

Unscented Kalman Filtering of Line Spectral Frequencies

Andrew Errity^{*}, John McKenna^{*} and Stephen Isard^{**}

^{*}School of Computing
Dublin City University, Dublin 9
{aerrity, john}@computing.dcu.ie

^{**}Centre for Speech Technology Research
University of Edinburgh
s.isard@ed.ac.uk

Abstract

We propose a new method for estimating Line Spectral Frequency (LSF) trajectories that uses unscented Kalman filtering (UKF). This method is based upon an iterative Expectation Maximisation (EM) approach in which LSF estimates are generated during a forward pass and then smoothed during a backward pass. The EM approach also provides re-estimated Kalman filter parameters for further forward-backward passes that improve estimation.

This approach exploits the non-independence of neighbouring spectra. We estimate LSFs as they have good interpolation and quantization properties. This allows us to estimate LSF trajectories that are guaranteed to result in stable filters. We analyse noisy synthetic speech using this technique. The results compare favourably with other methods.

1. Introduction

Linear Prediction has been in widespread use throughout speech processing yet has a number of undesirable properties. The Linear Prediction coefficients (LPC) have a large range and even a small change to the LPCs can result in a relatively large change in the corresponding filter's poles. Also, the LPCs have no physical significance unless they are converted to another form, which makes it difficult to identify a stable filter from LPCs. Thus, the LPCs are poorly suited to interpolation and quantization and better representations have been sought.

Line Spectral Frequencies [1] are a popular alternative representation. The LP polynomial can be decomposed into a pair of symmetric and anti-symmetric polynomials, the Line Spectral Pair (LSP) polynomials. This representation has the useful property that the roots of the symmetric and anti-symmetric polynomials are interlaced and all lie on the unit circle. When polynomials maintaining these properties are summed, they result in a polynomial that is minimum phase, guaranteeing a stable filter. The LSPs can be represented by the angle of the roots of the polynomials, referred to as Line Spectral Frequencies. As these roots occur in complex conjugate pairs, roots in the lower half of the complex plane can be ignored resulting in LSFs bounded between 0 and π .

This paper details a new approach to the problem of estimating LSFs from speech. This equates to the problem of estimating a hidden state given a series of noisy measurements. The Kalman filter (KF) has been proven to provide an optimal, in the minimum mean-square error (MMSE) sense, solution to linear forms of this problem. Kalman filtering has been successfully applied to several speech processing problems. Recently, KF methods based upon EM have been used to track LPCs for speaker-characterisation [2], and to track LSFs (using an estimated

linear relationship between LSFs and speech) for use in joint cost computation in unit-selection speech synthesis [3].

As the relationship between LSF values and speech is nonlinear, we must look to nonlinear variants of the KF. The most widely used nonlinear adaptation of the KF is the extended Kalman filter (EKF), which has been shown to have several flaws. The EKF has been used in the past to estimate formant trajectories [4]. The unscented Kalman filter [5,6] provides an alternative to the EKF. It substitutes a deterministic sampling method in place of the EKF's simple linearization approach. The UKF has recently been successfully applied to the problem of signal and parameter estimation for single and multi-microphone speech enhancement [7]. We propose using the UKF in an EM iterative approach to estimating LSF trajectories directly from speech.

The outline of this paper is as follows. Firstly, we detail the discrete-time nonlinear dynamic system model used to represent the problem and provide some background on the UKF. Then we describe the recursive procedure used to produce LSF estimates and refine our system parameters. Finally, we compare this approach to similar methods and report our findings.

2. Unscented Kalman Filtering of Speech

2.1. The Model

The filtering problem addressed in this paper is that of finding the best (MMSE) estimate of the unobserved LSFs given an observed speech signal. This can be modeled as the state-space equations (1) and (3).

$$s_n = h(x_n) + v_n \quad n = 1, 2, \dots, N \quad (1)$$

where, s_n , the measurement, is the speech sample at time n ; x_n , the (hidden) state, is the set of p LSF values which are related to s_n by the non-linear function $h(\cdot)$; v_n is the measurement noise, assumed Gaussian with probability distribution $p(v) \sim N(0, R)$. The measurement function $h(\cdot)$ relates LSF values to speech and is equal to

$$h(x_n) = H_n g(x_n) \quad (2)$$

where $g(x_n)$ is a nonlinear function transforming LSFs to LPCs; and H_n relates the LPC values to s_n using p preceding samples, $H_n = [s_{n-1} \dots s_{n-p}]$.

$$x_n = \Phi x_{n-1} + w_n \quad n = 1, 2, \dots, N \quad (3)$$

where, Φ is a linear function relating the previous state estimate to the current state estimate; w_n is the process noise (uncorrelated with v_n), with probability distribution, assumed Gaussian, $p(w) \sim N(0, Q)$. We also calculate a measure of our confidence in the estimate at each time step in the form of an estimate error covariance matrix P_{x_n} .

