Interpolation of spectral sequences based on Kalman filter

FUMIYASU IKARASHI

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 2001

Keywords: Illusion of Continuity, Phonetic Restration, Kalman filter, RTS algorithm.

1 Introduction

The world around us is a noisy environment. So when we hear some sound, the sound is often mixed or destroyed by noises. However in the situation, we can understand the meaning. This is becase we have an ability to select a target sound from other sounds, refered to "Cocktail-Party Effect" and to restore a missing sound , called "Illusion of Continuity" or "Phonetic Restration". Recently, many reserchers have made a effort to build computer models with such abilities in the framework of "Auditory Scene Analysis".
Aikawa *et al.*[1] reported that the dynamic process of perceiving frequency-modulated (FM) tones can be described by a second order AR model. Masuda *et al.*[2] proposed a model which extrapolates spectral sequencies by extending the FM-tracking model to spectral domain. Sakaguchi[3] also proposed a similar spectral extrapolating model which predicts and tracks spectral peaks represented by Auditory Cortex 1 model. Those models are not enough "Phonetic Restration" models, becase they provided only an extrapolation function, which often gives rise to temporal discontinuity of spectral sequencies.
Hence this paper proposes a spectral interpolating model. The model algorithm is based on Kalman filter set a second order AR model. The interpolation is realized by estimating a series of Line Spectral Frequencies using Kalman filter from both time directions.

2 Spectral sequence estimetion model

2.1 a concept of the model

This study aims at interpolating spectral sequences destroyed by bursts of noise. The interpolation is carried out by estimateing a series of LSFs (Line Spectral Frequencies) using Kalman filter. The concept of this model is illustlated in Figure 1. This model is composed of some parts, that is, Input, Analysis , Search, Interpolation, Synthesis and Output. The following list is a simple explanation of the processing flow.

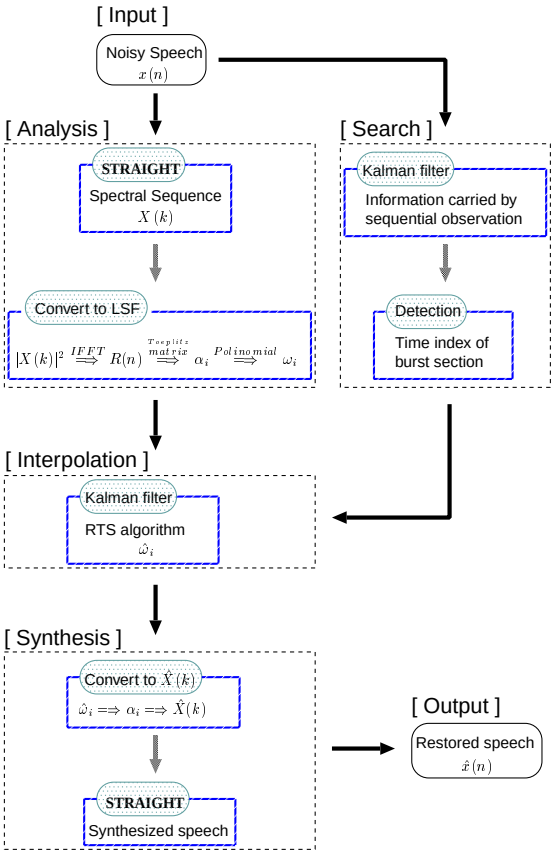

Figure 1: A interpolating model of spectral sequences based on Kalman filter

Input : Enter a noisy speech into this model.

Analysis : Convert the noisy speech to LSFs.

Search : Detect portions of noise bursts directly from the speech wave by Kalman filter.

Interpolation : Interpolate LSFs of noisy sections using RTS algorithm of Kalman filter.

Synthesis : Convert interpolated LSFs to speech.

Output : Emit the restored speech wave.

In the following sections, Kalman filter, Analysis, Search, Interpolation are explained in detail.

