Speech Analysis/Synthesis/Conversion by Using Sequential Processing

B. Panuthat [1], *T. Funada* [1], *N. Kanedera* [2]

[1] Faculty of Engineering, Kanazawa University, Ishikawa, Japan
[2] Ishikawa National College of Technology, Ishikawa, Japan

ABSTRACT

This paper presents a method for speech analysis/synthesis/ conversion by using sequential processing. The aims of this method are to improve the quality of synthesized speech and to convert the original speech into another speech of different characteristics. We apply the Kalman Filter for estimating the auto-regressive coefficients of 'time varying AR model with unknown input (ARUI model)', which we have proposed to improve the conventional AR model, and we use a band-pass filter for making 'a guide signal' to extract the pitch period from the residual signal. These signals are utilized to make the driving source signal in speech synthesis. We also use the guide signal for speech conversion, such as in pitch and utterance length. Moreover, we show experimentally that this method can analyze/synthesize/convert speech without causing instability by using the smoothed auto-regressive coefficients.

1. INTRODUCTION

In most of conventional speech analysis/synthesis methods, processing is executed frame by frame. In the method, the speech signal is assumed to be stable and the parameters are estimated from the whole sampling points in each frame. In this case, many types of window function are used to reduce the influence of end point discontinuity, which causes to produce high-frequency components. The effect of using the window function depends on the frame length and the window type. So that it is difficult to determine appropriate parameter values. Moreover, in the case of synthesis, interpolation must be necessary for smoothing parameters between neighboring frames.

On the other hand, in sequential processing, we do not need the window function nor the parameter interpolation. Thus, we expect that the sequential processing may convert/synthesize speech better than frame by frame processing. In this paper, we show (1) a procedure for estimating time varying auto-regressive coefficients by the Kalman Filter, which we have proposed to improve the conventional AR model, (2) a procedure for estimating the driving source signal, (3) a procedure to convert speech, such as in pitch and utterance length.

In this paper, we confirm experimentally (1) optimal values of the parameters of ARUI model for stable condition, (2) no difference in hearing test of the synthesized speech before and after smoothing the auto-regressive coefficients, (3) successful speech conversion under modification of the driving source signal and the auto-regressive coefficients.

2. TIME VARYING MODEL AND PARAMETER ESTIMATION

2.1 Time Varying AR Model

Consider a speech signal $Y(n)$ described by the conventional *p*-th order AR model

$$Y(n) = \sum_{i=1}^{p} a_i Y(n-i) + V(n) \tag{1}$$

where, $V(n)$ is a zero-mean Gaussian white noise. The auto-regressive coefficients a_i can be estimated from observation data in each frame by using the Yule-Walker equation under the condition that a_i is constant in each frame. However, since a_i changes with time, determining the appropriate values of the frame length and the frame period is difficult.

In the case of time varying AR model, we consider the auto-regressive coefficients as a stochastic process. Representing the time varying auto-regressive coefficients as $A_i(n)$, we rewrite Eq.(1) into Eq.(2).

$$Y(n) = \sum_{i=1}^{p} A_i(n) Y(n-i) + V(n) \tag{2}$$

In this case, we define a state variable vector $\mathbf{X}(n)$ by

$$\mathbf{X}(n) \equiv \left(A_1(n), A_2(n), \cdots, A_p(n)\right)^t \tag{3}$$

and, each $A_i(n)$ is assume to be represented by Gaussian Markov process in Eq.(4).

$$A_i(n+1) = \phi A_i(n) + W_i(n) \tag{4}$$

where ϕ is a constant value between 0 and 1, $W_i(n)$ is a zero-mean Gaussian white noise which is uncorrelated with $V(n)$.

