LF model based glottal source parameter estimation by extended Kalman filtering

Haoxuan Li, Ronan Scaife, Darragh O'Brien

Speech Research Group, Rince

Dublin City University, Glasnevin, Dublin 9

Email: haoxuan.li3@mail.dcu.ie, ronan.scaife@dcu.ie, dobrien@computing.dcu.ie

Abstract — **A new algorithm for glottal source parameter estimation of voiced speech based on the Liljencrants-Fant (LF) model is presented in this work. Each pitch period of the inverse filtered glottal flow derivative is divided into two phases according to the glottal closing instant and an extended Kalman filter is iteratively applied to estimate the shape controlling parameters for both phases. By searching the minimal mean square error between the reconstructed LF pulse and the original signal, an optimal set of estimates can be obtained. Preliminary experimental results show that the proposed algorithm is effective for a wide range of LF parameters for different voice qualities with different noise levels, and accuracy especially for estimation of return phase parameters compares better than standard time-domain fitting methods while requiring a significantly lower computational load.**

Keywords – **LF model, glottal source parameterisation, extended Kalman filter**

I INTRODUCTION

Glottal source modelling is an important topic in the area of speech signal processing and has been investigated over several decades [1, 2, and 3]. An accurate glottal source model can improve the naturalness of synthetic speech [4, 5] and the model parameters can be modified to implement speech transformation. One widely used glottal source model is the Lijencrants-Fant (LF) model [1], which is a four-parameter model suitable for voiced speech source modelling.

Much research has been done to fit the LF model to the inverse filtered glottal flow derivative waveforms to extract the model parameters. The algorithms proposed in [6, 7] are time-domain based methods. In such algorithms, generally one set of initial values of the model parameters is at first obtained, afterwards these parameters are re-estimated by minimising the non-linear least square error between the constructed LF model pulses and the inverse filtered residual signals. In [8], two transformed LF model parameters are calculated from the inverse filtered glottal flow, and the last parameter is estimated by choosing the value from a reasonable range to find the closest match between the spectra of the LF pulses to the source signal. In [9], the author proposed a purely frequency-domain based method. A code-book is built for the H1-H2 (first harmonic minus second harmonic) value and a large number of LF model parameter variations, and the initial estimates can be obtained by searching the code-book to find the closest match to the target signal spectrum. Subsequently an optimisation procedure is applied to refine the estimates. A final parameter is adjusted to minimise the error between the two spectra in higher frequencies. It is said to be robust to phase distortions.

In the work presented here, a new time-domain based LF model fitting algorithm is introduced. Instead of trying to extract the four typical parameters of the LF-model, two shape controlling parameters in the model equations are estimated directly by extended Kalman filtering (EKF). Subsequently the two parameters can be used for re-constructing the LF model. Compared to standard time-domain fitting methods which are based on the non-linear least square error criterion [10], the proposed algorithm offers a more flexible way to re-generate the LF pulses, also because of the fast convergence property of the Kalman Filtering technique, the computational load is lower without losing accuracy.

This paper is structured as follows: In Section II the background for the LF-model and extended Kalman filter is presented. Section III describes the implementation of the new algorithm. Experimental results are presented in Section IV to demonstrate the validity of the algorithm while conclusions are made in Section V.

II BACKGROUND

a) The LF-model

The LF model is a four-parameter model used for representing the glottal flow derivative (GFD) [1]. Typically the four parameters are three time points

t_e, t_p, t_a and one amplitude parameter E_e. If the start point of the cycle t_o is set to 0, and the end of the cycle is t_c, the time domain LF model can be constructed by the equation (1):

$$E(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega_g t), & 0 \le t \le t_e \\ -\frac{E_e}{\epsilon t_a}\left[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}\right], & t_e < t \le t_c \end{cases} \quad (1)$$

where t_e is the glottal closing instant, t_a is related to the return phase of the model, t_p is the positive peak of the glottal flow and generally it is the first zero-crossing point before t_e in the glottal flow derivative. E_0 and E_e are the positive and negative peak values of the derivative function, α, ω_g, and ε are the parameters controlling the shape of the model. A typical example can be seen in Fig. 1.

