Comparison of Time- and Frequency-domain Based LF-model Fitting Methods for Voice Source Parametrisation

Haoxuan Li, Darragh O'Brien and Ronan Scaife

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

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

Abstract — **The Liljencrants-Fant (LF) model is used to capture the shape parameters from the voice source. In this paper, two LF-model fitting approaches (one time-domain, one frequency-domain) are presented and compared by applying each to artificial and real speech source signals. Experimental results demonstrate that in most cases the time-domain method is superior to the frequency-domain based algorithm. By assessing approaches for estimating the LF-model parameters from a glottal source signal, this paper makes a contribution to the investigation of voice source parametrisation.**

Keywords — **LF-model, Voice source parametrisation**

I Introduction

Modelling of the voice source has been for many years an active research topic of digital speech signal processing. It has been shown that the Liljencrants-Fant (LF) model [1] can be used to simulate the glottal source in a HMM-based speech synthesiser to improve the naturalness of synthetic speech [2, 3]. By studying both voice source and vocal tract filter, it is possible to detect pathological voice [4]. The voice source may be used to investigate the characteristics of emotional speech [5]. In addition, voice source parameters can be used for speaker identification [6].

To extract the voice source, the first step is to remove the vocal tract contribution from the speech signal. A glottal inverse filtering algorithm [7, 8] can be used for this purpose. Once the glottal waveform is obtained, glottal model fitting methods can be applied to estimate the voice source parameters. Because of its wide use in representing the glottal flow for voiced speech, many voice source model fitting algorithms are based on the LF-model [1].

Generally, LF-model fitting methods can be categorised into two groups: time-domain (TD) and frequency-domain (FD) based methods. For TD approaches [9, 10], one or more multi-parameter optimisation procedures are applied to adjust the LF-model parameters to minimise the fitting errors between the LF-model pulses and the extracted glottal flow derivative (GFD). The disadvantage of TD based LF fitting algorithms is their poor glottal source parameter estimation for phase-distorted speech. In addition, because of source-tract interaction and limitations of inverse filtering techniques, in most cases the vocal tract effect cannot be completely removed. This may result in inaccurate estimation of the glottal opening instant, and the glottal source parameters. Although a low-pass filter can be used to remove ripples of incompletely cancelled formants and additive noise [11], the changed shape of the glottal waveform may lead to inaccurate source parameter estimation.

For frequency-domain based methods, firstly the glottal flow spectrum is calculated from the inverse filtered glottal flow signal. Subsequently, the glottal source parameters are estimated by minimising the spectral distance between the real glottal flow and the LF-model spectra. A typical FD based fitting method is given in [12], which initialises the LF parameter estimates from a codebook of differences between the magnitude of the first two

Authorized licensed use limited to: DUBLIN CITY UNIVERSITY. Downloaded on April 21,2021 at 13:20:59 UTC from IEEE Xplore. Restrictions apply.

harmonics (H1*-H2*) of the source spectrum. Initial LF parameter estimates are then refined in a two-step optimisation procedure. There are also limitations with this method. It has been shown in [13] that multiple sets of LF-model parameters can generate very similar H1*-H2* value, therefore the spectral optimisation procedure may become stuck in a local minimum caused by poor initialisation, and correspondingly inaccurate estimates will be obtained.

To the best of our knowledge, no previous work has focused specifically on the comparison of time- and frequency-domain based LF-model fitting methods. By presenting evaluation results for both artificial and real voice data, this paper compares a TD-LF model fitting approach [14] to a spectral fitting algorithm which is similar to [12]. The proposed time-domain fitting method uses the extended Kalman filter and dynamic programming to improve its accuracy of the estimates. Also, the spectral fitting approach utilises a codebook of the amplitudes of the first six harmonics rather than H1*-H2* for more robust initialisation. This study aims to give researchers working in this area an overview of the performance of the two fitting approaches for both artificial and real voice source data.

The structure of this paper is as follows: Section II gives the background for the LF-model and the two fitting methods. Section III presents the experimental and evaluation results. Section IV summarises the study and draws conclusions.

