# Speech analysis using the Robust Time-Weighted Kalman filtering

# 시간가중치의 로버스트 칼만필터를 이용한 음성분석

(Hong Sub Choi*, Souguil Ann*)

최홍섭*, 안수길*

## ABSTRACT

In this paper time-varying speech signal is analyzed by using the Kalman filtering methods. In general assuming that the speech process could be stationary in short time duration, frame-based analysis method, such as LPC(Linear Predictive Coding), SSLPC(Sample Selective LPC), has been utilized to obtain the useful information of speech signal, which is, however, not suitable for applying to the time-varying signal. Kalman filtering is generally considered to be an appropriate means for estimation of the time-varying AR(Autoregressive) speech model. Now we consider two limiting factors in using the conventional analysis method. First the familiar Kalman filter procedure has a infinite memory which degrades the ability of adaptive estimation of rapid changing parameter in the current speech. In addition to infinite memory effect, the second is that the sequential Kalman filtering method poorly estimates the parameter coefficients when periodic impulse trains are the excitation source, as in voiced speech. Therefore we propose the robust Kalman filter with time-weighted-error criterion which is applied to analyze the synthetic speech signal.

## 요 약

본 논문에서는 시변형 신호인 음성 신호의 분석에 칼만필터를 이용하였다. 일반적인 음성 분석은 프레임단위의 처리방법인 선형 예측 부호화 기법을 주로 이용하지만 음성의 시변 특성을 파악하는데에는 적절하지 못하다. 따라서 순차적인 추정기법으로 많이 이용되는 칼만 필터를 음성분석에 적용하였다. 또한 음성과 같은 시변신호에서는 과거 신호의 잡음의 분산값에 적당한 가중치를 부가하므로써 과거의 신호에 의해서 현재의 추정값에 미치는 영향을 줄였으며 이를 음성의 천이 구간에서의 파라메타 추정에 사용하였다. 그리고 음성신호 모델에서 생기는 모델링 오차는 일반적으로 백색 가우시안 잡음으로 가정하고 있으나 이는 자음과 같은 무성음에서 특징 파라메타 추정에는 오차가 적지만 모음등의 유성음에서는 음성 발생시의 여기신호인 펄스열에 의해서 많은 모델링 오차를 생기게 한다. 따라서 모델링 오차신호는 Non-Gaussian 확률 분포로 가정한 후 로버스트 칼만 필터를 사용하여 합성음에 대해 특징 파라메타를 추출하였다.

## I. Introduction

Several methods for estimating parameters of a speech production model from observed speech signals only have been developed in speech analysis. In the linear prediction analysis, the speech production process is assumed to be stationary. Accordingly, for any rapid variation of underlying

*서울대학교 공과대학 전자공학과

system parameters or excitation signals, such as stop consonants, fricative onsets, and transition between consonants and vowels, accurate parameter estimation cannot be obtained. Kalman filtering is generally considered to be an effective means of estimating the time-varying coefficients of an AR (autoregressive) speech model, since they overcome some drawbacks of frame-based analysis method[1]. The state-space representation is well suited to sequential estimation. However, the initial application of Kalman filtering methods has an infinite memory so that its ability to adapt to rapid changes in the current speech is affected by the entire history of the signal. So we used a concept of fading memory filter which was first developed for control state estimation[2]. And the familiar Kalman filter theory involoved the use of ideal assumption of the linear system and white noise process for the estimation of the coefficients for speech model. In other words, the sequential Kalman filtering methods produce poor coefficient estimates when periodic impulse trains are the excitation source, as in voiced sounds.

In this paper, we propose a method that is designed to enhance the accuracy of the parameter estimation by the robust Kalman filter which assigns less weight to the small portion of large residuals so that the outliers will not terribly influence the final estimate, while giving unity weight to the bulk of small to moderate residuals. The above procedure takes into account the non-Gaussian nature of the source excitation for voiced speech by assuming that the innovation is from a mixture distribution. Experiments were performed using synthetic speech with transition region between vowel and consonants.

## II. Kalman filter with time-weighted error criterion :

We will assume that speech can be adequately modeled by an AR model represented by the following equation :

$$s(k)=\sum_{i=1}^{p} a_i(k)s(k-i)+e(k) \tag{1}$$

where $a_i(k)$ are time varying coefficients and $e(k)$ represents the error signal.