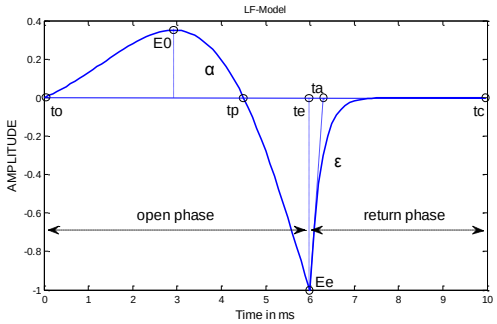

Fig. 1: A typical LF model pulse

E_e as the negative peak of the pitch period can be identified relatively easily, so it is generally the first parameter for estimation, subsequently the timing point t_e can be found. For the other parameters, they are correlated as:

$$\int_0^{t_c} e(t)dt = 0,$$

$$E_0 = -\frac{E_e}{e^{\alpha t_e}\sin(\omega_g t_e)},$$

$$\omega_g = \frac{\pi}{t_p},$$

$$\varepsilon N_a = 1 - e^{-\varepsilon(t_c - t_e)}. \quad (2)$$

Based on these correlations, the commonly used time domain LF model fitting approach is to first estimate E_e, t_e, t_p and t_a, next calculate the amplitude parameter E_0 and the shape controlling parameters α, ω_g, ε. The LF pulse can be generated eventually by equation (1).

A more convenient approach is to directly estimate the shape controlling parameters. It can be observed in (1) that the two parameters α and ε are independent, α controls the open phase while ε is responsible for the return phase of the model (see Fig. 1, the 'return phase' used here including the generally defined return phase and closed phase). In terms of the state-space theory of Kalman filtering, if α and ε are the process state vectors to be estimated, the inverse filtered glottal derivative signal can be used as the measurement of non-linear functions related to track these parameters.

b) The Extended Kalman Filter (EKF)

The Extended Kalman Filter [11] is applicable when the relationship between the process to be estimated and the measurement is non-linear. As with the basic KF, EKF makes use of past measurements to produce an a priori estimate, subsequently current measurement is used to update and generate the posteriori estimate.

The process model and the measurement model are given by:

$$x_k = f(x_{k-1}, k) + w_k,$$

$$z_k = h(x_k, k) + v_k, \quad (3)$$

where x_k is the state vector of the process model at step k. z_k is the measurement, w_k and v_k are random variables representing the process and measurement noise with Gaussian distribution $p(w) = N(0, Q)$ and $p(v) = N(0, R)$. At last f and h are non-linear functions controlling the process.

The EKF time update equations can be expressed by the following two equations:

$$\hat{x}_k^- = f(\hat{x}_{k-1}, k),$$

$$P_k^- = F_k P_{k-1} F_k^T + Q, \quad (4)$$

where $\hat{x}_k^-$ is the a priori estimate at step k, $\hat{x}_{k-1}$ is the posteriori estimate at step $k-1$, P_k^- and P_{k-1} are the related error variances. F_k is the partial derivative function of f with respect to x which is $F_k = \frac{\partial f}{\partial x}(\hat{x}_{k-1}, k)$.

Subsequently the EKF measurement update equations are:

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R)^{-1},$$

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - h(\hat{x}_k^-, k)),$$

$$P_k = (I - K_k H_k) P_k^-, \quad (5)$$

where K_k is the Kalman gain, H_k is the partial derivative function of h with respect to x which is $H_k = \frac{\partial h}{\partial x}(\hat{x}_{k-1}, k)$.

It can be seen that once a set of initial parameters [$\hat{x}_0$, P_0, Q and R] are given, the EKF runs iteratively, and eventually the optimal estimates for the state vector of the process model can be obtained, although sufficient data is necessary for the convergence of the estimation process.

III IMPLEMENTATION

This work is mainly focused on the estimation of glottal source shape controlling parameters, so it is assumed that the glottal opening instants (GOIs) and glottal closing instants (GCIs) are already known, (e.g. [12] introduced an automatic algorithm for identifying GOIs and GCIs), subsequently the inverse filtered glottal flow derivative can be divided

into single pitch periods by GOIs and each of them will be analysed individually.

a) The LF-model Equation Re-written

It is convenient to convert the LF-model timing parameters to a ratio format for discrete time series. For a single cycle, if the pitch period is T_0, the start point t_o is set to 0, the three LF timing parameters can be converted to: $T_e = t_e/T_0$, $T_p = t_p/T_0$, $T_a = t_a/T_0$, and because t_c is the end of this cycle, T_c is set to 1. If there are N samples in this cycle, together with equation (2), equation (1) can be re-written as:

$$E(k) = \begin{cases} -\dfrac{E_e}{e^{\alpha T_e}\sin\left(\frac{\pi}{T_p}\cdot T_e\right)} e^{\frac{\alpha k}{N}} \sin\left(\frac{\pi}{T_p}\cdot\frac{k}{N}\right), \\ \qquad\qquad 0 \le k \le T_e N \\ -\dfrac{E_e}{\varepsilon T_a}\left[e^{-\varepsilon\left(\frac{k}{N}-T_e\right)} - e^{-\varepsilon(1-T_e)}\right], \\ \qquad\qquad T_e N < k \le N \end{cases} \tag{6}$$

where k is the k^{th} sample of the sequence.

