AUTOMATIC LF-MODEL FITTING TO THE GLOTTAL SOURCE WAVEFORM BY EXTENDED KALMAN FILTERING

Haoxuan Li, Ronan Scaife and Darragh O'Brien

Speech Research Group, RINCE Institute,
Dublin City University, Glasnevin, Dublin 9, Ireland

ABSTRACT

A new method for automatically fitting the Liljencrants-Fant (LF) model to the time domain waveform of the glottal flow derivative is presented in this paper. By applying an extended Kalman filter (EKF) to track the LF-model shape-controlling parameters and dynamically searching for a globally minimal fitting error, the algorithm can accurately fit the LF-model to the inverse filtered glottal flow derivative. Experimental results show that the method has better performance for both synthetic and real speech signals compared to a standard time-domain LF-model fitting algorithm. By offering a new method to estimate the glottal source LF-model parameters, the proposed algorithm can be utilised in many applications.

Index Terms— LF-model, glottal source, extended Kalman filter

1. INTRODUCTION

The voice source is an important hidden property of human speech signals. Accurate parameterisation of the voice source can be used for many practical applications, such as improving the naturalness of HMM-based speech synthesis [1], speaker identification [2] and voice transformation [3].

Automatic voice source parameterisation generally includes two steps. Firstly, a source and vocal tract separation algorithm [4] is applied to the speech signal to remove the effect of the vocal tract, and the glottal source waveform is obtained. Subsequently, a parametric model, for instance the Liljencrants-Fant (LF) model [5], which is widely used to represent the glottal pulse for voiced speech, is fitted to the extracted glottal source waveform to estimate the glottal source model parameters.

A typical time-domain LF-model fitting algorithm is given in [6]. Once an initial estimate of LF-model parameters has been obtained, a multi-parameter non-linear optimisation procedure is used to minimise the fitting error between the LF-model pulse and the extracted glottal flow derivative to achieve an optimal estimate. However, an inaccurate initial estimate of the glottal opening instant can result in performance degradation of the fitting algorithm: the extracted glottal source parameters can be sub-optimal, because the optimisation procedure may become stuck in a local minimum [7].

This study is an extension of our previous work [8], which showed that the extended Kalman filter [9] can be used to track the LF-model shape-controlling parameters for both the open phase and return phase. In this paper a new dynamic programming procedure is applied to search for the optimal glottal opening instant location for each pitch period. In addition further evaluation has been carried out and is presented here. To demonstrate the validity of the proposed algorithm, it is compared to a standard time-domain LF-model fitting method. Comparisons are made not only for synthetic speech of different voice qualities, but also for real speech segments. Results show our approach performs better than the traditional approach yielding a more accurate fit to the inverse filtered glottal flow derivative.

2. BACKGROUND

2.1. LF-model representation

The LF-model [5] is a four-parameter model of the differentiated glottal flow. If the number of samples of a pitch period is N, and k is the k^{th} sample, a discrete form of the LF-model is given by:

$$r_o(k) = -\frac{E_e}{e^{\alpha T_e}\sin\left(\frac{\pi}{T_p}T_e\right)} e^{\frac{\alpha k}{N}} \sin\left(\frac{\pi}{T_p}\cdot\frac{k}{N}\right) = h_o(\alpha, k),\ 0 \le k \le T_e N$$

$$r_r(k) = -\frac{E_e}{\varepsilon T_a}\left[e^{-\varepsilon\left(\frac{k}{N}-T_e\right)} - e^{-\varepsilon(1-T_e)}\right] = h_r(\varepsilon, k),\ T_e N < k \le N \quad (1)$$

where E_e is the amplitude parameter, T_p, T_e and T_a are the three timing parameters (normalised by pitch period N), T_e is the open quotient, T_p and α affect the asymmetry property of the open phase component r_o, and T_a and ε control the shape of return phase component r_r.