Considering $A_i(n)$ and $Y(n)$ as the system state variable and the system output, respectively, we can rewrite Eqs.(2),(4) in terms of a system dynamic equation as Eq.(6) by using the definition of Eq.(5).

$$\left.\begin{aligned} \mathbf{C}(n) &\equiv (Y(n-1), Y(n-2), \cdots, Y(n-p))^t \\ \mathbf{W}(n) &\equiv (W_1(n), W_2(n), \cdots, W_p(n))^t \\ \Phi &\equiv \phi \mathbf{I}_p \end{aligned}\right\} \quad (5)$$

$$\left.\begin{aligned} Y(n) &= \mathbf{C}^t(n)\mathbf{X}(n) + V(n) \\ \mathbf{X}(n+1) &\equiv \Phi\, \mathbf{X}(n) + \mathbf{W}(n) \end{aligned}\right\} \quad (6)$$

where $(\cdot)^t$ means the transpose of vector $\cdot$. The state vector $\mathbf{X}(n)$ can be estimated by an iterative algorithm of the Kalman filter as Eq.(7).

$$\left.\begin{aligned} \mathbf{K}_n &= \mathbf{P}_n\mathbf{C}(n)\left(\mathbf{C}^t(n)\mathbf{P}_n\mathbf{C}(n) + \gamma\right)^{-1} \\ \hat{\mathbf{X}}_{n|n} &= \Phi\,\hat{\mathbf{X}}_{n|n-1}\mathbf{K}_n\left(y(n) - \mathbf{C}^t(n)\hat{\mathbf{X}}_{n|n-1}\right) \\ \hat{\mathbf{X}}_{n+1|n} &= \hat{\mathbf{X}}_{n|n} \\ \mathbf{G}_n &= \mathbf{P}_n - \mathbf{K}_n\mathbf{C}^t(n)\mathbf{P}_n \\ \mathbf{P}_{n+1} &= \Phi\,\mathbf{G}_n\,\Phi^t + \mathbf{Q} \end{aligned}\right\} \quad (7)$$

where, $n = 0,1,\cdots$, and $\hat{\mathbf{X}}_{j|n}$ is the estimator of $\mathbf{X}(j)$ when the observation data are from $y(0)$ to $y(n)$. The matrix $\mathbf{P}_n$ is the covariance matrix of $\hat{\mathbf{X}}_{n|n-1}$, and $\mathbf{G}_n$ is the covariance matrix of $\hat{\mathbf{X}}_{n|n}$. The estimation of $\mathbf{X}(n)$ can be archived by sequential calculation if the values of $\gamma = \mathbf{E}\left[V(n)^2\right]$, $\mathbf{Q} = \mathbf{E}\left[\mathbf{W}(n)\mathbf{W}(n)^t\right]$ and the initial values, such as P_0, $\mathbf{C}(0)$, $\hat{\mathbf{X}}_{0|-1}$ are given.

The estimate $\bar{y}(n)$ of original signal $y(n)$ can be obtained from Eq.(8).

$$\bar{y}(n) = \sum_{i=1}^{p} \hat{a}_i(n)\bar{y}(n-i) + \hat{v}(n) \quad (8)$$

Where, $\hat{v}(n) \equiv y(n) - \mathbf{C}(\mathbf{n})^t\,\hat{\mathbf{X}}_{n|n-1}$ is the predictive residual signal of the observation data $y(n)$, and $\hat{a}_i(n)$ is the i-th component of the estimate $\hat{\mathbf{X}}_{n|n}$.

2.2 Time Varying ARUI Model

In the previous section, we assumed that the phoneme information $A_i(n)$ and the sound source (prosodic) information $V(n)$ are separable and $V(n)$ is a zero-mean Gaussian white noise. However, if $Y(n)$is in the case of voiced speech, $V(n)$ is the glottal volume signal. Therefore, the assumption concerning $V(n)$ is not valid for separating phoneme and prosodic information.

Thus, we add a new input signal $X(n)$ into Eq.(2) in order to separate the prosodic characteristic from $V(n)$. Eq.(2) can be rewritten as

$$Y(n) = \sum_{i=1}^{p} A_i(n)Y(n-i) + X(n) + V(n) \quad (9)$$

where $\mathrm{X}(n)$,that can not be observed from speech signal, is turbulent noise in the case of unvoiced and glottal wave in the case of voiced. We name $\mathrm{X}(n)$ as 'unknown input' and call this model 'AR model with unknown input (ARUI model)'.