2.2 Kalman filter

Dynamic system of Kalmn filter is expressed as follow.

$$\boldsymbol{x}_{k+1} = \boldsymbol{F}_k \boldsymbol{x}_k + \boldsymbol{G}_k \boldsymbol{w}_k, \qquad [\text{ state equation }] \tag{1}$$

$$\boldsymbol{z}_k = \boldsymbol{H}_k \boldsymbol{x}_k + \boldsymbol{v}_k, \qquad [\text{ observation equation }] \tag{2}$$

where $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ are assumed to be white noise sequences described by their second-order statistics

$$
\begin{aligned}
E[\boldsymbol{w}_k] &= 0 && (3)\\
E[\boldsymbol{v}_k] &= 0 && (4)\\
E[\boldsymbol{w}_k {\boldsymbol{w}_i}^T] &= \begin{cases} \boldsymbol{Q}_k, & i = k \\ 0, & i \neq k \end{cases} && (5)\\
E[\boldsymbol{v}_k {\boldsymbol{v}_i}^T] &= \begin{cases} \boldsymbol{R}_k, & i = k \\ 0, & i \neq k \end{cases} && (6)\\
E[\boldsymbol{w}_k {\boldsymbol{v}_i}^T] &= 0, \quad \text{for all } k \text{ and } i && (7)
\end{aligned}
$$

and matrices $\boldsymbol{F}_k$, $\boldsymbol{G}_k$, $\boldsymbol{H}_k$ are known.
Given Equation (1) and (2), Klaman filter provides the linear, minimum mean-squared error estimator of the state $\boldsymbol{x}_k$ given the mesurements $\{\boldsymbol{z}_1,\ \boldsymbol{z}_2,\ \cdots,\ \boldsymbol{z}_k\}$. A commonly used form of Kalman filter is the following.

$$
\begin{aligned}
\hat{\boldsymbol{x}}_{k|k} &= \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k \left[\boldsymbol{z}_k - \boldsymbol{H}_k \hat{\boldsymbol{x}}_{k|k-1}\right] && : \text{estimated value of } \boldsymbol{x}_k && (8)\\
\hat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{F}_k \hat{\boldsymbol{x}}_{k|k} && : \text{prediction value of } \boldsymbol{x}_k && (9)\\
\boldsymbol{K}_k &= \boldsymbol{P}_{k|k-1} {\boldsymbol{H}_k}^T \left[\boldsymbol{H}_k \boldsymbol{P}_{k|k-1} {\boldsymbol{H}_k}^T + \boldsymbol{R}_k\right]^{-1} && : \text{Kalman gain} && (10)\\
\boldsymbol{P}_{k|k} &= \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k \boldsymbol{H}_k \boldsymbol{P}_{k|k-1} && : \text{estimated value of } \boldsymbol{P}_k && (11)\\
\boldsymbol{P}_{k+1|k} &= \boldsymbol{F}_k \boldsymbol{P}_{k|k} {\boldsymbol{F}_k}^T + \boldsymbol{G}_k \boldsymbol{Q}_k {\boldsymbol{G}_k}^T && : \text{prediction value of } \boldsymbol{P}_k && (12)\\
\hat{\boldsymbol{x}}_{0|-1} &= E\left\{\boldsymbol{x}_0\right\} = \bar{\boldsymbol{x}}_0 && : \text{initial value of } \boldsymbol{x}_k && (13)\\
\boldsymbol{P}_{0|-1} &= E\left\{\left[\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0\right]\left[\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0\right]^T\right\} = \boldsymbol{P}_{x_0} && : \text{initial value of } \boldsymbol{P}_k && (14)
\end{aligned}
$$

$\hat{\boldsymbol{x}}_{k|k}$ is refered to as the filtered estimation of $\boldsymbol{x}_k$ and $\hat{\boldsymbol{x}}_{k+1|k}$ is the one-step predictor of $\boldsymbol{x}_{k+1}$. Figure 2 shows a processing flow of Kalman filter.