2.2. Shape-controlling parameter tracking by EKF

For a given pitch period of the glottal flow derivative signal, the LF-model parameters are constant. Accordingly, the

© EURASIP, 2012 - ISSN 2076-1465

state-space process model and measurement model for the shape controlling parameters α and ε can be written as:

$$x_k = x_{k-1},$$
$$r_k = h(x_k, k) + v_k, \quad (2)$$

where k is the k^{th} speech sample, x is the constant state parameter standing for α or ε, r is the measurement given by r_o or r_r, h is the related non-linear function h_o or h_r defined in (1), and v is the observation noise with Gaussian distribution $p(v) = N(0, R)$. Accordingly, the EKF time update equations are as follows:

$$\hat{x}_k^- = \hat{x}_{k-1},$$
$$P_k^- = P_{k-1}, \quad (3)$$

where $\hat{x}^-$ and $\hat{x}$ are a priori and a posteriori estimates of x, and P^- and P are the corresponding error covariances. The EKF measurement update equations are given by:

$$K_k = P_k^- H(\hat{x}_k^-)(H(\hat{x}_k^-)P_k^- H(\hat{x}_k^-) + R)^{-1},$$
$$\hat{x}_k = \hat{x}_k^- + K_k(r_k - h(\hat{x}_k^-, k)),$$
$$P_k = (1 - K_k H(\hat{x}_k^-))P_k^-, \quad (4)$$

where K is the Kalman gain and $H(\hat{x}_k^-) = \frac{\partial h}{\partial x}(\hat{x}_k^-, k)$. It can be seen that once an initial set of parameters $[x_0, P_0, R]$ is given, the extended Kalman filter runs recursively to track the true values of α and ε respectively across a single pitch period by using samples of the two phases.

3. ALGORITHM

Firstly an iterative closed phase inverse filtering approach [4] is applied to extract the glottal flow derivative (GFD) signal. Afterwards, the initial glottal opening instants t_0 are obtained from a threshold based procedure [7]. The GFD waveform is divided into individual pitch periods. The new fitting algorithm is applied to each period as shown in Table 1. t_0, t_p and t_e are relative sample numbers.

Step 1: the negative peak point t_e and its absolute amplitude E_e are found by searching the waveform, and the GFD signal is separated into the open phase and return phase.

Step 2: The initial t_p position is obtained by identifying the first zero-crossing point before t_e [6].

Step 3: The optimal fitting of the open phase mainly depends on the values of T_p and T_e, which are calculated by $T_p = (t_p - t_0)/N$, $T_e = (t_e - t_0)/N$. We set the dynamic range of t_0 from 1 to the point which is $0.15N$ samples (a reasonably small open quotient value) before t_e. To locate the optimal t_0, a rectangular window across t_0 to t_e is used to extract the glottal open phase. Subsequently the windowed GFD open phase is used by the EKF to track the open phase shape-controlling parameter α and calculate the mean squared fitting error. Unlike the linear Kalman filter algorithm, the EKF has to be accurately initialised to ensure the obtained linearised models are valid. Therefore we choose to use multiple initial values of α for the EKF (100 α_0 used in this work, experiments shown that using a larger number will not significantly improve the accuracy but increase the computational load), and the one giving the minimal mean squared fitting error ($MMSE_f$) is taken as the optimal selection. Each time t_0 changes, the $MMSE_f$ obtained from the corresponding EKF operation is stored, and a global $MMSE_f$ is found after all iterations. Subsequently, an optimal t_0 instant is located and T_e and an initial T_p is calculated as the output.

Table 1. Proposed new time-domain LF-model fitting algorithm

For each pitch period of GFD signal $r[k]$ (k=1:N) **do**

1. Find negative peak t_e and its amplitude E_e
2. Find t_{p0} which is the first zero-crossing point before t_e
3. **For** $t_0 = 1: t_e - 15\%N$ **do**
   - GFD open phase $r_o = r[k]$ (k=t_0: t_e)
   - $T_{p0} = (t_{p0} - t_0)/N$
   - $T_e = (t_e - t_0)/N$
   - EKF for α with multiple initial values $\alpha_0 = 0: 1: 100$
   - Find and store minimal squared fitting error $MMSE_f$
   - Find optimal t_{0opt} which has a global $MMSE_f$
   - Calculate T_{p0}, T_e by t_{0opt}, set $r_o = r[k]$ (k=t_{0opt}: t_e)
4. **For** $T_p = T_{p0} - 5\%: T_{p0} + 5\%$ **do**
   - EKF for α as in Step 3
   - Find optimal T_p which has a global $MMSE_f$
   - Output T_p
5. GFD return phase $r_r = r[k]$ (k=t_e: N)
   - EKF for ε with multiple initial values $\varepsilon_0 = 1: 1: 200$
   - Find optimal ε_{0opt} which has a $MMSE_f$
   - Calculate and output T_a by ε_{0opt}

Step 4: A similar procedure to Step 3 is used to refine the estimate of T_p: a reasonable range $T_{p0} \pm 5\%$ is applied to the EKF to find the optimal T_p.