II Background

a) LF-model Representation in Discrete Form

The LF-model [1] is a four-parameter model of the differentiated glottal flow. If the number of samples in a pitch period is N, and k is the k^{th} sample, a discrete form of the LF-model is given by:

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

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

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 the return phase component r_r.

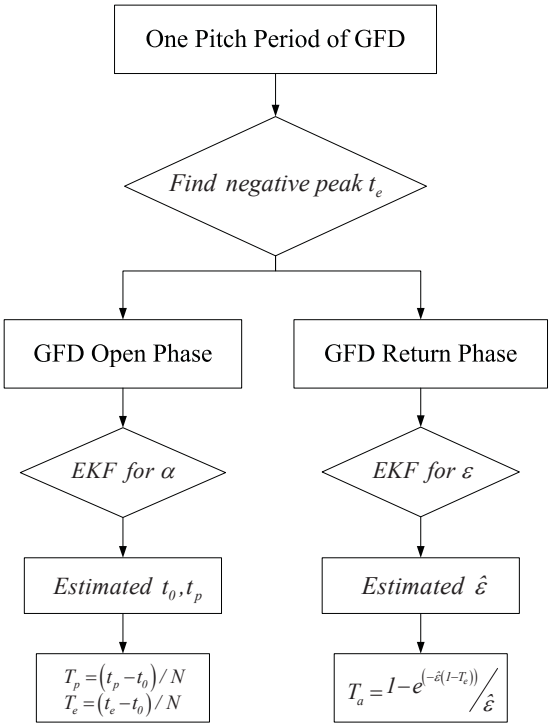

Fig. 1: Flow chart of the Time-domain LF-fitting algorithm.

b) Time-domain LF-fitting by EKF

In [14] we have shown that the extended Kalman filter can be used to track the two LF-model shape-controlling parameters. For a given pitch period of the glottal flow derivative signal, the LF-model parameters are constant. Accordingly, the state-space process model and measurement model for the shape controlling parameters α and ε can be written as:

$$\begin{aligned} x_k &= x_{k-1}, \\ r_k &= h(x_k, k) + v_k \end{aligned} \tag{2}$$

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 one of the related non-linear functions 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:

$$\begin{aligned} \widehat{x}_k^- &= \widehat{x}_{k-1}^-, \\ P_k^- &= P_{k-1}^- \end{aligned} \tag{3}$$

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

Authorized licensed use limited to: DUBLIN CITY UNIVERSITY. Downloaded on April 21,2021 at 13:20:59 UTC from IEEE Xplore. Restrictions apply.

equations are given by:

$$\begin{aligned} K_k &= P_k^- H(\widehat{x}_k^-)(H(\widehat{x}_k^-)P_k^- H(\widehat{x}_k^-) + R)^{-1}, \\ \widehat{x}_k &= \widehat{x}_k^- + K_k(r_k - h(\widehat{x}_k^-, k)), \\ P_k &= (1 - K_k H(\widehat{x}_k^-))P_k^- \end{aligned} \tag{4}$$

where K is the Kalman gain, and $H(\widehat{x}_k^-) = \frac{\partial h}{\partial x}(\widehat{x}_k^-, k)$.

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. In addition, multiple initial values for α and ε are used to help find the estimates with minimal mean squared fitting errors. The tracking procedure is applied, combined with a dynamic programming procedure to find the optimal glottal opening instant t_0, and the asymmetry point t_p. The flow chart for estimating the three LF-model parameters by the proposed TD LF-fitting approach is presented in Fig. 1.

c) Codebook Based Spectral LF-fitting

The frequency-domain LF-model fitting approach used here is similar to the algorithm proposed in [12] but with a modified initialisation procedure. Firstly, a codebook is generated of two thousand LF-model parameter sets along with the corresponding amplitudes of the first six harmonics. For each pitch period of the glottal flow derivative (GFD), a 256-point Hamming window (with the glottal closing instant in the centre) is applied and the GFD spectrum is obtained by the Fast Fourier Transform. Subsequently, the mean squared error (MSE) between the first 6 harmonic amplitudes of the GFD spectrum and those in the codebook is calculated. The set of LF-model parameters generating the minimal MSE is selected as the initial estimate.