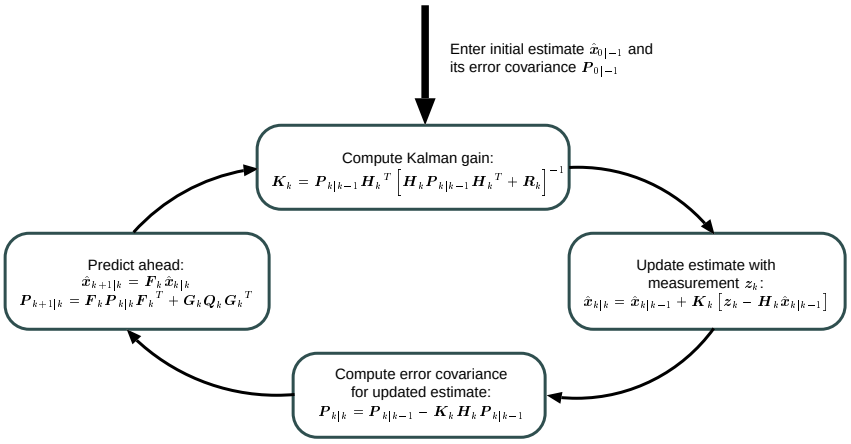

Figure 2: Kalman filter loop

2.3 Analysis

In Analysis, LSFs are computed from input noisy speech. The speech wave is analyzed by STRAIGHT and transfered to spectral sequences. Then the spectral sequences are converted to LSFs.

2.4 Search

In Search, the time index of burst section is estimated directly from input noisy speech by using Kalman filter based on a time varying AR model. In the folowing, the algorithm is described. Speech samples $Y(n)$ are represented using linear predictive coefficients $A_i(n)$ $(i = 1, \cdots, p)$.

$$Y(n) = \sum_{i=1}^{p} A_i(n) Y(n-i) + V(n) \tag{15}$$

The coefficients $A_i(n)$ are assumed to be time variant and to obey the following stochastic process,

$$A_i(n+1) = \phi A_i(n) + W_i(n) \tag{16}$$

where $0 < \phi < 1$, and $W_i(n)$ is white noise. Then Equations of Kalman filter are decided.

$$\boldsymbol{X}(n+1) = \boldsymbol{\Phi}\boldsymbol{X}(n) + \boldsymbol{W}(n) \tag{17}$$

$$Y(n) = \boldsymbol{C}^t(n)\boldsymbol{X}(n) + V(n) \tag{18}$$

where

$$\boldsymbol{C}^t(n) \equiv (Y(n-1), Y(n-2), \cdots, Y(n-p))^t \tag{19}$$

$$\boldsymbol{X}(n) \equiv (A_1(n), A_2(n), \cdots, A_p(n)) \tag{20}$$

$$\boldsymbol{W}(n) \equiv (W_1(n), W_2(n), \cdots, W_p(n))^t \tag{21}$$

$$\boldsymbol{\Phi} \equiv \phi \boldsymbol{I}_p \tag{22}$$

Now "Information carried by sequential observation" is defined as

$$i(n) = \frac{1}{2} \log_2 \frac{|\boldsymbol{P}_n|}{|\boldsymbol{G}_n|} \tag{23}$$

$\boldsymbol{P}_n$ is a priori error covarience and $\boldsymbol{G}_n$ is a posteriori error covarience. Both $\boldsymbol{P}_n$ and $\boldsymbol{G}_n$ are sequentially given in recursive calculations of Kalman filter based on Equation(17),(18). The information $i(n)$ indicates "unexpectedness" of observation $Y(n)$ at time n. So the $i(n)$ enable to detect the section of bursts noise.