2.2. The Unscented Kalman Filter

The well-known Kalman filter can be used for linear estimation of a hidden state given a sequence of observations. The extended Kalman filter was developed for use in the nonlinear case. In the EKF, a simple linearization of the state and measurement functions is performed allowing the traditional KF equations to be applied. Unfortunately, the EKF has two major flaws:

- Derivation of the required Jacobian matrices is often complex and can lead to implementation difficulties
- Linearization approximations made can introduce large errors and may lead to filter instabilities

In order to address these flaws the unscented Kalman Filter was developed [5], [6]. The principle behind the UKF is the assumption that it is easier to approximate a Gaussian distribution than it is to approximate a non-linear function. In the UKF, an unscented transform (UT) is used to propagate the random state variable through the non-linear functions (in place of the linearization procedure of the EKF). In the UT a number of weighted *sigma* points are carefully chosen such that they capture the mean and covariance of the state variable as follows (the subscript n is dropped for notational clarity)

$$W_0^m = \lambda / (p + \lambda)$$

$$W_0^c = \lambda / (p + \lambda) + (1 - \alpha^2 + \beta) \quad (4)$$

$$W_i^m = W_i^c = 1 / (2(p + \lambda)) \quad i = 1, \dots, 2p$$

$$\mathcal{X}_0 = \bar{x}$$

$$\mathcal{X}_i = \bar{x} + \left(\sqrt{(p + \lambda) P_x} \right)_i \quad i = 1, \dots, p \quad (5)$$

$$\mathcal{X}_i = \bar{x} - \left(\sqrt{(p + \lambda) P_x} \right)_{i-p} \quad i = p + 1, \dots, 2p$$

where, $\lambda = \alpha^2(p + \kappa) - p$; $(\dots)_i$ is the i th matrix row (or column); α is a parameter used to scale the spread of the *sigma* points around \bar{x} , which we set equal to 0.001; β allows any prior knowledge of the distribution of x to be included, we set this to 2 as this is optimal for Gaussian distributions; κ is a further scaling parameter which we set equal to zero.

These *sigma* points are then transformed through the true non-linear functions without approximation. The *a priori* statistics of the transformation of the original variable can then be estimated from the transformed *sigma* points.

$$S_i = h(\mathcal{X}_i) \quad i = 0, \dots, 2p \quad (6)$$

$$\bar{s} \approx \sum_{i=0}^{2p} W_i^m S_i \quad (7)$$

$$P_s \approx \sum_{i=0}^{2p} W_i^c [S_i - \bar{s}][S_i - \bar{s}]' \quad (8)$$

The UKF adapts the KF measurement update equations to use *a priori* estimates produced by UT to estimate the *a posteriori* state and covariance (for a detailed discussion refer to [6]).

2.3. The EM Algorithm

On our first iteration through the speech, we must choose some initial filter parameters. The initial state estimate, x_n , was set equal to LSF values calculated from LP autocorrelation analysis of a 25ms window taken from the signal start. R was set equal to the standard deviation of the error residual found by inverse filtering this window. Q was set equal to the identity matrix times 10^{-5} , an empirically chosen small number. Φ was chosen as the identity matrix as we assume no

prior knowledge of the vocal tract configuration, i.e. we initially assume that it remains approximately the same from one sample to the next and allow Q to model any drift. Our confidence in the initial estimate, P_{x_0} , is maintained at a reasonable baseline level for all iterations. A value of 0.01 was found to be appropriate given the accuracy of our initial state estimate.