b) Estimating α by EKF

According to (6), one glottal cycle is separated into an open phase, where $E_o = E(0, T_e N)$, and a return phase, where $E_r = E(T_e N, N)$. One EKF is applied to the open phase to estimate α, the other EKF is used for estimating ε in return phase.

Considering α is the single constant to be estimated, the process model and the measurement model can be expressed by:

$$\alpha_k = \alpha_{k-1},$$
$$E_k = h_o(\alpha_k, k) + v_k, \tag{7}$$

where E_k is the k^{th} sample of E_o , v_k is the measurement noise with Gaussian distributions $p(v) = N(0, R_o)$, and h_o is a non-linear function defined in the upper equation of (6).

Based on (4) the related EKF time update equations are as follows:

$$\hat{\alpha}_k^- = \hat{\alpha}_{k-1},$$
$$P_k^- = P_{k-1}, \tag{8}$$

and the measurement update equations are:

$$K_k = P_k^- H_o(\hat{\alpha}_k^-)(H_o(\hat{\alpha}_k^-)P_k^- H_o(\hat{\alpha}_k^-) + R_o)^{-1},$$
$$\hat{\alpha}_k = \hat{\alpha}_k^- + K_k(E_k - h_o(\hat{\alpha}_k^-, k)),$$
$$P_k = (1 - K_k H_o(\hat{\alpha}_k^-))P_k^-, \tag{9}$$

where $H_o(\hat{\alpha}_k^-) = \frac{\partial h_o}{\partial \alpha}(\hat{\alpha}_k^-, k)$.

For the running of EKF, the next step is to set the initial values. Experiments show that $R_o = 0.01$, $P_0 = 1$ is a reasonable choice. However because of the limited number of samples, the initial estimate of α_0 is more important. Instead of making use of an additional procedure such as the EM algorithm [13] to iteratively re-estimate α_0, a reasonable range (0-100, increased by 1 for each iteration) of the values for α_0 is used currently. Subsequently the estimated α for each α_0 is used to re-construct the open phase of the LF-model and the mean square error (MSE) to the original signal is calculated. The best estimate of α is the one having the minimal mean square error. Fig. 2 shows an example of reconstructed LF-model open phase according to different α_0 values. It can be observed that when $\alpha_0 = 15$, the estimate of α by EKF is the closest match and provides the initial estimate.

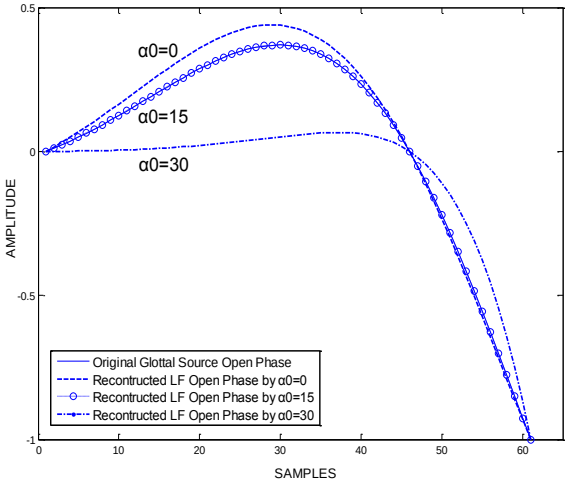

Fig. 2: Reconstructed LF-model open phase according to different $\boldsymbol{\alpha_0}$ values.

It can be observed in equation (6) that for estimating α, T_p should be known, which means the timing parameter t_p must be extracted. The initial estimate of t_p is to find the first zero-crossing point before the glottal closing instant t_e [6]. A reasonable range $[T_p - 5\%, T_p + 5\%]$ (increased by 1% each time) is used to refine the estimate with EKF iteratively. Eventually the two parameters t_p and α_0 which give the minimal MSE between the reconstructed and the original signal for all iterations are the optimal estimates.

c) Estimating ε by EKF

Estimation of the return phase shape controlling parameter ε is relatively simple because there is no need to find additional timing parameters. The process model and measurement model are given by:

$$\varepsilon_k = \varepsilon_{k-1},$$
$$E_k = h_r(\varepsilon_k, k) + v_k, \tag{10}$$

E_k is the k^{th} sample of E_r, v_k is the measurement noise with Gaussian distribution $p(v) = N(0, R_r)$, and h_r is the non-linear function defined by the lower equation in (6).