Assuming the predictor coefficients are constant over an analysis interval (during the closed glottis interval), parameter estimation problem for the system (1) is represented in state-space notation as follows.

$$a(k)=\Phi_k a(k-1) \tag{2}$$

$$s(k)=\mathbf{s}^T(k-1)\mathbf{a}(k)+e(k) \tag{3}$$

where $\mathbf{a}(k)$ is the p-dimensional parameter vector and $\mathbf{s}(k-1)$ is the p-dimensional vector of past observations, which are respectively given,

$$\mathbf{a}(k)=\begin{bmatrix} a_1(k) \\ a_2(k) \\ \vdots \\ a_p(k) \end{bmatrix}, \qquad \mathbf{s}(k-1)=\begin{bmatrix} s(k-1) \\ s(k-2) \\ \vdots \\ s(k-p) \end{bmatrix}.$$

$e(k)$ has the variance $E[e(k)e(j)]=r_k\delta_{kj}$ where $r_k$ is assumed to be known.

$\Phi_k$ is the time-varying parameter transition matrix from $k-1$ instant to $k$ instant, which is assumed to be the identity matrix in most applications.

Then time-weighted Kalman filter is derived as follows.

The usual least-square error criterion can be given as

$$J_k=\sum_{i=1}^{k} r_i^{-1}(s(i)-\mathbf{s}^T(i-1)\mathbf{a}(i))^2 \tag{4}$$

Thus we consider the time-weighting in such a way that the error criterion is weighted so as to decrease the importance of past samples, in other words, backwards increasing the variance $r_i$ of the measurement noise $e(t)$. Therefore now define $\tilde{J}_k$ as a new error criterion, such that

$$\tilde{J}_k=\sum_{i=1}^{k} c^{i-k} r_i^{-1}(s(i)-\mathbf{s}^T(i-1)\mathbf{a}(i))^2 \tag{5}$$

where $c$ is a weighting constant, $c>1$. An expression for the time-weighted Kalman filter algorithm under the new error criterion can be stated as follows[3].

$$\bar{\mathbf{a}}(k)=\Phi_k\hat{\mathbf{a}}(k-1) \tag{6}$$

$$\hat{\mathbf{a}}(k)=\bar{\mathbf{a}}(k)+\mathbf{k}_k v(k) \tag{7}$$

$$v(\mathrm{k})=s(k)-\mathbf{s}^T(k-1)\bar{\mathbf{a}}(k) \tag{8}$$

$$\mathbf{k}_k=M_k\mathbf{s}(k-1)[\mathbf{s}^T(k-1)M_k\mathbf{s}(k-1)+r_k]^{-1} \tag{9}$$

$$M_k=c\Phi_k P_{k-1}\Phi_k^T \tag{10}$$

$$\begin{aligned}P_k&=E[(\mathbf{a}(k)-\hat{\mathbf{a}}(k))(\mathbf{a}(k)-\hat{\mathbf{a}}(k))^T]\\&=M_k-\mathbf{k}_k\mathbf{s}^T(k-1)M_k\end{aligned} \tag{11}$$

where the coefficients $\bar{\mathbf{a}}(k)$, $\hat{\mathbf{a}}(k)$ are the predicted value at $k-1$ instant and the estimated value at $k$ instant, respectively, and $P_k$ is the estimation error covariance.

## III. Robust Time-Weighted Kalman Filter for the Non-Gaussian noise process

In general, the distribution probability density of $e(n)$ in system (1) is not known precisely, but it is easy to find that the error is composed of two parts: one is the error due to fitting vocal tract structure by improper model parameters as well as random noise interferences. This error can be considered as a Gaussian process with a relatively small variance, which exists everywhere in speech signals. Another error components due to glottal source excitation usually appears as a few impulses, and is essentially a non-Gaussian process with a much larger variance.

Thus the error $e(t)$ can be assumed to have a $\varepsilon$-contaminated Gaussian mixture distribution as follows,

$$f_e=(1-\varepsilon)N(\cdot\,|0,1)+\varepsilon N(\cdot\,|0,\sigma_e^2) \tag{12}$$

where $N(x/\mu,\ \sigma^2)$ is a normal density with mean $\mu$ and variance $\sigma^2$ and $\varepsilon$ is the mixing parameter $(0<\varepsilon<1)$. The $\varepsilon$-contaminated normal mixture density is also classified as the term heavy-tailed densities which we mean any distribution whose tail is heavier than some nominal Gaussian distribution. However the fact is well known that the behavior of linear least squares estimates can be quite bad when plant or observation noise are non-Gaussian, particulary when the non-Gaussian is of a heavy-tailed variety giving rise to occasionally very large values[5]. For situations in which large disturbances occur infrequently and at random times, it would be desirable to use a robust Kalman filter which is more or less desensitized to the influence of heavy-tailed distributions. Before proceeding further we present a brief recap on min-max robust stochastic approximation (SA) estimation.