The KF can be used in an EM iterative approach [8] which, having performed a forward and backward pass through all the data, yields maximum likelihood estimates for the filter parameters for use in the next iteration as well as handling missing data samples. This EM approach can be adapted for use in the non-linear case. During the ‘expectation’ step a forward pass is made through the data using the UKF equations to produce state estimates and estimate error covariance matrices at each time step. The square-root implementation of the UKF [9] is used as it provides numeric stability and guarantees positive semi-definiteness of the state covariances. We add further robustness to our UKF by rejecting sample points that produce a measurement estimate error greater than 3 times the standard deviation of the measurement noise. Such sample points are treated as missing data.

Backward passes are performed using the Rauch-Tung-Striebel [10] smoothing equations to produce state estimates based on all observations; these equations are valid as our process model is linear. The initial state estimate calculated by the backward pass, x_0^N (where the superscript denotes the number of data samples available), is used as the x_0 value for the next iteration. The ‘maximization’ step involves re-estimating the UKF parameters using the smoothed state estimations. Re-estimation of the measurement noise covariance, R , requires calculation of the backward pass measurement prediction, $h(x_n^N)$, and its covariance matrix $P_{s_n^N}$. An unscented transform of x_n^N through the non-linear function $h(\cdot)$ is performed during the backward pass. This results in an estimate of $h(x_n^N)$ and $P_{s_n^N}$ which can be used to re-estimate R as

$$R = \frac{1}{N} \sum_{n=1}^N \left[(s_n - h(x_n^N))(s_n - h(x_n^N))' + P_{s_n^N} \right] \quad (9)$$

In order to produce re-estimates of the process parameters, Φ and Q , a lagged estimate error covariance, $P_{x_{N-1}^N}$, must be calculated. This is calculated during the backward pass and requires that the final lagged covariance be calculated first. This has been shown to be

$$P_{x_{N-1}^N} = \Phi P_{x_{N-1}} - Kh(\Phi P_{x_{N-1}}) \quad (10)$$

$h(\Phi P_{x_{N-1}})$ is difficult to calculate as $\Phi P_{x_{N-1}}$ is asymmetric. To overcome this, at the very last sample, we assume Φ equals the identity matrix; the effects of this simplification are negligible, particularly over a single pair of neighbouring LSFs. To calculate $h(P_{x_{N-1}})$ we perform an unscented transform of x_{N-1} , with estimate error covariance $P_{x_{N-1}}$, through $h(\cdot)$ and find that

$$h(P_{x_{N-1}}) = P_{s_{N-1}} \left(P_{x_{N-1}} P_{s_{N-1}}^{-1} \right)' \quad (11)$$

where, P_s is the measurement estimate error covariance and $P_{x\bar{s}}$ is the covariance of the state estimate error and measurement estimate error. Both P_s and $P_{x\bar{s}}$ can be calculated using the UKF equations. Thus in the non-linear case the final lagged estimate error covariance is

$$P_{xN,N-1}^N = \Phi P_{xN-1,N-1} - KP_{sN-1} \left(P_{sN-1,N-1} P_{sN-1}^{-1} \right) \quad (12)$$

Subsequent values of the lagged covariance, and hence the re-estimated Φ and Q can be calculated using the standard equations.

The performance of the re-estimated parameters can be measured by computing a log-likelihood score during the forward pass [8]. After only 3-4 iterations, changes in the log likelihood value become negligible. At this point satisfactory estimates result.

3. Evaluation & Discussion

3.1. Analysis of Synthesized Speech

In order to evaluate the performance of our approach we generated a suite of synthetic data to perform analysis over. The suite is similar to that used in [11].

We generated sets of glottal excitations and formant filters that produced sets of synthetic male and female speech for three monophthongs (/a/, /i/ and /u/) and three diphthongs (/ai/, /au/ and /ui/). The excitation had two components: an LF-modeled glottal pulse train and added noise bursts. The noise bursts were Gaussian white noise modulated by a rectangular window, centered on the instant of glottal closure. F0 was set to 100Hz and 200Hz for the male and female sets respectively. The male excitation had LF parameter values that correspond to modal speech and the female speech was given breathy parameters. The duration of the noise burst was varied in 3 steps: 0%, 60% and 100% of the pitch period. The intensity of the noise was varied in 4 steps; the power ratio, or harmonics-to-noise ratio (HNR), between the glottal pulses and the noise bursts was set to ∞ dB (no added noise), 20dB, 10dB and 5dB.

