A Reinvestigation of the Extended Kalman Filter applied to Formant Tracking

Joseph Timoney[1], Tom Lysaght[1], Victor Lazzarini[2] And Ruiyao Gao[3]

[1] *Dept. of Computer Science,* [2] *Department of Music, NUI Maynooth, Maynooth, Co. Kildare,*
[3] *ABB Ltd., Finnabair Industrial Park, Dundalk, Co. Louth, Ireland*
contact author: jtimoney@cs.may.ie,

Abstract. This paper examines the application of the Extended Kalman Filter to formant tracking. The derivation of the Jacobian matrix for the Extended Kalman filter procedure is given. Additionally, it demonstrates how robustness can be incorporated to the procedure. Results are presented to illustrate the formant tracking ability of the non-robust and robust Extended Kalman filter algorithms.

I. Introduction

A key feature in the analysis of speech signals is the formants that describe the time-varying resonant frequencies of the vocal tract. Many approaches have been suggested in the literature for the tracking of formants, with the most popular techniques being derived on a frame-based Linear Prediction coefficient (LPC) analysis of the speech signal. The major drawback of these frame-based techniques is that continuity of the formant estimates across frames must be imposed [1]. An alternative to a frame-based analysis is to use a technique that will track the formants on a sample-by-sample basis, whereby continuity should be ensured because of the proximity of the values. A good choice for this is the well-known Kalman filter which assumes the system to be linear and dynamic. However, its disadvantage for formant tracking is that it does not provide an estimate of the formant values; rather it returns the LPC parameters from which they can be derived afterwards. If the actual formant values are needed there and then the problem is non-linear and in that case an Extended Kalman Filter (EKF) can be applied [2]. However, attempting to implement the algorithm as described in [2] was problematic because the derivation of a required linearization term was unclear. This paper aims to rectify these problems and to produce a complete description of the EKF applied to formant tracking. Additionally, it present results for evaluation purposes. Furthermore, a modification of the EKF algorithm is also made to investigate whether more recent results in robust Kalman filtering [3] confers any performance gains.

II. Method

The EKF uses a non-linear state space description of the system as given by

$$\mathbf{x}_{k+1} = \mathbf{f}(k, \mathbf{x}_k) + \mathbf{w}_k \quad (1)$$

$$\mathbf{y}_k = \mathbf{h}(k, \mathbf{x}_k) + \mathbf{v}_k \quad (2)$$

where, $\mathbf{y}_k$ is the observed output, $\mathbf{x}_k$ is the state, $\mathbf{f}(.)$ and $\mathbf{h}(.)$ describe non-linear, and possibly time-varying, state transition and measurement matrices respectively. $\mathbf{w}_k$ and $\mathbf{v}_k$ are independent zero-mean white Gaussian noise processes with covariance matrices $\mathbf{Q}_k$ and $\mathbf{R}_k$ respectively.

The state transition and measurement matrices are linearized with respect to the most recent state estimate, either $\hat{\mathbf{x}}_k$ or $\hat{\mathbf{x}}_k^-$, by first constructing the Jacobian matrices of partial derivatives

$$\mathbf{F}_{k+1,k} = \left.\frac{\partial \mathbf{f}(k,\mathbf{x}_k)}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_k} \quad \text{and} \quad \mathbf{H}_k = \left.\frac{\partial \mathbf{h}(k,\mathbf{x}_k)}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_k^-} \tag{3}$$

These matrices are then employed in a first-order Taylor series approximation of the non-linear functions $\mathbf{f}(.)$ and $\mathbf{h}(.)$ around $\hat{\mathbf{x}}_k$ or $\hat{\mathbf{x}}_k^-$ respectively. The EKF recursion equations can then be written as follows

Initialization: for k=0, set the initial state estimate as

$$\hat{\mathbf{x}}_0 = E[\mathbf{x}_0] \tag{4}$$

and the initial error covariance as

$$\mathbf{P}_0 = E\left[(\mathbf{x}_0 - E[\mathbf{x}_0])(\mathbf{x}_0 - E[\mathbf{x}_0])^T\right] \tag{5}$$