Step 5: To fit the return phase, t_0-1 zeros (associated with the closed phase of the previous pitch cycle and not used in the open phase fitting procedure) are appended to the current pitch period of the GFD signal to ensure there is a sufficient number of samples for the EKF. Subsequently an optimal estimate of the return phase controlling parameter ε is obtained by using multiple initial values of ε (a number of 200 ε_0 used in this work) for initialising the EKF tracking procedure and searching for the global MMSE. The

return phase timing parameter T_a is calculated from the LF-model constraint [5] $T_a = 1 - e^{(-\hat{\varepsilon}(1-T_e))}/\hat{\varepsilon}$.

4. EVALUATION

To evaluate the newly proposed time-domain LF-model fitting method (NTDF), it was compared with a standard time-domain LF-model fitting algorithm (STDF). The latter and our new method were applied to both synthetic and real speech data. The evaluation results are presented below.

4.1. Synthetic Speech

Three sets of LF-model parameters of different voice qualities [10] were used to generate the glottal source signals, and the corresponding glottal pulse trains were obtained by concatenating ten identical pitch periods. Afterwards the three sets of LF-model pulse trains were passed through three all-pole vocal tract filters modeling three vowel sounds (formant frequencies and bandwidths were taken from [11]), and a total of nine sustained synthetic speech segments were created. In addition, for breathy voice, simulated noise of 30dB SNR was added to mimic real breathy speech quality. All vowel segments were inverse filtered by iterative closed phase inverse filtering [4] to extract the glottal flow derivatives. The GFD signals were divided into individual pitch periods by the initial estimation of glottal opening points. Subsequently, the proposed fitting approach (NTDF) and the standard time-domain LF-model fitting algorithm (STDF) were applied respectively to all pitch cycles of the GFD signals. The Root-Mean-Square errors (RMSE) for the estimated LF-model timing parameters for both algorithms with respect to the true values were calculated, and the results are presented in Fig. 1. It is observed that for modal and vocal fry voice qualities the RMSE scores are consistently lower for the proposed fitting algorithm compared to STDF. For breathy voice the results are less clear. The estimated T_p and T_e for breathy vowels /IH/, /UH/ by NTDF are more accurate, however for T_a the standard fitting method performs better. This may be explained by imperfect inverse filtering caused by short duration of the closed phase for breathy voices, but requires further investigation. In addition, the running time for NTDF is about 1/3 faster than STDF for the current algorithm configuration and further improvement of the computational complexity can be obtained by choosing more appropriate initial values.

Overall these experimental results demonstrate the validity of the proposed LF-model fitting algorithm to estimate glottal LF-model parameters for a wide range of synthetic speech signals, and it is superior to the standard time-domain fitting method in most cases.