The refinement of the estimate uses a two-step optimisation procedure. The first step is to adjust the initial estimates of the three LF-model parameters by minimising the differences between the first six harmonics of the GFD and the LF-model spectrum, H_{GFD} and H_{LF}, by the Nelder-Mead multidimensional unconstrained non-linear algorithm. The differences for the first two harmonics are doubled (they are strongly weighted as they are more related to the glottal contribution) and the cost function is given by:

$$\begin{aligned} D_1 = 2 \cdot \sum_{n=1}^{2} (H_{GFD}(n) - H_{LF}(n))^2 \\ + \sum_{n=3}^{6} (H_{GFD}(n) - H_{LF}(n))^2. \end{aligned} \tag{5}$$

The second step is to adjust the estimate of the return phase parameter T_a, leaving T_p and T_e unchanged. As the LF-model return phase is an exponential decay function, T_a mainly contributes to higher frequency components. Therefore, the Nelder-Mead algorithm is applied to minimise a Itakura-Saito spectral distance [15] between the reconstructed LF-model spectrum from the LF parameter estimates and the GFD spectrum:

$$D_2 = \frac{1}{N} \cdot \sum \left(\frac{P_{GFD}(\omega)}{P_{LF}(\omega)} - log\left(\frac{P_{GFD}(\omega)}{P_{LF}(\omega)}\right) - 1\right). \tag{6}$$

The frequency ω has a range from after the 6^{th} harmonic to half the sampling frequency, P is the amplitude and N is the number of frequency samples corresponding to ω. The input and output of each stage of the algorithm are presented in Fig. 2.

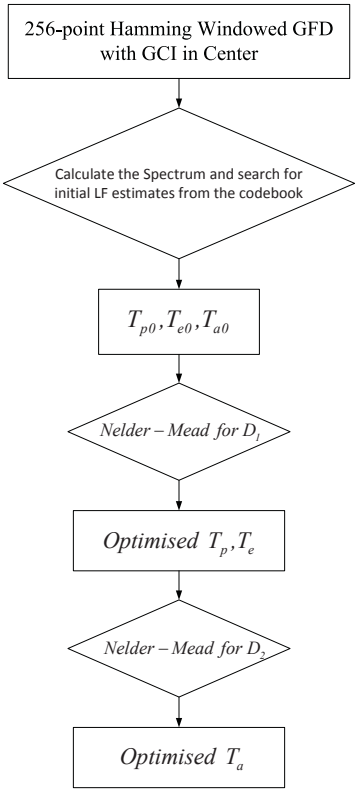

Fig. 2: Flow chart of the spectral LF-fitting algorithm.

III Evaluation

To compare the two LF-model fitting methods, both artificial and real glottal source signals were used. The experiment details are presented below.

a) Artificial Glottal Source

50 sets of LF-model pulses (each set consisting of 10 identical pitch periods) were randomly generated from the ranges presented in Table 1 and corresponding to a wide range of voice qualities.

Authorized licensed use limited to: DUBLIN CITY UNIVERSITY. Downloaded on April 21,2021 at 13:20:59 UTC from IEEE Xplore. Restrictions apply.