Let $\Gamma$ be a class of estimates, $\Omega$ a class of distributions, and $V(T,F)$ the asymptotic variance of $T\in\Gamma$ when the distribution is $F\in\Omega$. If $E_0$ and $F_0$ satisfy

$$\min_{T\in\Gamma}\max_{F\in\Omega}V(T,F)=V(F_0,T_0)=\max_{F\in\Omega}\min_{T\in\Gamma}V(T,F), \tag{13}$$

we refer to $E_0$ as a min-max robust estimate. $F_0$ is referred to as the least favorable distribution[6]. And SA-estimates are based on Robbins-Monro type stochastic approximation algorithms of the form,

$$T_n=T_{n-1}+\frac{G}{n}\Psi(y_n-T_{n-1}), \tag{14}$$

where $G$ is an appropriate gain constant and $\Psi(\cdot)$ is an appropriate influence function. Selecting the influence function is important because the robustness properties totally depends upon the choice of the influence function. In the literature, there are many functions developed. Among them we can consider well known influence functions defined by

$$\Psi_\epsilon(t) = \begin{cases} t & |t| \leq K \\ K \cdot sgn(t) & |t| > K \end{cases} \tag{15}$$

with K depending upon $\varepsilon$.

$$\Psi_p(t) = \begin{cases} \frac{1}{sy_p} tan\left[\frac{t}{2sy_p}\right] & |t| \leq y_p \\ \frac{tan\left[\frac{1}{2s}\right]}{sy_p} \cdot sgn(t) & |t| > y_p \, , \end{cases} \tag{16}$$

The effect of using an influence function is to assign less weight to the small portion of large residuals so that the outliers will not terribly influence the final estimate, while giving unity weight to the bulk of small or moderate residuals. Since conventional Kalman filter weights all the residuals equally, the large variance process will dominate the estimation accuracy. For non-Gaussian obsrvation noise and Gaussian plant noise, the estimation of $\boldsymbol{a}(k)$ can be solved by robust Kalman filter with influence function $\Psi_p(t)$ in (16) for the residual process with density which goes like $cos^2(t)$ in the middle and has exponential tails. The algorithm is expressed as follows [4]:

$$\bar{\boldsymbol{a}}(k) = \Phi_k \hat{\boldsymbol{a}}(k-1) \tag{17}$$

$$\hat{\boldsymbol{a}}(k) = \bar{\boldsymbol{a}}(k) + M_k \boldsymbol{s}^T(k) \Psi[s(k) - \boldsymbol{s}^T(k-1)\bar{\boldsymbol{a}}(k)] \tag{18}$$

$$M_k = c\Phi_k^T P_{k\ 1} \Phi_k \tag{19}$$

$$P_k = M_k - M_k \boldsymbol{s}^T(k) \boldsymbol{s}(k) M_k E_{F_0} \Psi'[s(k) - \boldsymbol{s}^T(k-1)\bar{\boldsymbol{a}}(k)] \tag{20}$$

$$E_{F_0}\{\Psi'_p(v)\} = (s \cdot y_p)^{\ 2} \left\{1 - p(1 + tan^2(\frac{1}{2s}))\right\} \tag{21}$$

where $y_p$ is defined by $\Phi(-y_p) = \frac{p}{2}$ and p=0.317, s=0.67.

## Ⅳ. Experimental results :

In order to assess the validity of the proposed methods, the speech analysis system was simulated on a digital computer. The synthetic speech signal is created by an all-pole filter ($p$=8) with known time-varying coefficients excited by an impulse train of 100 samples period. The coefficients during a transition interval are varying in a linear interpolation fashion. The speech signal is deemphasized by a simple one pole filter to add the glottal effects. And it was sampled at a frequency of 10 KHz. The analysis interval was set at 25.6ms (256 samples). In this simulation the following values are used :

$$\Phi_k = I,$$
$$\hat{\boldsymbol{a}}(0) = 0,$$
$$P_0 = \begin{bmatrix} 100 & & 0 \\ & \ddots & \\ 0 & & 100 \end{bmatrix},$$
$$r_k = 1.0,$$

c is decided according to the experimental results between $1 < c < 2$.