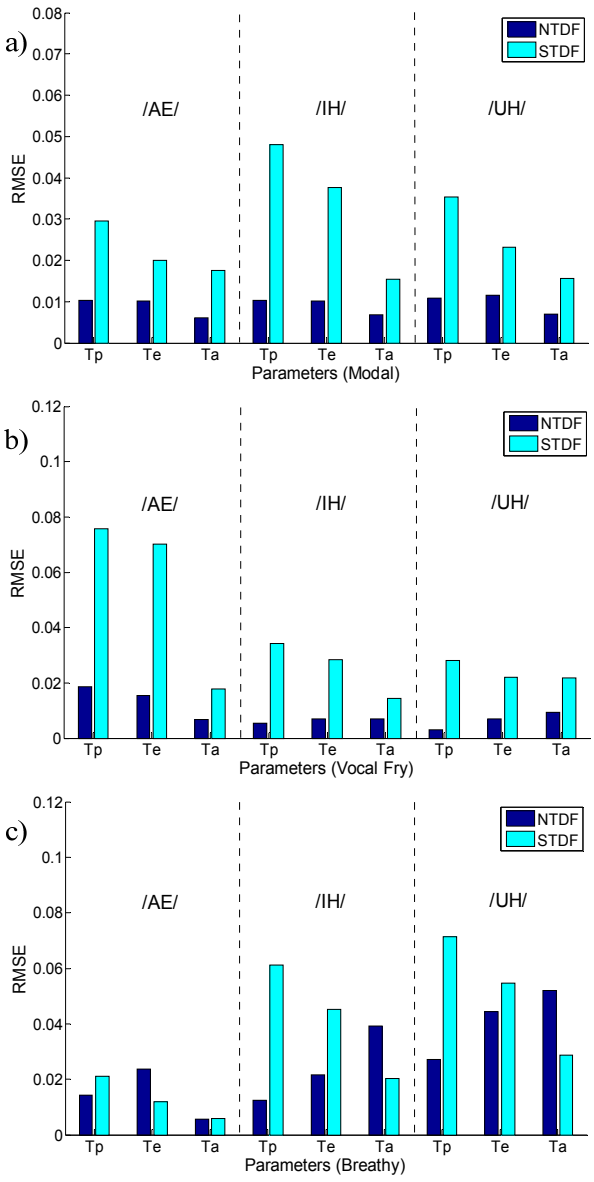

Figure 1. RMSE scores of estimated LF-model parameters for a) modal voice, b) vocal fry voice and c) breathy voice

4.2. Real Speech

Two segments of real speech were extracted from the CMU-ARCTIC database [12] for speakers *bdl* (a male voice) and *slt* (a female voice). Both segments were inverse filtered to extract the glottal flow derivative signals. Afterwards the two LF-model fitting algorithms were applied. The original speech waveforms, the GFD waveforms and the fitted LF-model pulses are presented in Figs. 2 and 3. A single pitch period of GFD and fitted LF-model waveforms are shown in Fig. 4. In the absence of a priori knowledge of the glottal source component for real speech, it is difficult to measure

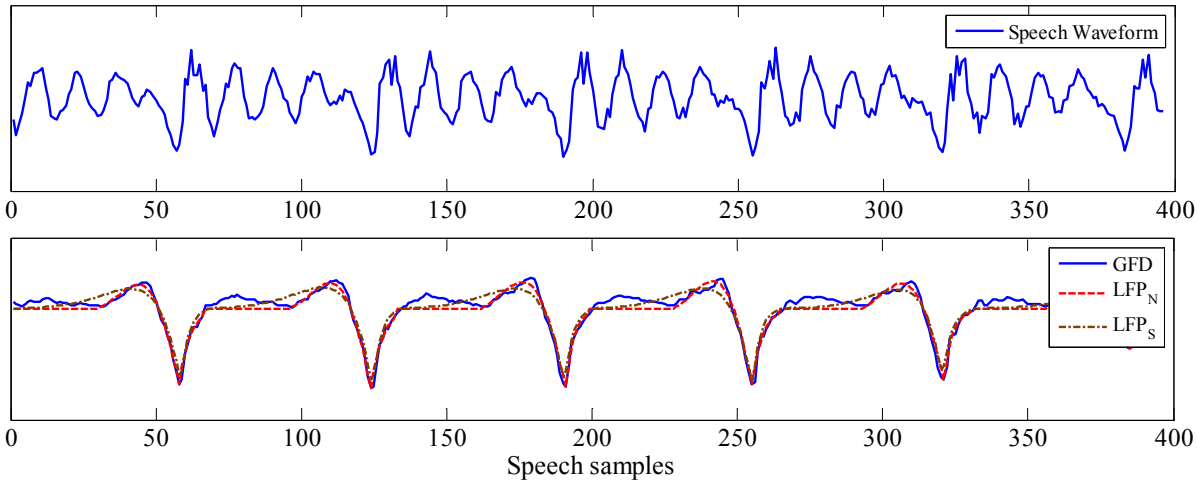

Figure 2. Top: male speech waveform, Bottom: GFD signals and fitted LF-model pulses (LFP) from the two algorithms