The EKF time update equations are:

$$\hat{\varepsilon}_k^- = \hat{\varepsilon}_{k-1},$$
$$P_k^- = P_{k-1}, \tag{11}$$

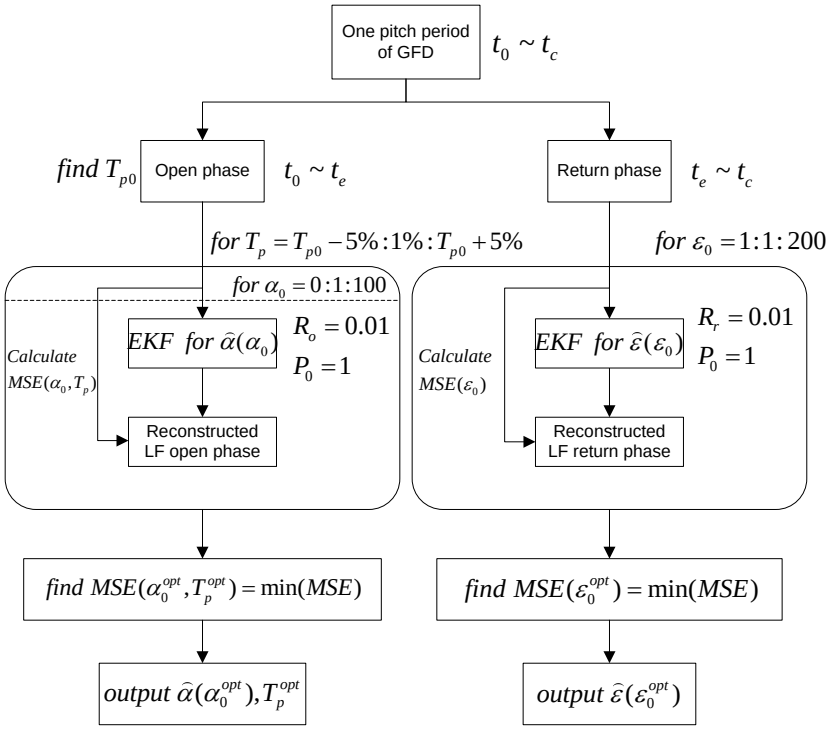

Fig. 3: Flow chart for the time-domain shape controlling parameters estimation algorithm

and the three measurement update equations can be written as:

$$K_k = P_k^- H_r(\hat{\varepsilon}_k^-)(H_r(\hat{\varepsilon}_k^-)P_k^- H_r(\hat{\varepsilon}_k^-) + R_r)^{-1},$$

$$\hat{\varepsilon}_k = \hat{\varepsilon}_k^- + K_k(E_k - h_r(\hat{\varepsilon}_k^-, k)),$$

$$P_k = (1 - K_k H_r(\hat{\varepsilon}_k^-))P_k^-, \qquad (12)$$

with the definition of $H_r(\hat{\varepsilon}_k^-) = \frac{\partial h_r}{\partial \varepsilon}(\hat{\varepsilon}_k^-, k)$.

As mentioned earlier for the initial values of EKF, $R_r = 0.01$, $P_0 = 1$ is a good choice. Also because $\varepsilon \approx 1/T_a$, and generally T_a changes from 1% to 20% [8], the range of values for ε_0 could be (1-200, increased by 1 for each iteration) for the purpose of handling most situations. The optimal estimate of ε is obtained by fitting the re-built LF return phase signal to the original signal iteratively to find the one which gives the minimal mean square error (MSE).

Fig. 3 shows a flow chart which describes the complete estimation procedure.