2.5 Interpolation

In Interpolation, LSFs of noisy portion are interplorated using Kalman filter. To realize the interpolation, RTS algorithm is adopted. RTS alogrithm is refered to "fixed-interval smoothing". The time interval of mesurements is fixed, and then optimal estimates at all interior points are sought. RTS algorithm consists of two steps; forward sweep and backward sweep. The forward sweep is equivalent to recursive quations (8) - (14). With each step of the forward sweep, it is neccessary to save $\hat{\boldsymbol{x}}_{k|k-1}$, $\hat{\boldsymbol{x}}_{k|k}$, $\boldsymbol{P}_{k|k-1}$, $\boldsymbol{P}_{k|k}$. These are needed for the backward sweep. After

completing the forward sweep, the backward sweep begins with "initial" conditions $\hat{\boldsymbol{x}}_{N|N}$, $\boldsymbol{P}_{N|N}$ obtained as the final computation in the forward sweep. The recursive equations for the backward sweep are

$$\hat{\boldsymbol{x}}_{k|N} = \hat{\boldsymbol{x}}_{k|k} + \boldsymbol{A}_k \left[\hat{\boldsymbol{x}}_{k+1|N} - \hat{\boldsymbol{x}}_{k+1|k}\right] \tag{24}$$

$$\boldsymbol{A}_k = \boldsymbol{P}_{k|k}\boldsymbol{F}^T\boldsymbol{P}^{-1}_{k+1|k} \tag{25}$$

$$(\ k = N-1, N-2, \cdots, 0\ )$$

With each step of the backward sweep, the estimated values by the forward sweep is updated to yield an improved smoothed estimate, which is based on all the measurement data. This RTS algorithm is shown in Figure 3.
After the backward sweep, LSFs except for the estimates of noisy section are replaced by LSFs before filtering. Then the interpolation finishes.

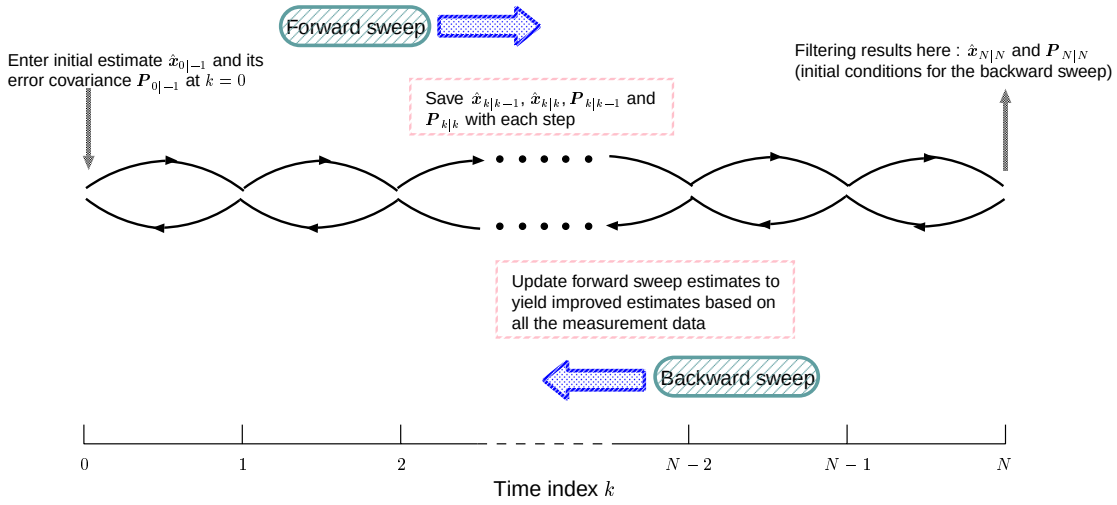

Figure 3: Processing flow of RTS alogorithm

3 Setting a second order AR model to Kalman filter

In order to apply Kalman filter for an interpolation of LSFs, a model which describes behavior of a time series of LSFs is needed. In this paper, it is assumed that the model is described as a second order AR model. And a second order AR model is set to Kalman filter.