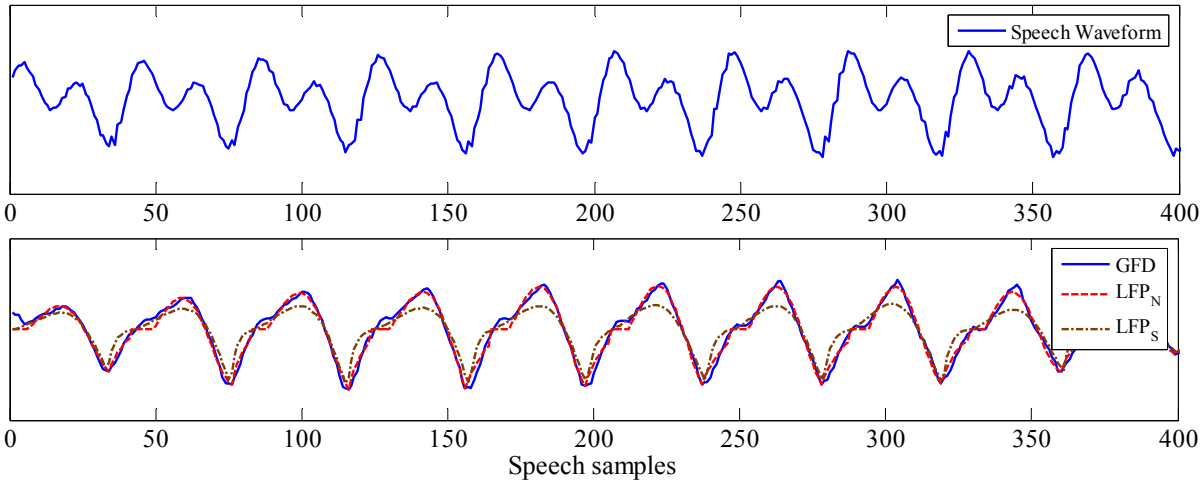

Figure 3. Top: female speech waveform, Bottom: GFD signals and fitted LF-model pulses (LFP) from the two algorithms

the accuracy of the estimated source parameters. Instead we compare the goodness of fit to the estimated GFD signals of the two algorithms. Therefore the mean squared error ($MSE = E[(r - r_{LF})^2]$) between the estimated GFD signals r and the reconstructed LF-model pulses r_{LF} across the full speech segments were calculated and the results are presented in Table 2. It can be observed from the waveforms and the MSE scores that for both male and female speech segments, the proposed algorithm outperforms standard time-domain fitting approach by generating smaller MSE scores. For NTDF, the male subject has a larger MSE than female is because of the ripples appearing in closed phases.

Table 2. MSE scores for real speech segments from two automatic time-domain LF model fitting algorithms

	BDL	SLT
NTDF	0.1851	0.0670
STDF	0.3448	0.4732

5. CONCLUSION

A new automatic time-domain method to fit the LF-model to the inverse filtered glottal flow derivative signal for voice source parameterisation is presented in this paper. An extended Kalman filter is used to track the two shape-controlling parameters with dynamic searching procedures to find a globally optimal fit of the LF-model pulse to the differentiated glottal flow signal, such that the corresponding LF-model timing parameters can be accurately extracted. Comparisons were made between the proposed method and a standard time-domain algorithm by applying them to both synthetic speech and real speech signals. Results demonstrate the effectiveness of the new fitting algorithm. For synthetic speech the estimated LF-model parameters are more accurate in most cases, and for real speech the reconstructed LF-model pulses are better fitted to the glottal flow derivative signals.

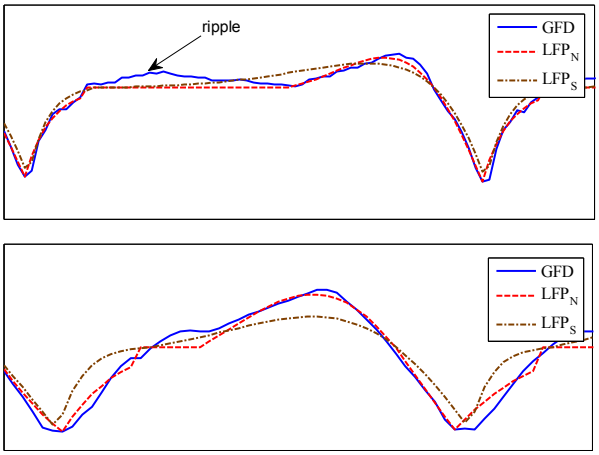

Figure 4. Single pitch period of GFD and fitted LF-model waveforms for top: male and bottom: female

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of Haoxuan Li by the China Scholarship Council and the European Regional Development Fund (ERDF) in carrying out the work presented in this paper.

REFERENCES

[1] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source", in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4704-4707.

[2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification", *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 5, pp. 569-586, 1999.

[3] M. Tooher, I. Yanushevskaya, and C. Gobl, "Transformation of LF parameters for speech synthesis of emotion: regression trees", in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, ISCA, 2008.

[4] E. Moore and M. Clements, "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information", in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 101-104, 2004.

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

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

[8] H. Li, R. Scaife and D. O'Brien, "LF model based glottal source parameter estimation by extended Kalman filtering", in *Proceedings of the 22nd IET Irish Signals and Systems Conference,* 2011.

[9] G. Welch and G. Bishop, "An introduction to the Kalman filter", University of North Carolina at Chapel Hill, Chapel Hill, NC, 1995.

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

[11] 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.

[12] "CMU-ARCTIC speech synthesis databases", available at http://festvox.org/cmu-arctic/index.html.