2.3 Estimation of Unknown Input

Assuming the unknown input $\mathrm{X}(n)$ follows Gaussian-Makov process, $\mathrm{X}(n)$ can be estimated by the Kalman filter we describe as follows.

Since $\mathrm{X}(n)$ can be shown as Eq.(10), the coefficient $A_i(n)$ can be estimated same as Eq.(4).

$$X(n+1) = \psi X(n) + W_{p+1}(n) \quad (10)$$

where ψ is a constant value between 1 and 0, and $W_{p+1}(n)$ is a white noise that is uncorrelated with $W_i(n)$ $(i = 1,2,\cdots,p)$. The system dynamic equation can be represented same as Eq.(6) by using Eqs.(4), (9), (10), and changing Eq.(3), (5) to Eq.(11), where Φ in Eq.(6) is defined as Eq.(12).

$$\left.\begin{aligned} \mathbf{X}(n) &\equiv (A_1(n), A_2(n), \cdots, A_p(n), X(n))^t \\ \mathbf{C}(n) &\equiv (Y(n-1), Y(n-2), \cdots, Y(n-p), 1)^t \\ \mathbf{W}(n) &\equiv (W_1(n), W_2(n), \cdots, W_p(n), W_{p+1}(n))^t \end{aligned}\right\} \quad (11)$$

$$\Phi \equiv \begin{pmatrix} \phi \mathbf{I}_p & 0 \\ 0 & \psi \end{pmatrix} \quad (12)$$

Sequential estimation of $\mathbf{X}(n)$ can be calculated by the Kalman filter same as Eq.(7). Synthesized speech can be obtained by Eq.(13), where the source signal $\hat{x}(n)$ is the estimate of the (p+1)-th element of $\mathbf{X}(n)$.

$$\bar{y}(n) = \sum_{i=1}^{p} \hat{a}_i(n)\bar{y}(n-i) + \hat{x}(n) \quad (13)$$

From now, $\hat{x}(n)$ is named as 'U-Input'.

3. SPEECH ANALYSIS/SYNTHESIS

We use 101 syllables and two sentences of Japanese language for analysis/synthesis.

3.1 Experimental Details

Figure 1 presents two kinds of analysis/synthesis systems ((a) and (b)) that we propose in this paper.

System (a): Estimating the U-Input by sequential analysis and using it as a driving source signal for synthesis.

System (b): Band pass filtering the estimated U-Input in the system (a), and using it as a driving source signal for synthesis.

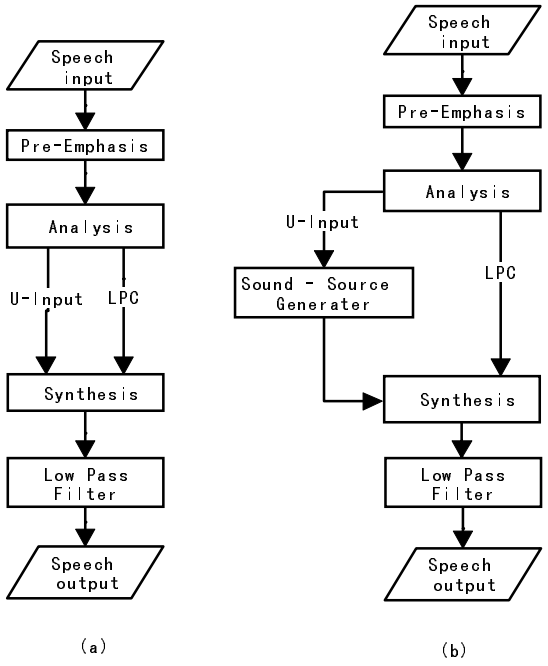


Figure 1. Block diagram of sequential speech analysis/synthesis system.

The parameter values of analysis/synthesis of both systems are assigned as follows:

The analysis order p is between 10 and 29, variance γ of the input white noise $V(n)$ is between 800 and 3000, parameter ψ of the unknown input is between 0.6 and 0.95.