3.1 a second order AR model

A second order AR model is defined as follows.

$$y(n) = gx(n-1) - \alpha_1 y(n-1) - \alpha_2 y(n-2) \tag{26}$$

$$\lambda = 2\pi\frac{f_n}{f_s} \tag{27}$$

$$\alpha_1 = -2e^{-\lambda\zeta}\cos\lambda\sqrt{1-\zeta^2} \tag{28}$$
$$\alpha_2 = e^{-2\lambda\zeta} \tag{29}$$
$$g = 1+\alpha_1+\alpha_2 \tag{30}$$

where f_n is the natural frequency, f_s is the sampling frequency, ζ is the damping factor. λ is a normalized frequency between $-\pi$ and π. α is a linear prediction coefficient. g is the gain constant of the system.

3.2 Setting to Kalman filter

A second order AR model is regarded as a kind of predictor. To point out this, $y(n)$ is replaced by $\hat{x}(n)$. If $e(n)$ is assumed to be error between mesurements $x(n)$ and estimates $\hat{x}(n)$, then

$$e(n) = x(n) - \hat{x}(n) \tag{31}$$

Equation(31) is substituted to Equation(26),

$$\hat{x}(n) = gx(n-1) - \alpha_1\hat{x}(n-1) - \alpha_2\hat{x}(n-2) \tag{32}$$
$$x(n) - e(n) = gx(n-1) - \alpha_1\{x(n-1) - e(n-1)\} - \alpha_2\{x(n-2) - e(n-2)\} \tag{33}$$
$$x(n) \approx (g-\alpha_1)x(n-1) - \alpha_2 x(n-2) + \alpha_1 e(n-1) + \alpha_2 e(n-2) \tag{34}$$

Here this formula can be set to Kalman filter,

$$\begin{bmatrix} x(n+1) \\ x(n) \end{bmatrix} = \begin{bmatrix} g-\alpha_1 & -\alpha_2 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} x(n) \\ x(n-1) \end{bmatrix} + \begin{bmatrix} \alpha_1 & \alpha_2 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} e(n) \\ e(n-1) \end{bmatrix} \tag{35}$$
$$z(n) = \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} x(n) \\ x(n-1) \end{bmatrix} + v(n) \tag{36}$$

The matrices of Equation(1),(2) are corresponding to

$$\boldsymbol{F} = \begin{bmatrix} g-\alpha_1 & -\alpha_2 \\ 1 & 0 \end{bmatrix} \tag{37}$$
$$\boldsymbol{G} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 0 & 0 \end{bmatrix} \tag{38}$$
$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 \end{bmatrix} \tag{39}$$

Subscripts of these matrices are omitted. Because these are constant at any time.

4 Simulation of interpolation of LSFs

4.1 Simulation data

A simulated sound data was synthesized with Klatt formant synthesizer. The sound data was conected vowel which has three formants. The synthesis coditions are shown in Table 1 and 2.

Table 1: Synthesis conditions

Sampling frequency		8 kHz
Fundamental frequency		140 Hz
Duration time		500 ms
Band size :	Formant1	80 Hz
	Formant2	120 Hz
	Formant3	150 Hz

Table 2: Transition of each formant

Time [ms]	Formant1 [Hz]	Formant2 [Hz]	Formant3 [Hz]
0∼100	800	1200	2500
100∼300	800 $\Rightarrow$ 250	1200 $\Rightarrow$ 2500	2500 $\Rightarrow$ 3000
300∼500	250	2500	3000

The tansition section (250 $\sim$ 350 ms) of the sound was mixed by white noise whose SNR was -10 dB.

4.2 Experimental conditions

Expeimental parameters are shown in Tabele 3.

Table 3: Experimental conditions

Acoustic parameter	LSF (30th)	
STRAIGHT (Analisis)	Analiysis frame length	40 ms
	Analiysis frame shift	1 ms
	FFT frame length	1024 point
Time variant AR model (Serch)	Order	15
	Autoregressive parameter	$\phi = 0.95$
	Covaiance of observation noise	200
Second AR model (Interpolation)	Damping factor	$\zeta = 0.99$
	Natural frequency	$f_n = 1$ kHz
	Sampling frequency	$f_s = 8$ kHz

$\zeta = 0.99$ and $f_n = 1$ kHz have already decided on LSFs of cleen speech.

