ROBUST TRACKING OF GLOTTAL LF-MODEL PARAMETERS BY MULTI-ESTIMATE FUSION

Haoxuan Li, Darragh O'Brien and Ronan Scaife

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

ABSTRACT

A new approach to robust tracking of glottal LF-model parameters is presented. The approach does not rely on a new glottal source estimation algorithm, but instead introduces a new extensible multi-estimate fusion framework. Within this framework several existing algorithms are applied in parallel to extract glottal LF-model parameter estimates which are subsequently passed to quantitative data fusion procedures. The preliminary implementation of the fusion algorithm described here incorporates three glottal inverse filtering methods and one time-domain LF-model fitting algorithm. Experimental results for both synthetic and natural speech signals demonstrate the effectiveness of the fusion algorithm. The proposed method is flexible and can be easily extended for other speech processing applications such as speech synthesis, speaker identification and prosody analysis.

Index Terms— LF-model, glottal source, data fusion

1. INTRODUCTION

Robust estimation of glottal source parameters is important for a number of applications. To improve naturalness, modern speech synthesis systems require a flexible and realistic parametric model to represent the glottal source rather than a simple pulse-train [1]. Glottal source parameters can be used to identify speakers [2]. Prosody analysis and modification needs not only intonation information but also voice source control rules for quality improvement [3].

Speech is generally considered to be the convolution of two components: the glottal source and the vocal tract impulse response. Thus, to track the glottal source parameters we need to decompose speech into its two components. The most widely used method is glottal inverse filtering (GIF). Firstly the vocal tract filter coefficients are estimated, and used to remove the vocal tract resonances from the speech signal [4, 5] to yield the glottal source waveform. Subsequently the source component parameters are obtained by fitting the Liljencrants-Fant (LF) model [6] to the glottal source signal [7, 8].

The diversity and complexity of human speech (and extraneous factors such as recording devices and ambient noise), pose significant challenges to any single glottal source estimation algorithm. For example, the performance of closed phase inverse filtering (CPIF) may be influenced by the selection of the closed glottis interval or the validity of the zero excitation assumption, and iterative adaptive inverse filtering (IAIF) does not perform well for higher fundamental frequencies [10]. As such it is difficult to find a consistently accurate and reliable estimate of the glottal source component. Instead, as in the approach presented here, it may be more reasonable to draw on different algorithms in parallel to extract the multiple source parameter estimates and to combine them by quantitative data fusion techniques. To the best of our knowledge, only limited research has been carried out into this approach. One example is the method proposed by Drugman [11], which combines glottal source estimates from two different algorithms. However only the glottal open phase parameters are estimated and combined and the limited variation of parameters across continuous pitch cycles for individual speakers is not considered and the fusion rules are straightforward.

This paper presents a general framework for tracking the complete set of glottal source LF-model timing parameters of voiced speech segments by fusing the estimates obtained from multiple algorithms. In the following sections the structure of the fusion algorithm is presented, a preliminary implementation of the fusion algorithm is described, and experimental results are provided for both synthetic and real speech signals.

2. FUSION FRAMEWORK

Our fusion framework is depicted in Fig. 1. A voiced speech segment is firstly divided into overlapping frames $F_1, \dots, F_L$. Next, two or more speech decomposition (SPD) algorithms are applied to each frame in parallel. These algorithms may include Linear Prediction-based glottal inverse filtering techniques [4, 5 and 12] or any other source-vocal tract separation methods [10, 13]. Each SPD algorithm separates the speech signal into glottal source and vocal tract components. Generally the vocal tract component can be