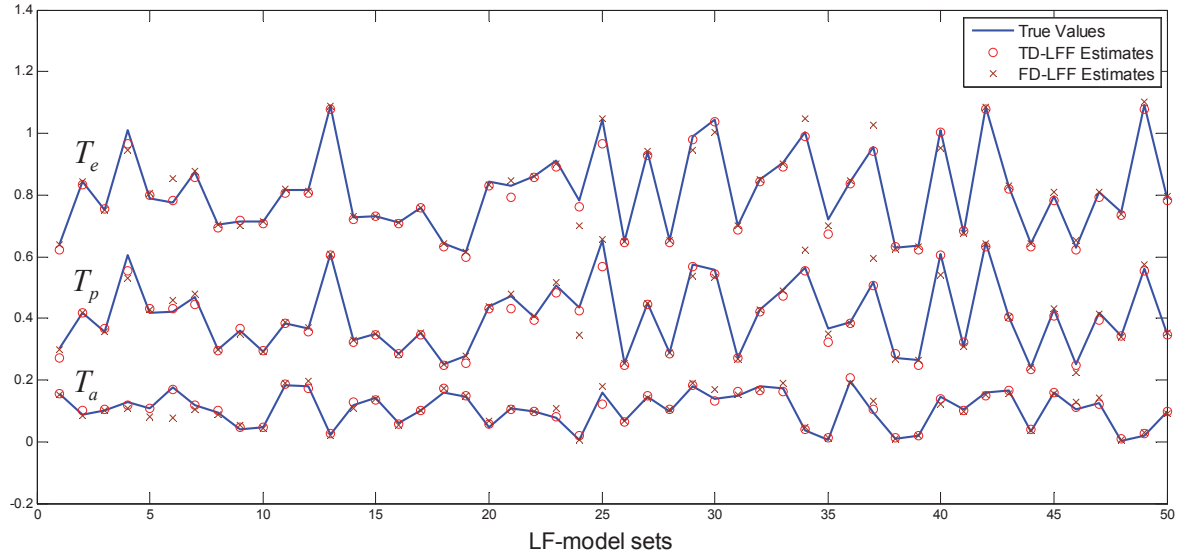

Fig. 3: Artifical glottal source LF-model parameter true values and the estimates by TD-LFF and FD-LFF algorithms

Table 1: Range of LF-model parameters

LF Parameter	Range
T_p	[0.2, 0.72]
T_e	[0.3, 0.8]
T_a	[0, 0.2]

Table 2: RMSE scores for the three LF-model parameters by TD- and FD-LF fitting methods applied to artificial speech data

LF Parameter	TD-LFF	FD-LFF
T_p	0.0196	0.0257
T_e	0.0182	0.0258
T_a	0.0086	0.0191

Subsequently, the time-domain and the frequency-domain LF-model fitting (TD-LFF and FD-LFF) algorithms were applied to extract the estimates of the three LF parameters.

The results are presented in Fig. 3 (all values of T_e were increased by 0.3 for better illustration). In addition, the root-mean-square errors (RMSE) between the estimates and their true values were calculated and are presented in Table 2. It can be observed that for a "clean", artificial glottal source signal, both TD- and FD-LF-model fitting algorithms can generate acceptable estimates. It can also be seen that in most cases, TD-LFF outperforms FD-LFF.

Table 3: RMSE scores for the three LF-model parameters by TD- and FD-LF fitting methods applied to real speech data

LF Parameter	TD-LFF	FD-LFF
T_p	0.0543	0.0852
T_e	0.0527	0.1017
T_a	0.0111	0.0274

b) Real Glottal Source

The real speech data [1] are based on the all-voiced utterance "We were away a year ago". The inverse filtered glottal source waveform and corresponding hand-labelled LF-model parameters for one male speaker was selected, and 100 pitch periods of the source signal were extracted excluding poor inverse filtering results.

The TD- and FD-LF fitting methods were applied to the glottal source waveform. The hand-labelled LF-model parameters and the estimates by both algorithms are presented in Fig. 4. The RMSE scores between the hand-labelled data and estimates were calculated and are presented in Table 3. It can be observed that overall TD-LFF has a better performance than FD-LFF. For pitch periods 2-12 and 18-27, FD-LFF generated inaccurate T_p and T_e while the estimates obtained by TD-LFF are very close to the true values. For the remaining 79 pitch periods, performance of the two approaches varies. Fig. 5 shows an example where

[1]Supplied by Dr. Yanushevskaya of Phonetics & Speech Laboratory, Centre for Language and Communication Studies, Trinity College Dublin

Authorized licensed use limited to: DUBLIN CITY UNIVERSITY. Downloaded on April 21,2021 at 13:20:59 UTC from IEEE Xplore. Restrictions apply.