IV Evaluation

There are two parts to the evaluation and preliminary results are presented. Firstly several sets of synthetic LF model pulses for different voice qualities and noise levels are generated for testing the accuracy of the proposed algorithm. Subsequently, the proposed algorithm is integrated into a speech analysis toolbox called aparat [10] and compared to its original LF fitting algorithm. Three sets of synthetic vowels are used for this experiment.

a) Accuracy Test for Synthetic LF pulses

Three sets of LF model parameters including modal, vocal fry and breathy voice quality shown in [14] were used to generate the LF pulses. Table 1 shows these sets and the related true values for α and ε calculated by equations (1) and (2). Afterwards three sets of 10 pitch period signals were obtained by concatenating identical pulses. To each signal was added Gaussian white noise with SNR 45 dB (moderate noise level) and 30 dB (high noise level) giving 6 sets of synthetic glottal flow derivative signals for testing. The proposed algorithm was applied and the mean error rates of the estimated results are shown in Table 2. Fig. 4 shows an example of the reconstructed LF pulses fitted to original ones of Modal voices with 30 dB SNR, for clarity only the first two pitch cycles are shown.

Table 1: LF model parameters for three voices

Voice	T_p(%)	T_e(%)	T_a(%)	α	ε
Modal	45.66	57.50	0.91	6.0240	109.8901
Vocal fry	18.99	25.14	0.83	9.1815	120.4819
Breathy	52.89	75.75	8.19	1.0490	11.4500

It can be observed that the error rates of the estimated shape controlling parameters are reasonably low even for high noise level. For small T_a, the estimate of ε is more accurate. This might be because the GFD signal goes abruptly from the negative peak to zero and lasts till the opening of the next pitch cycle, therefore better estimates can be obtained by EKF. It can be seen from Fig. 4 that the reconstructed LF pulses by the estimated shape controlling parameters are fitted well to the original ones. These results demonstrate the validity of the novel LF fitting algorithm for synthetic LF pulses.

Table 2: Mean error rates for estimated α and ε

Voice	Error Rate (α)	Error Rate (ε)
Modal(45dB)	6.53%	0.09%
Modal(30dB)	6.69%	0.10%
Vocal fry (45dB)	2.42%	0.39%
Vocal fry (30dB)	2.57%	0.79%
Breathy(45dB)	4.52%	3.51%
Breathy(30dB)	6.37%	4.22%

b) Algorithm Comparison

Aparat [10] is a speech analysis tool which by default uses IAIF [15] for extracting glottal flow signals. The proposed algorithm was integrated into aparat: the new LF fitting method is called the extended Kalman filtering method (EKFM) and the original one which is called the standard time-domain method (SDTDM). LF parameters, formant frequencies and bandwidths were taken from [16, 17] to generate three vowel sounds /aa/, /ih/, /uh/ by putting 5 pitch periods of the constructed LF pulses through a formant synthesizer. LF parameters for the three vowels are shown in Table 3.

Table 3: LF parameters used for the three vowels

Vowel	T_p	T_e	T_a	α	ε
/aa/	0.50	0.65	0.050	3.1230	19.9816
/ih/	0.45	0.60	0.075	2.1870	13.2672
/uh/	0.45	0.60	0.005	4.3035	200

The output speech signals were inverse filtered (the IAIF settings were manually adjusted), and the LF model parameters were estimated using both SDTDM and EKFM. Tables 4 and 5 show the mean values of the estimates and the running time for the analysis of the two algorithms.

Table 4: Means of estimates and running time of SDTDM

Vowel	T_p	T_e	T_a	α	ε	R-time
/aa/	0.5068	0.6633	0.0204	3.7820	49.0196	0.96 sec
/ih/	0.5005	0.6569	0.0213	3.7170	46.9484	0.95 sec
/uh/	0.4466	0.5903	0.0198	3.9360	50.5051	0.97 sec

Table 5: Means of estimates and running time of EKFM

Vowel	T_p	T_e	T_a	α	ε	R-time
/aa/	0.5062	0.6543	0.0455	3.3460	21.9670	0.45sec
/ih/	0.5000	0.6500	0.0623	2.8345	15.9918	0.42 sec
/uh/	0.4430	0.5949	0.0054	4.1545	185.185	0.46 sec

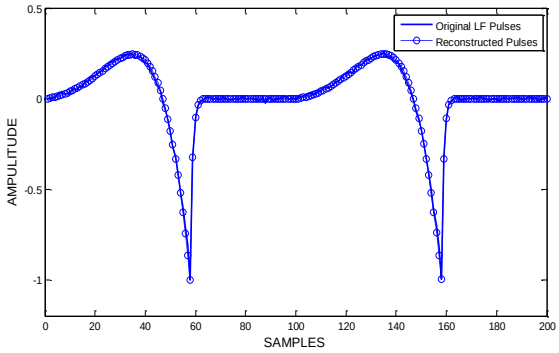

Fig. 4: Pulses fitted to original LF model of Modal Voice (30dB)