One synthetic excitation was generated for each condition and applied to sets of 10 LP coefficients (that represented 5 formants) for each of the 6 vocalic sounds. There were 7 versions of each vocalic sound for both male and female, giving a total of 84 synthesized vocalic segments. Each segment was 0.2s in length with a sampling frequency of 11025Hz.

We performed analysis over two such data sets, giving a total of 168 segments. An analysis order of 12 was used to simulate the real-life situation where the order that should be applied to real speech is generally unknown. Three iterations of our UKF EM approach were performed on each segment. The resulting LSF trajectories were found to be smooth and reasonable estimates of the true values. This can be seen in the equivalent formant trajectories (Fig. 1). The LSF estimates were found to produce a stable LP filter in all segments.

Our UKF encountered numerical instabilities when analyzing a small number of the segments (1.79%). We plan to investigate means of increasing the robustness of our approach in the future.

It should also be noted that this system is best suited to offline processing due to its computational complexity.

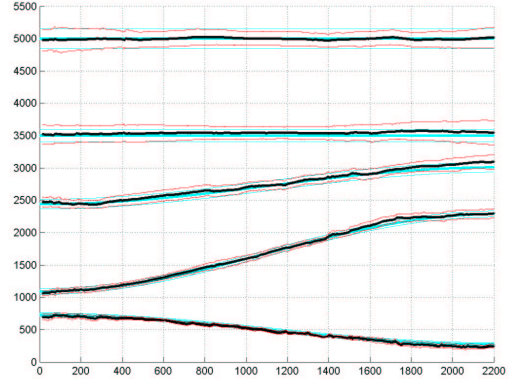


Figure 1: Formant estimation from a synthesized male diphthong /ai/ with 10dB HNR applied to 60% of the pitch period. Bandwidth delimiters are shown with thin lines. Lighter lines indicate true formants; darker lines represent estimates.

3.2. Comparison with Other Methods

We used the previously proposed KF method for estimating LP coefficients [2] to analyse the data set generated for the above experiment. As before, an analysis order of 12, and 3 EM iterations were used. LPC estimates were calculated for each segment in the data set.

As LP has been shown to possess poor interpolation qualities we expected that the LPC trajectories produced by KF would lead to instabilities in the resulting LP filter. This proved to be true, as only 28.57% of the segments contained no unstable LP coefficients. Thus, UKF of LSFs performs significantly better than KF of LPCs in terms of producing a stable filter.

In order to perform a comparison, the LPC estimates produced by KF (KFLPC) and LSF estimates produced by UKF (UKFLSF) were each converted to cepstral coefficients, c_k , of order 22. All unstable poles produced by KFLPC were reflected inside the unit circle prior to cepstral conversion. A truncated cepstral distance measure (13) was then used to evaluate the distance from each method's estimates to the true cepstral values (calculated from the true LPCs),

$$\sum_{k=1}^K (c_k - \hat{c}_k)^2 \quad (13)$$

where \hat{c}_k are the cepstral coefficients derived from the estimates. The means of the cepstral distances of each method for the entire data set are reported in Table 1. The UKFLSF approach is seen to be comparable to the KFLPC method in terms of spectral accuracy.

We also used discrete all-pole (DAP) modeling [12] to analyse our test set. This was done on a frame-by-frame basis with frame length 25ms and a frame shift of 10ms. Linear interpolation between frames was carried out on LSFs calculated from DAP derived LPCs. The DAP analysis produced a number of unstable filters, the poles of which were reflected back inside the unit circle prior to conversion to LSFs. The results are summarised in Table 1 and show that the UKFLSF outperforms DAP modeling in terms of spectral accuracy.

In order to compare trajectory smoothness, we first down-sampled the UKFLSF and KFLPC values at the frame rate

used for DAP analysis. For each method, we then calculated the mean square difference (MSD) between sets of cepstral coefficients in neighbouring frames as follows

$$\left(\sum_{m=2}^M (c_m - c_{m-1})^2 \right) / (M-1) \quad (14)$$

where M is the total number of frames and c_m is a vector containing the cepstral coefficients for frame m .

The results are presented in Table 1. It can be seen that the KF produces the smoothest monophthong estimates. The MSD value for the actual diphthong data was found to be 0.0459. Thus, the UKF estimates most accurately reflect the smoothness of the real trajectories.