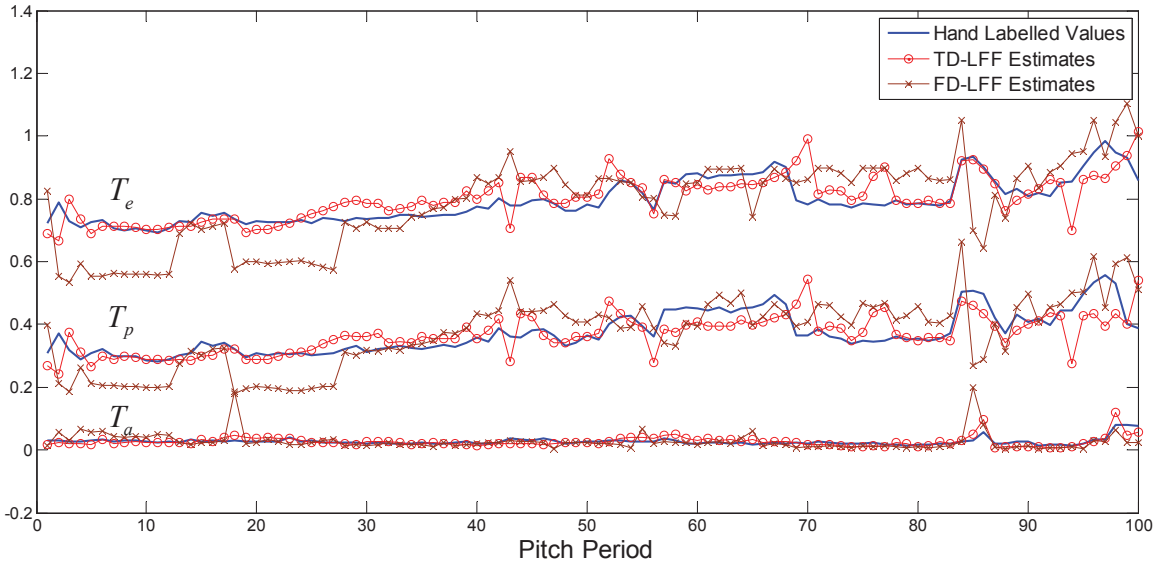

Fig. 4: Real glottal source LF-model parameter hand-labelled values and the estimates by TD-LFF and FD-LFF algorithms

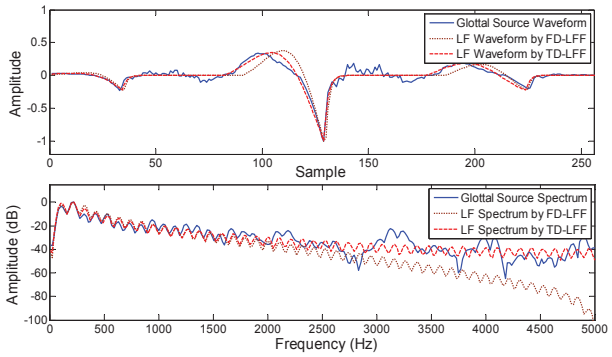

Fig. 5: An example where the LF-model is better fitted by TD-LFF to the real glottal source

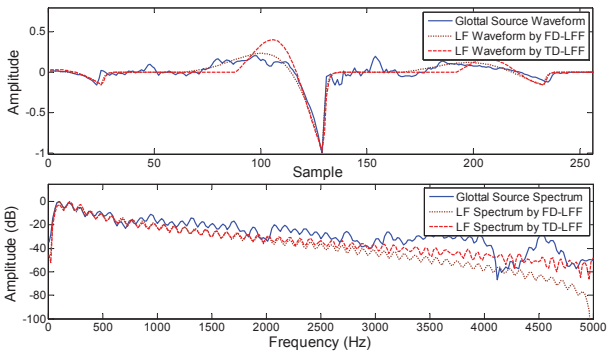

Fig. 6: An example where the LF-model is better fitted by FD-LFF to the real glottal source

TD-LFF performs better. It can be observed that the LF-model waveform obtained by FD-LFF is poorly fitted to the glottal waveform open phase. An inaccurate fit of the third harmonic in the spectrum plot is also visible. Fig. 6 illustrates a case where FD-LFF outperforms TD-LFF. Due to the weak amplitude of the glottal waveform and a large number of ripples appearing in the open phase, TD-LFF failed to locate the glottal opening instant. In addition, it can be observed in the spectrum plot that the first and third harmonic of the estimated LF-model by TD-LFF is not very well fitted to the glottal spectrum compared to FD-LFF.