It can be observed for the open phase parameters T_p and T_e, the estimates from SDTDM and EKFM are close. Compared to the true values, T_p and T_e for /ih/ by both methods have a shift of about 0.5%, which is because of the inaccurate identification of the glottal opening instants. Compared to SDTDM, EKFM gives more accurate estimates for α. For the return phase parameters T_a and ε of all the three vowels, SDTDM gives similar values of estimates, this might be because T_a is a parameter which changes a lot across different speech signals and cannot be accurately tracked by the non-linear least square error criterion without a proper initial value (initial T_a is set to 0.02 in aparat), sufficient number of iterations and so forth; while estimates for these two parameters by EKFM are more accurate compared to true values, which is because the fast convergence property of EKF and a reasonable range of initial values are used. In addition, by observation the running time of EKFM is more than 50% less compared to SDTDM based on current settings for the initial value ranges and iterations of α_0, ε_0 and T_p. This is a significant improvement of the computational load, although further promotion can be obtained when a more effective procedure to estimate the initial values of EKF is available. In summary, these preliminary experimental results demonstrate that the novel LF model based glottal source parameter estimation algorithm outperforms standard time-domain glottal source parameterisation method for accuracy especially of the return phase parameters, while with an significant improvement in computational load.

V Conclusion and Future Work

A new algorithm to estimate the LF model based time-domain shape controlling parameters of the inverse filtered glottal flow derivative by extended Kalman filtering is described. Each pitch period of GFD is divided into an open phase and a return phase from the glottal closing instant, subsequently EKF is iteratively applied to track the shape controlling parameters for both phases to obtain an

optimal fit. Preliminary experimental results demonstrate that the proposed algorithm is effective for synthetic LF model pulses of different voices and noise levels. In addition, comparison shows that its accuracy, especially for return phase parameters, is better than standard time-domain LF-model fitting algorithm but with a much lower computational load. Therefore, it is believed that this method could be used in applications which require a fast glottal source parameterisation.

Clearly better inverse filtering algorithms will provide more accurate glottal flow signals. Therefore, in future work different source-vocal tract separation techniques will be used to estimate the glottal source for both synthetic and real speech. Subsequently the novel LF fitting algorithm can be applied and the statistical significance of improvements can be tested.

REFFERENCES

[1] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow", STL-QPSR, vol. 4, no. 1985, pp. 1–13, 1985.

[2] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Am, vol. 49, no. 2, pp. 583–590, 1971.

[3] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", the Journal of the Acoustical Society of America, vol. 87, pp. 820, 1990.

[4] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis", in Proc. Of the 6th ISCA Workshop on Speech Synthesis, Germany, 2007.

[5] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis", in 9th Annual Conference of the International Speech Communication Association, 2008.

[6] H. Strik, B. Cranen, and L. Boves, "Fitting a LF -model to inverse filter signals", in ESCA 3rd European Conference on Speech Communication and Technology: EUROSPEECH '93, Berlin , pp. 103–106, 1993.

[7] H. Strik and L. Boves, "Automatic estimation of voice source parameters", in Proceedings International Conference on Spoken Language Processing (ICSLP)'94, pp. 155–158, 1994.

[8] J. C. Kane and C. Gobl, "Automatic parameterisation of the glottal waveform combining time and frequency domain measures", in Proceedings of 6th Maveba International Workshop, 2009 .

[9] J. Kane, M. Kane, and C. Gobl, "A spectral LF model based approach to voice source parameterisation", in Eleventh Annual Conference of the International Speech Communication Association, 2010.

[10] M. Airas, "TKK Aparat: An environment for voice inverse filtering and parameterization, volume 33", Logopedics Phoniatrics Vocology, pp. 49–64, 2008.

[11] G. Welch and G. Bishop, "An introduction to the Kalman filter", University of North Carolina at Chapel Hill, Chapel Hill, NC, vol. 7, no. 1, 1995.

[12] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals", Proceedings of Interspeech 2009.

[13] 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, no. 4, pp. 253–264, 1982.

[14] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization", IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 2, pp. 492–501, 2006.

[15] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Communication, vol. 11, no. 2-3, pp. 10 9–118, 1992.

[16] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of z-transform (ZZT) decomposition of speech for source-tract separation", in Proc. International Conf. Speech, Language Processing, 2004.

[17] O. O. Akande and P. J. Murphy, "Estimation of the vocal tract transfer function with application to glottal wave analysis", Speech Communication, vol. 46, no. 1, pp. 15-36, May, 2005.