© EURASIP, 2012 - ISSN 2076-1465

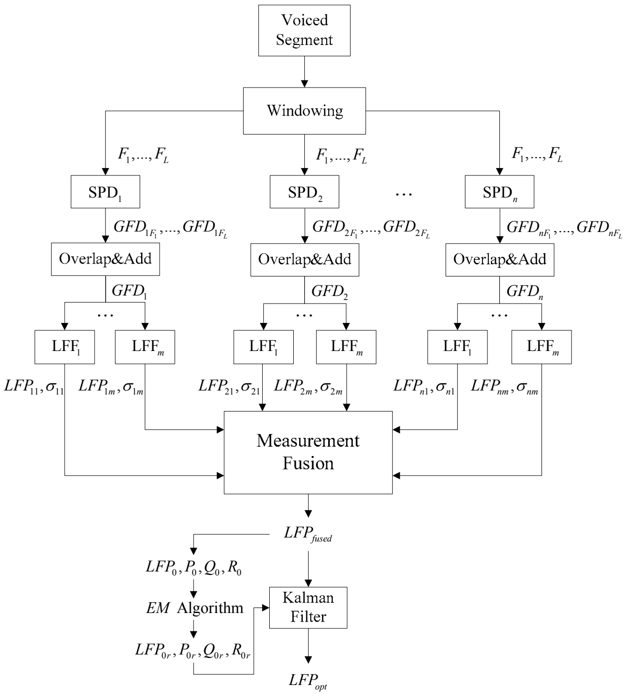

Figure 1. A general framework of the multi-estimate fusion algorithm

described by an all-pole filter with the filter coefficients obtained by an optimisation algorithm such as linear prediction [4]. With the extracted vocal tract coefficients the speech signal can be inverse filtered to cancel the vocal tract effect. Once the vocal tract component is removed the glottal flow derivative (GFD) signal is obtained. For each of the overlapping frames the corresponding inverse filtered GFD signals obtained from the same SPD algorithm are concatenated by an overlap-and-add procedure and the outputs of this stage are n GFD (where n is the number of decomposition algorithms) signals for the original voiced segment.

Next the glottal source parameters are estimated. Because of its effectiveness for approximately 83% of natural phonations [14], the LF-model [6] is currently used in the proposed framework for representing the glottal source. Each GFD signal is divided into consecutive pitch periods. Subsequently, one or more LF-model fitting (LFF) algorithms are applied to each pitch period of the GFD signals. The LFF algorithm is used to estimate the glottal LF-model parameters (LFP) by fitting the LF-model to the GFD signal. Given n speech decomposition algorithms and m LF-model fitting algorithms, a total of $n\times m$ sets of LF-model parameter estimates are obtained for each pitch period. In addition, for each set of estimated LF parameters, an error covariance σ is calculated.

Finally, the fusion procedure is applied. For a single pitch period the $n\times m$ sets of estimated LF-model paramters are combined by the generalised Millman's fusion formula [15] given in (1):

$$LFP_{fused} = a_{11}LFP_{11} + \cdots + a_{ij}LFP_{ij} + \cdots + a_{nm}LFP_{nm},$$
$$a_{11} + a_{12} + \cdots + a_{ij} + \cdots + a_{nm} = 1,$$
$$a_{ij} = \frac{1}{\sigma_{ij}}\left(\frac{1}{\sigma_{11}} + \cdots + \frac{1}{\sigma_{ij}} + \cdots + \frac{1}{\sigma_{nm}}\right)^{-1}, \qquad (1)$$

where i=1,…n, j=1,…m, and a_{ij} is the weighting factor of the corresponding set of LFP, which is calculated from the error covariances. It can be observed that the smaller the covariance, the more weight is given to the set of estimates. The measurement fusion procedure is applied across all pitch periods in the speech signal to produce a single, fused set of LF-model parameter estimates.

In order to obtain reliable parameter trajectories, it is necessary to smooth the fused LF-model parameters across all pitch periods. It is reasonable to assume limited variation in glottal source parameters across adjacent pitch periods especially for sustained vowel sounds. Tooher [16] shows that the variation of LF-model parameters can be regarded as a linear process. Thus, assuming that the true glottal source parameters are the system state to be tracked, and the fused LF-model parameter estimates are the measurement, the corresponding state-space process and measurement equations can be described by (2):

$$rLFP_k = rLFP_{k-1} + w_k,$$
$$LFP_k = rLFP_k + v_k, \qquad (2)$$

where $rLFP$ is the vector of real glottal source parameters, LFP is the vector of fused voice source estimates, w and v are the process noise and measurement respectively, with Gaussian distributions $p(w) = N(0,Q)$, $p(v) = N(0,R)$.