IV Discussion & conclusion

In this paper we examined two glottal source LF-model fitting methods (one time-domain, one frequency-domain) and compared them by applying each algorithm to artificial and real glottal source signals. Experimental results show that, for an artificial glottal source signal, both methods can generate accurate LF-model estimates, although the time-domain based approach performs slightly better. For inverse filtered real speech glottal source with hand-labelled source parameter data, the time-domain method shows generally more reliable estimates of the LF parameters compared to the spectral fitting method. It can be observed that the spectral fitting approach is more sensitive to the goodness of the inverse filtered glottal source. Even small fitting errors to low frequency harmonics of the glottal spectrum may result in inaccurate estimates of the open quotient parameter T_e and the asymmetrical parameter T_p. The time-domain method is more likely to generate poor estimates for glottal waveforms of low amplitude and with a large number of ripples in the open phase.

Improved performance of the LF-model fitting

Authorized licensed use limited to: DUBLIN CITY UNIVERSITY. Downloaded on April 21,2021 at 13:20:59 UTC from IEEE Xplore. Restrictions apply.

algorithms requires further developments not only of the fitting method, but also in speech decomposition. It is obvious that the cleaner the glottal source waveform, the easier it is to fit the LF-model to it. Although the experimental results presented here show that the time-domain fitting method outperforms the frequency-domain method in most cases, more evaluation is necessary across different speakers and utterances to test the performance of the two methods. It is interesting to note that no single method consistently outperforms the other. This suggests that for robust source parameter estimation, a hybrid approach that combines estimates from both TD- and FD-domain methods is worthy of further investigation.

V Acknowledgements

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] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow", STL-QPSR, vol. 4, no. 1985, pp. 113, 1985.

[2] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "HMM-based speech synthesis with an acoustic glottal source model", *Proc. of The First Young Researchers Workshop in Speech Technology*, Dublin, Ireland, April, 2009.

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

[4] T. Drugman, T. Dubuisson, T. Dutoit, "On the Mutual Information between Source and Filter Contributions for Voice Pathology Detection", in *Proceedings of Interspeech*, 2009.

[5] I. Yanushevskaya, M. Tooher, C. Gobl, and A. N Chasaide, "Time- and amplitude-based voice source correlates of emotional portrayals", in *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII '07)*, Vol. 4738, pp. 159-170, 2007.

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

[7] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", *Speech Communication*, vol. 11, no. 23, pp. 109118, 1992.

[8] 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.

[9] E. L. Riegelsberger, A. K. Krishnamurthy, "Glottal source estimation: Methods of applying the LF-model to inverse filtering", in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*, vol. 2, pp. 542-545, 1993.

[10] H. Strik, B. Cranen, and L. Boves, "Fitting a LF-model to inverse filter signals", in *ESCA 3rd European Conference on Speech Communicati-on and Technology: EUROSPEECH 93*, Berlin, pp. 103106, 1993.

[11] H. Strik, "The effect of low-pass filtering on estimated voice source parameters", in *5th European Conference on Speech Communication and Technology*, 1997.

[12] 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.

[13] N. Henrich, C. d' Alessandro, and B. Doval, "Spectral correlates of voice open quotient and glottal flow asymmetry:theory, limits and experimental data", in *Seventh European Conference on Speech Communication and Technology*, 2001.

[14] 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.

[15] P. Chu, D. Messerschmitt, "A frequency weighted Itakura- Saito spectral distance measure", in *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.30, no.4, pp. 545-560, Aug 1982.

Authorized licensed use limited to: DUBLIN CITY UNIVERSITY. Downloaded on April 21,2021 at 13:20:59 UTC from IEEE Xplore. Restrictions apply.