4.3 Result

Figure 4 shows the interpolated result. The noisy section was restored to maintain spectral structures before and after the section.

5 Conclusion

In this paper, a method which interpolates noisy sections using Kalman filter was proposed. The model was able to restore a sound with noisy portion.

References

[1] Aikawa,K., Kwahara,H., Tsuzaki,M. , A neural matrix model for active tracking of frequency-modulated tones , ICSLP 96 , vol.1, pp578-581

[2] Masuda-Katsuse,I., Kwahara,H., Aikawa,K. , Speech Segregation Based on Continuity of Spectral Shapes , CASA 97 , pp39-45

[3] Sakaguchi,N. , Study on prediction of perceived spectral sequences based on spectral peak-tracking model frequency-modulated tones , JAIST master thesis , 1998

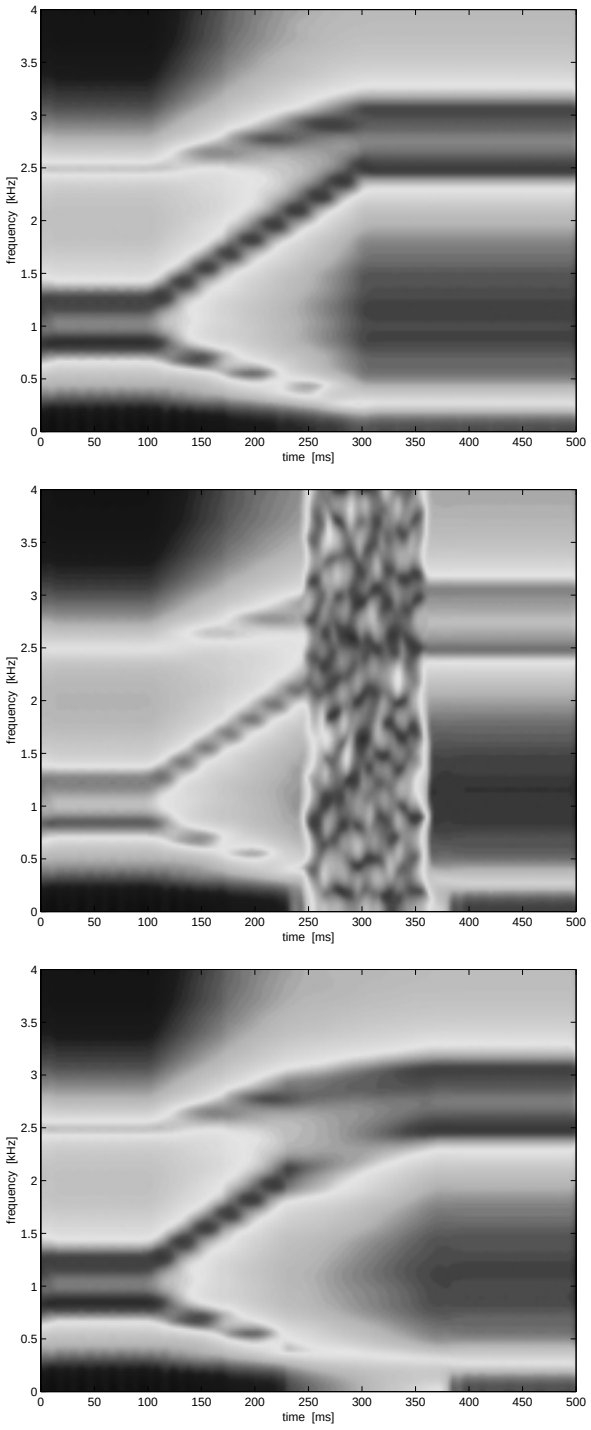

Figure 4: Sound spectrogram : top) cleen speech, middle) input noisy seech, bottom) restored speech