Three types of experiments have been executed as follows:

Type 1: Experiment according to the flow of the diagram (a) without using Pre-Emphasis block and Low Pass filter block. We take this experiment for only comparing with the type 2 experiment.

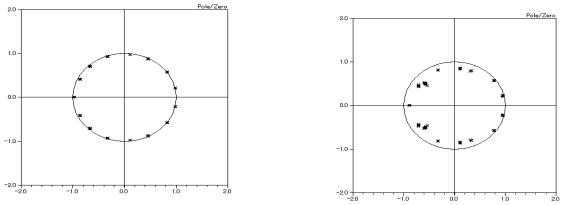
Type 2: Experiment according to the full flow of the diagram (a).

Type 3: Experiment according to the flow of the diagram (b). In addition, in Sound – Source Generator block the band pass filter, pass band of which is within pitch frequency range for extracting the driving source signal.

3.2 Experimental Results

For type 1, some poles of estimated auto-regressive coefficients stay outside the unit circle, which will make estimation unstable. For reducing the oscillation we have to set ϕ less than 0.6 and γ more than 2500. But the power envelope of synthesized signal deviates from the original one.

For type 2, by using Pre-Emphasis block more poles stay within unit circle, which means few oscillation occur. Figure 2 shows the pole position before pre-emphasis and after pre-emphasis. From the result we obtain that if we set ϕ as 0.6-0.85, γ as 500-3000, p as 10-29, the oscillation does not occur. Low Pass filter is used to make the shape of synthesized signal similar to original one. Figure 3 shows the difference of estimated auto-regressive coefficients when the value γ is 500 and 2500. Greater value of γ will make the estimated auto-regressive coefficients smoother than smaller value. Moreover, we obtained from this experiment that the quality of synthesized signal does not depend on the value of γ and shows almost same quality as original one in non-official hearing test.



Before Pre-Emphasis After Pre-Emphasis

Figure 2 Poles of estimated auto-regressive coefficients.

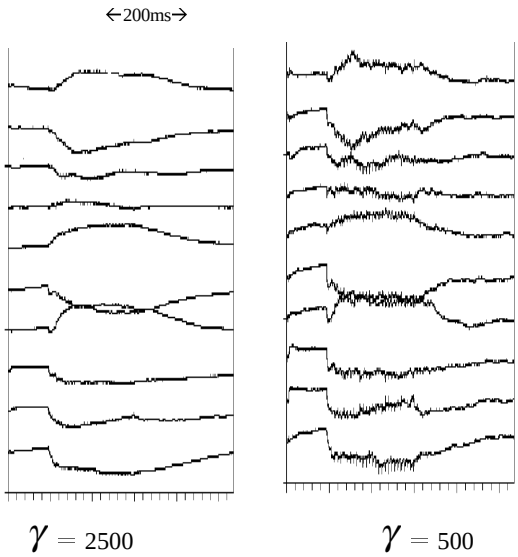


Figure 3 Time varying patterns of auto-regressive coefficients

For type 3, U-Input signal is passed through the band pass filter to extract the glottal signal as shown in figure 4. We use the output signal of this filter as a driving source signal for synthesis. The quality of synthesized speech of this type is not good compared with type 2 because of the lack of individuality. The shape of power envelop of synthesized signal is not same as the original one. However, we can understand the meaning of that speech well in hearing test. ← 10ms →

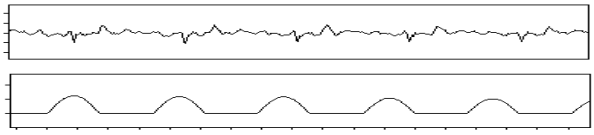


Figure 4 U-input(upper) and extracted glottal signal that is the output of Band Pass filter(lower).

4. SPEECH CONVERSION

We use 101 syllables and two sentences of Japanese language same as the speech analysis/synthesis.