Computation: for k=1,2, …. compute:

a) The propagation of the state estimate: $\hat{\mathbf{x}}_k^- = \mathbf{f}(k, \hat{\mathbf{x}}_{k-1}^-)$ (6)

b) The propagation of the Error Covariance: $\mathbf{P}_k^- = \mathbf{F}_{k,k+1}\mathbf{P}_{k-1}\mathbf{F}_{k,k+1}^T + \mathbf{Q}_k$ (7)

c) The Kalman gain matrix: $\mathbf{G}_k = \mathbf{P}_k^-\mathbf{H}_k^T\left[\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k\right]^{-1}$ (8)

d) The State Estimate Update: $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{G}_k\mathbf{y}_k - \mathbf{h}(k, \hat{\mathbf{x}}_k^-)$ (9)

e) The Error Covariance update: $\mathbf{P}_k = (\mathbf{I} - \mathbf{G}_k\mathbf{H}_k)\mathbf{P}_k^-$ (10)

In the formant tracking problem the state vector is assumed a set of m formants frequencies and their respective bandwidths [2],

$$\mathbf{x} = [F_1, F_2, \ldots, F_m, B_1, B_2, \ldots, B_m]^T$$

It is assumed that the dynamic model for state update is linear, that is the formant values and bandwidths at time k+1 are equal to the current values plus some deviation

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{w}_k \tag{11}$$

Thus, the state transition function $\mathbf{f}(.)$ and its corresponding linearized matrix $\mathbf{F}_{k+1,k}$ are given by the identity matrix $\mathbf{I}$. It is assumed that the speech signal is produced by an all-pole model of order n, with LPC parameters $a_1, \ldots, a_n$ giving an output value

$$y(k) = -a_1 y(k-1) - \cdots - a_n y(k-n) + v_k \tag{12}$$

Letting

$$\mathbf{S} = [y(k-1)\ y(k-2) \ldots y(k-n)]^T \quad \text{and} \quad \mathbf{a} = [a_1\ a_2 \ldots a_n]^T \tag{13}$$

means that (12) can also be rewritten as

$$y(k) = -\mathbf{S}^T\,\mathbf{a} \tag{14}$$

The vocal tract transfer function of the all-pole model can be expressed as the cascade of m=n/2 resonators, each one representing a single formant resonance,

$$\frac{1}{1+a_1 z^{-1}+\cdots+a_n z^{-n}}=\left(\frac{1}{1+c_1 z^{-1}+d_1 z^{-2}}\right)\cdots\left(\frac{1}{1+c_m z^{-1}+d_m z^{-2}}\right) \tag{15}$$

The formants and bandwidths are contained in the resonator coefficients and are given by

$$c_j=-2e^{-\pi B_j T}\cos\left(2\pi F_j T\right) \tag{16}$$

$$d_j=e^{-2\pi B_j T} \tag{17}$$

From (3), to apply the EKF the matrix $\mathbf{H}_k$ must be computed for steps (c) and (e), as described by equations (8) and (10), of the procedure

$$\mathbf{H}_k=-\mathbf{S}^T\frac{\delta\mathbf{a}}{\partial\mathbf{x}_k} \tag{18}$$

A recursive equation for the computation of (18) was proposed in [2]. Within this equation there are three terms. It is with the definition of the update of the second and third terms, $\mathbf{a}^*$ and $\mathbf{a}^{**}$, in [2] where the difficulty occurred as the procedure is not entirely clear. An attempted implementation of the EKF using a best possible interpretation of this recursion equation was seen not to produce good formant tracks but ones whose values fluctuated in an unstable manner. A subsequent expansion by hand of the recursive equation was also found not to correspond with the true value of $\mathbf{H}_k$ if computed directly from (18). Therefore a new expression for the recursion was required. Analysis following an expansion of $\frac{\delta\mathbf{a}}{\partial\mathbf{x}_k}$ for m formants lead to the observation that each of columns i, denoted as $\mathbf{col}(i)$, of this (2m x 2m) matrix could be generated by the expressions

$$\underset{i=1,\ldots m}{\mathbf{col}(i)}=\frac{\partial c_i}{\partial F_i}.\,\mathbf{C}_i \text{ and } \underset{i=m+1,\ldots 2m}{\mathbf{col}(i)}=\frac{\partial c_{i-m}}{\partial B_{i-m}}.\,\mathbf{C}_{i-m}+\frac{\partial d_{i-m}}{\partial B_{i-m}}.\,\mathbf{D}_{i-m} \tag{19}$$

where the '.' denotes the multiplication of each element and the vectors $\mathbf{C}_i$ and $\mathbf{D}_i$ are given by