	KFLPC MSE	UKFLSF MSE	DAP MSE	KFLPC MSD	UKFLSF MSD	DAP MSD
Mono	0.3568	0.4152	0.4137	0.0063	0.0129	0.0224
Diph	0.3969	0.3546	0.4385	0.038	0.0424	0.0608

Table 1: MSE and MSD of the truncated cepstral distance for UKFLSF, KFLPC and DAP modeling.

In Table 2 we present a comparison between our unscented Kalman filtering of LSFs approach and the recently proposed Direct LSF (DLSF) estimation algorithm [13]. DLSF directly adapts a series of cascaded second order systems for estimation and has been shown to outperform other cascaded recursive least squares (CRLS) methods and the autocorrelation method. This experiment involved estimating the LSF values of a known AR process excited by Gaussian noise. Analysis was performed over 100 non-overlapping segments of the generated signal, each of length 240 samples. The mean square error was calculated for each segment and the mean across all segments is reported in Table 2. It can be seen that UKFLSF outperforms DLSF for each LSF value.

Method	UKFLSF	DLSF
LSF/2 π	M.S.E (10^{-3})	M.S.E (10^{-3})
0.0445	0.00066	0.0228
0.0671	0.00226	0.0404
0.1026	0.00076	0.028
0.1241	0.00271	0.0692
0.2276	0.00128	0.058
0.2505	0.00035	0.0185
0.2711	0.00300	0.1517
0.3088	0.00044	0.0354
0.3219	0.00172	0.1307
0.3763	0.01545	0.0654

Table 2: UKFLSF compared with DLSF.

4. Conclusions & Future Work

The unscented Kalman filtering approach to LSF estimation has been shown to be comparable to, or better than, other methods of tracking LP representations. The UKFLSF technique has been found to perform the task of tracking non-stationary filters, i.e. diphthongs, particularly well. In addition, it yields parameters that are guaranteed to produce stable filters.

In the future, we would like to improve the robustness of our current technique. The LP model assumes the vocal tract to be a hollow tube open at one end; however, during the glottal open-phase the sub-glottal cavity breaks this model. We wish to apply the UKF technique to closed-phase analysis of real speech, similar to that of [2], as this is where accuracy and smoothness is most important. We also wish to extend our investigations to non-vowel sounds such as nasals.

5. Acknowledgements

Andrew Errity is supported by a scholarship from the Irish Research Council for Science, Engineering and Technology.

6. References

- [1] F. Itakura, "Line spectrum representation of linear predictive coefficients," *J. Acoustic Soc. America*, vol. 57(1), p. 535, 1975.
- [2] J. McKenna and S. Isard, "Tailoring Kalman filtering towards speaker characterisation," in *Proc. of Eurospeech '99*, 1999, vol. 6, pp. 2793-2796.
- [3] J. Vepa and S. King, "Kalman-filter based Joint Cost for Unit-selection Speech Synthesis," in *Proc. of Eurospeech 2003*, 2003.
- [4] G. Rigoll, "A new algorithm for estimation of formant trajectories directly from the speech signal based on extended Kalman filtering," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, 1986.
- [5] S. J. Julier and J. K. Uhlmann, "A New Extension of the Kalman filter to nonlinear systems," in *Proc. of AeroSense: The 11th Int. Symp. On Aerospace/Defence Sensing, Simulation and Controls*, 1997, pp. 153-158.
- [6] E. A. Wan and R. van der Merwe, "The Unscented Kalman Filter for Nonlinear Estimation," in *Proc. of IEEE Symp. 2000 (AS-SPCC)*, 2000.
- [7] S. Gannot and M. Moonen, "Sequential-Joint Estimation of Signal and Parameters Using the Unscented Kalman Filter with Application to Single- and Multi-Microphone Speech Enhancement," in *Internal report, K.U. Leuven*, 2001.
- [8] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3(4), pp. 253-264, 1982.
- [9] R. van der Merwe and E. A. Wan, "The square-root unscented Kalman filter for state and parameter-estimation," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, 2001, pp. 3461-3464.
- [10] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, pp. 1445-1450, 1965.
- [11] B. Yegnanarayana, C. d'Allesandro and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. on Speech and Audio Processing*, vol. 6(1), pp. 1-11, 1998.
- [12] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, pp. 411-423, 1991.
- [13] V. Namburu, "Speech Coder Using Line Spectral Frequencies of Cascaded Second Order Predictors," M.S. Thesis, Virginia Tech, 2001.