a) LF model fitting (b) composite source model fitting

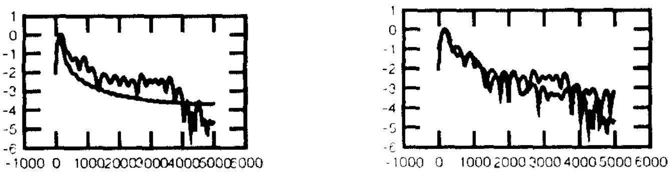

Fig. 5 Experiment of model fitting in frequency domain
(a) LF model fitting (b) composite source model fitting

of the proposed composite source model are estimeted using EM(estimate maximize) algorithm. Experimental results indicate that the proposed method gives more accurate source signals , formants and bandwidths. And, the proposed composite source model as a new voice source modeling method can represent the various voice source characteristics. The proposed methods for voice source estimation can be used to improve the speech quality in speech synthesis and coding techniques.

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