$$\mathbf{C}_i=\left[\prod_{l\neq i,l=1}^{m}\begin{bmatrix}1 & c_l & d_l\end{bmatrix}\;0\right]^T \text{ and } \mathbf{D}_i=\left[0\;\prod_{l\neq i,l=1}^{m}\begin{bmatrix}1 & c_l & d_l\end{bmatrix}\right]^T \tag{20}$$

where the product symbol ($\mathbf{\Pi}$) denotes a polynomial multiplication (vector convolution), and the final and first entry in $\mathbf{C}_i$ and $\mathbf{D}_i$ is zero respectively. This means that the partial derivative in (18) is

$$\frac{\delta\mathbf{a}}{\partial\mathbf{x}_k}=\left[\mathbf{col}(1)\quad\mathbf{col}(2)\ldots\mathbf{col}(m)\quad\mathbf{col}(m+1)\ldots\mathbf{col}(2m)\right] \tag{21}$$

To create a robust EKF, the procedure outlined in [3] can be applied to (7) and (8) to give

$$\mathbf{P}_k^-=(1/\lambda)\mathbf{F}_{k,k+1}\mathbf{P}_{k-1}\mathbf{F}_{k,k+1}^T+\mathbf{Q}_k \tag{22}$$

$$\mathbf{G}_k=\mathbf{P}_k^-\mathbf{H}_k^T\left[\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k w^{-1}\left(\hat{\varepsilon}(k)\right)\right]^{-1} \tag{23}$$

where λ is a forgetting factor and $w^{-1}\left(\hat{\varepsilon}(k)\right)$ is a weighting function derived from the residual model error.

III. Experiments

To test the EKF approach ideally a database of speech utterances with hand-labelled formants would be used, but because this was unavailable, the approach of [4] was followed to create synthetic sounds. Forty-eight speech utterances were processed using Wavesurfer software [5], and the pitch and formants extracted. Only formant data in voiced speech regions were retained and then interpolated. These formants were then used to generate synthetic speech. Both EKF algorithms were applied to the synthetic sounds. $\mathbf{Q}$ and $\mathbf{R}$ were diagonal matrices whose covariance was 50 and 0.1 respectively for the first test, and 50 and 0.005 respectively for the second test. The initial formant values were set to be $\{400, 1500, 2500, 3500\}$Hz, all having a bandwidth of 50Hz. The covariance error matrix was diagonal and its entries were initialised at 100. In the case of the robust EKF λ was chosen to be 0.9975 as lower values sometimes resulted the matrix inversion in (23) being ill-conditioned. The histogram peak and standard deviation of the absolute error difference was found between the synthetic and tracked formant values for both tests. Table 1 presents the results.

	Peak Error $\mathbf{R} = 0.1 \times \mathbf{I}$	Std. Error $\mathbf{R} = 0.1 \times \mathbf{I}$	Peak Error $\mathbf{R} = 0.005 \times \mathbf{I}$	Std. Error $\mathbf{R} = 0.005 \times \mathbf{I}$
EKF	17.24 Hz	201 Hz	6.41 Hz	161.63Hz
Robust EKF	14.41 Hz	193.4 Hz	8.78 Hz	159.8 Hz

Table 1 Histogram Peak and Standard Deviation of Absolute Error of both EKF approaches

IV. Conclusions

From the results it can be seen that the user-defined measurement error covariance is the most important parameter in determining the accuracy of the EKF formant tracking. Taking account of robustness does help to improve algorithm performance but not significantly and not completely. Setting the measurement error covariance to be $\mathbf{R} = 0.005 \times \mathbf{I}$ most frequently gives a value for formant error that is reasonably small but it appears that the standard deviation of the error is large and would suggest that the EKF approach is not ideal for formant tracking. However, some of the blame can be attributed to the formant values used to create the synthetic speech, as they were derived using another tracking algorithm [5] and thus were not always reliable or smooth, being subject to considerable fluctuations on occasion. Future work will seek proper hand-labelled formant data to verify to what degree this was a source of error.

References

[1] Timoney, J., and Lysaght, T., "Instantaneous frequency approaches for speech and audio signal analysis," *Sounds Electric 3*, Maynooth, 2001.

[2] Rigoll, G., "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended kalman filter", *ICASSP 86*, Tokyo, Japan, 1986.

[2] Yang, T. et al., "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended kalman filter", *Signal Processing*, vol. 63, 1997, pp. 151-156.

[4] Malkin, J., et al, 'A Graphical Model for Formant Tracking', *ICASSP 05*, Philadelphia, USA, 2005.

[5] Available at http://www.speech.kth.se/wavesurfer/