In order to quantitatively evaluate the estimation error of the spectral envelope obtained by the robust Kalman method and the conventional Kalman method we use the following spectral distortion measures :

$$D = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(10log_{10}f(\omega_i) - 10log_{10}\hat{f}(\omega_i))^2}$$

where $\hat{f}(\omega_i)$ denote estimated spectral densities obtained by the proposed method. The frequency range corresponding to half of the sampling frequency (i.e. 10 KHz) is divided into $L$ (i.e. $L$=80) equal frequency portions. And Fig.1 shows the comparison of the performance of the robust time-weigted Kalman filter (RTKF) algorithm and that of conventional Kalman filter (CKF) algorithms with noise-free synthetic speech signal. The spectral estimation accuracy of the robust time-weighted Kalman filter is superior to that of other method. Especially in the transition region (100-150 samples), RTKF algorithm perform better than CKF algorithm. Note also, however, the estimation accuracy are a little degraded in the transition region than in steady state region.

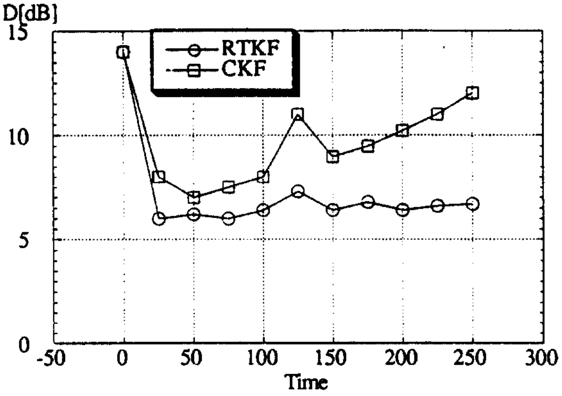

Fig. 1. the comparison of the performance of the robust time-weigted Kalman filter (RTKF) algorithm and conventional Kalman filter (CKF) algorithms.

## V. Conclusion :

We have presented the effect of a robust Kalman filter with time-weighted criterion on the time-varying spectral estimation performance. Using the robust concept in the statistic field, a new Kalman filter is designed to have the ability to be less sensitive to the non-Gaussian noise. And the robust Kalman filter is modified to easily track the parameter variation by adopting a fading memory means. We find that the proposed method has the good simulation results especially in time-varying transition region.

## References

1. J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, Vol.67, pp.1586-1604, Dec. 1979.
2. A.H.Jazwinski, *Stochastic Processes and Filtering Theory*, New York, Academic, 1970.
3. G.A.Mack and V.K.Jain, "Speech parameter estimation by time-weighted-error Kalman filtering," *IEEE Trans. ASSP*, Vol.ASSP-31, pp.1300-1303, Oct. 1983.
4. C.J.Masreliez and R.D.Martin, "Robust Bayesian estimation for the linear model and robustifying the Kalman filter," *IEEE Trans. Automatic Control*, Vol. AC-22, pp.361-371, June 1977.
5. P.J.Huber, "The 1972 Wald lecture-Robust statistics : A review," *Ann. Math. Statist.*, Vol.43, 1972.
6. P.J.Huber, "Robust estimation of a location parameter," *Ann.Math. Statist.*, Vol.35, 1964.

▲Hong Sub, Choi

was born in Seoul, Korea, on October 3, 1957.

He received the B.S., M.S. degree in electronics engineering from Seoul National University, Seoul, in 1985 and 1987, respectly. He is currently working toward the Ph.D. degree at Seoul National University. His research interests include the degital signal processing, speech analysis and recognition.

▲Souguil, Ann

has been professor at the Department of Electronics Engineering, Seoul National University, Seoul, Korea since 1969. He received his B. Sc., M. S. and Ph. D. degrees in electronics engineering from Seoul National University in 1957, 1959 and 1974, respectively. He was lecturer at the Department of Electronics Engineering, Korea Military Academy during 1957-1959. He worked at CEN Saclay Research Institute, France, as a research member during 1959-1960. From 1960 to 1963 he lectured at the Department of Electronics Engineering, Seoul National University and from 1964 to 1968 he worked on a ground tracking station project at Centre Nationale d'Etudes Spatielles, Paris, France. He was at the Aerospace Department, Schlumberger France, as a research member. He is currently serving as Director of Region 10, IEEE, and is consultant to Korea Telecom and the Ministry of Science and Technology, Government of Korea. Dr. Ann is member of Board of Directors, IEEE, and senior member of IEEE.