4.1 Experimental Details

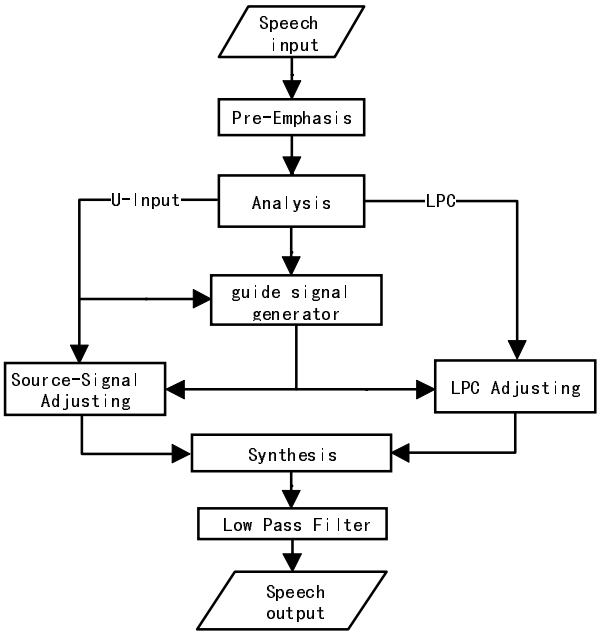


Figure 5 the block diagram of speech conversion.

Figure 5 presents speech conversion systems, which we propose in this paper. To get speech conversion system, three blocks are added into system (a) in figure 1. These three blocks are (1) Guide signal generator block: we use this signal for finding pitch period and determining the intervals of the original signal that are cut/added for converting speech. Figure 6 shows the details of block diagram of this generator. (2) Source-signal adjusting block: we combined this block for adding/cutting the source signal. (3) LPC adjusting block: this block is included for adding/cutting/smoothing the time patterns of auto-regressive coefficients.

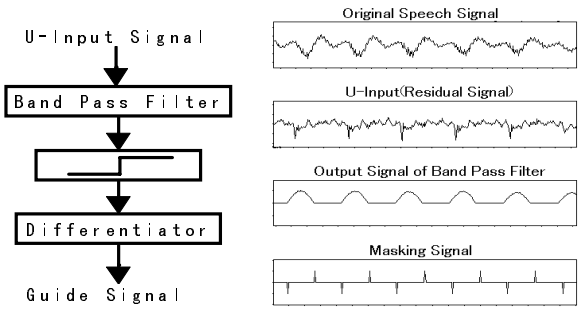


Figure 6 Block diagram of guide signal generator

We added these three blocks for converting the pitch period of utterance and the speed of utterance.

4.2 Experimental Results

From the experiments, after we add Source-Signal Adjusting block, LPC Adjusting block and Guide signal generator block, we can convert the speech pitch/length successfully. Moreover, we confirmed that smoothing the auto-regressive time patterns does not cause any difference in quality of synthesized speech compared with that of non-smoothing.

5. CONCLUSION

This paper shows that the sequential processing can analyze/synthesize/convert speech signal successfully and determine the pitch period by using band pass filter. The experiment in this paper also confirms that the smoothed time pattern of the auto-regressive coefficients does not cause any difference in quality of synthesized speech by hearing test.

6. REFERENCES

[1] G.Kitagawa, W.Gersch: “A Smoothness Priors Time-Varying AR Coefficient Modeling of Non-stationary Covariance Time Series”, IEEE Trans.,AC-30,1(1985).

[2] A.Matthias,etal: “Adaptive AR Modeling of Nonstationary Time Series by Means of Kalman Filtering”, IEEE Trans.BME-45,5(1998).

[3] T.Funada, B.Panuthat: “Sequential Processing for Speech Analysis/synthesis using AR model with Unknown Input”, Acoustical Society of Japan,July 16,1998.

[4] Samuel D. Stearns, Ruth A.David: “Signal Processing Algorithms in Fortran and C”, Prentice-Hall International edition, 1993.

[5] Robert Grover Brown: ”Introduction to random signal analysis and Kalman Filter”, John Wiley & sons, 1983.

[6] Dimitrie C. Popescu, Ilija Zeljkovic: “Kalman Filtering of Colored Noise for Speech Enhancement”, IEEE Trans., pp.997-1000, 1998.