Equation (2) is based on the following two premises: 1) that the variation of the voice source parameters across adjacent pitch periods is small, so that the true glottal source parameters of the k^{th} pitch cycle can be represented by source parameters of the $(k-1)^{th}$ pitch cycle plus process noise; 2) that the estimated source parameters can be considered as a summation of true parameter values and measurement noise. With such correlation, it is reasonable to use a Kalman filter (KF) to track the glottal source parameters. The KF performs best when the process and measurement noise covariances are known. However, for real speech signals there is no such a priori information available. The expectation-maximisation (EM) algorithm [17] is a machine learning technique for optimisation by recursively adjusting the estimates to maximise the corresponding log-likelihood. It is applied here to refine KF parameters given a set of "reasonable" initial values. Afterwards, with these re-estimated parameters the KF is utilised to estimate the optimal glottal source parameters LFP_{opt} across pitch periods of the full voiced segment.

3. PRELIMINARY IMPLEMENTATION

Our preliminary implementation of the multi-estimate fusion algorithm is described as follows. The input voiced speech segment is divided into frames of length 40ms with 50% overlap. Each frame is processed by three glottal inverse filtering algorithms (SPD in Fig. 1).

The first algorithm is iterative adaptive inverse filtering (IAIF) [5], which is based on the assumption that the glottal flow waveform can be represented by a first-order all-pole model. The IAIF algorithm operates by repeatedly removing the glottal and radiation effects using low order Linear Prediction analysis and inverse filtering. This removes the overall spectral tilt of the speech and allows estimation of the vocal tract filter using high order linear prediction analysis. The estimated vocal tract filter is used to inverse filter the original speech signal to extract the glottal flow derivative.

Weighted recursive least square with variable forgetting factor (WRLS-VFF) analysis [12] is the second speech decomposition method. The approach assumes that the speech signal is generated by an ARMA model. WRLS-VFF analysis operates by recursively minimising the prediction error for speech samples and allowing the variation of the forgetting factor. During the analysis process, the forgetting factor and the ARMA coefficients are obtained. Generally the maximum prediction error occurs at the glottal closure instant. Accordingly, the ARMA coefficients at the instant of glottal closure can be used to do the inverse filtering.

Iterative closed phase inverse filtering (ICPIF) [4] is also used in this preliminary implementation. Typically closed phase inverse filtering [10] operates on the assumption that for several milliseconds after the glottal closing instant the glottis remains closed and during this time the speech signal is due solely to the decaying vocal tract response. Thus, linear predictive analysis performed across this time interval models only the vocal tract filter and excludes any components due to the glottal source. The glottal waveform can be determined by inverse filtering the entire pitch period with the coefficients obtained from the closed phase. The first-order autocorrelation parameter can be used to measure the smoothness of the estimated glottal source waveform [4] and an iterative analysis procedure is applied to select the smoothest glottal flow derivative (GFD) signal by utilising different closed phase intervals. In our implementation of ICPIF, the initial glottal closing instants are found by analysing the variable forgetting factor λ obtained from WRLS-VFF [12].

GFD signals obtained from each SPD algorithm are concatenated by an overlap-and-add procedure to generate the entire glottal flow derivative signal for the original input speech segment. Afterwards, a new time-domain LF-model fitting (NTDLFF) algorithm [7] is applied to the three sets of GFD signals to extract the glottal source shape parameters period by period. These parameters are the open quotient $O_q = T_e/T_0$, the asymmetry coefficient $\alpha_m = T_p/T_e$ and the return phase parameter $R_a = T_a/T_0$. The error covariance for individual pitch periods is calculated from the fitting errors between the inverse filtered GFD signal and reconstructed LF-model. Thus three sets of estimated LFP sequences are combined by the measurement fusion procedure. Subsequently Kalman filtering with the EM algorithm is applied to track the parameter trajectories. The preliminary implementation of the fusion algorithm is presented in Fig. 2.

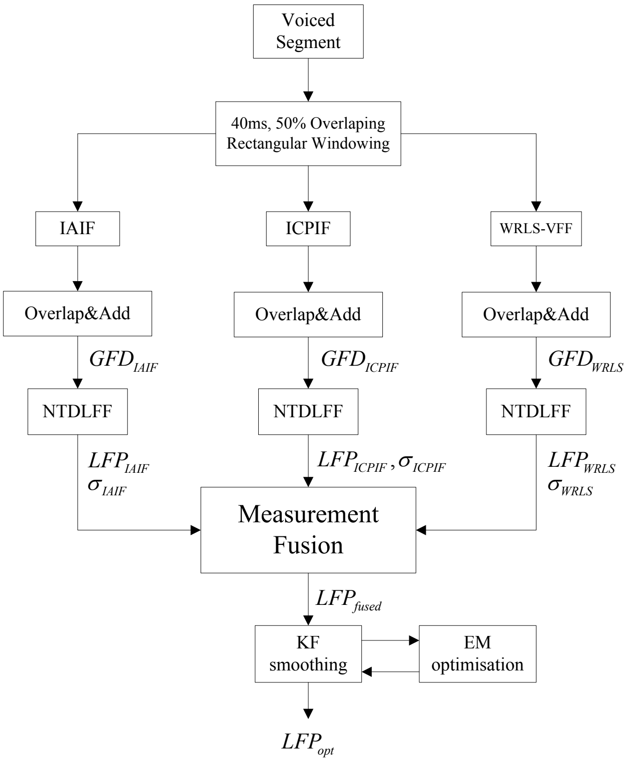

Figure 2. Preliminary implementation of multi-estimate fusion algorithm

4. EVALUATION

4.1. Synthetic Speech

To test the validity of the fusion algorithm, a segment of synthetic speech was generated as follows: 1) 50 LF-model pulses were created from a set of LF parameters $T_p = 0.48$, $T_e = 0.65$, $T_a = 0.035$ and $T_0 = 1$. 2) The first 20 were passed through a formant synthesizer for the vowel /AH/ and the last 20 pulses for the vowel /IH/ (thus two sustained vowel segments were obtained.) 3) A “coarticulatory” segment was generated by synthesizing the middle 10 pulses with line spectral frequencies calculated by linear interpolation from /AH/ to /IH/. 4) The three segments were concatenated.

The multi-estimate fusion algorithm was applied to this synthetic speech segment. The root mean square error (calculated by $RMSE(\hat{x}) = \sqrt{E((\hat{x}-x)^2)}$, where x is the true value and $\hat{x}$ is the corresponding estimate) of the estimated LF-model parameters by each algorithm and by

Table 1. RMSE scores of LF-model parameters estimated by different algorithms for clear synthetic speech

	O_q	α_m	R_a	cov_m
IAIF	0.0298	0.0174	0.0247	0.0317
ICPIF	0.0311	0.0194	0.0289	0.0626
WRLS-VFF	0.0422	0.0205	0.0383	0.0691
ME-FUSION	0.0272	0.0142	0.0239	

the fusion method are presented in Tables 1, with the corresponding mean error covariances cov_m in the last column. It can be observed that the fusion algorithm shows consistently smaller RMSE scores compared to other methods. For all three LF-model parameters, both IAIF and ICPIF performed well, and a relatively bigger RMSE was generated by WRLS-VFF. It is worth mentioning that more weight was given to the IAIF estimates by the fusion procedure, due to its producing the smallest mean error covariance.

4.2. Natural Speech Segments

10 voiced speech segments were randomly extracted from the CMU-ARCTIC database [18] (5 from the male voice bdl, 5 from the female voice slt), including sustained vowel sounds and voiced transitions, for robustness testing. There is no a priori information available for the glottal source parameters for this real speech, therefore it is difficult to quantify estimation errors. The coefficient of variation (*CV=Standard deviation/Mean*) used in [8] can describe the robustness of the estimated parameters, so it was applied here for the evaluation. Due to space limitations, only the mean scores of the results across all segments for each speaker are presented in Fig. 3 and Fig. 4. Although it is apparent that results vary across parameters and between speakers, the CV scores are consistently lower for the multi-estimate fusion method compared to the other glottal inverse filtering algorithms used alone.

5. CONCLUSIONS & FUTURE WORK

A general multi-estimate fusion framework for tracking glottal LF-model parameters was presented and our preliminary implementation of the algorithm was described. Experimental results for both synthetic and real speech showed that by combining the estimates extracted from different algorithms with a quantitative data fusion technique, the tracked LF-model parameters were more robust than those obtained from any single algorithm used in isolation.

In our preliminary implementation only one LF-model fitting algorithm was applied; to improve the performance of the fusion algorithm, different methods will be used in future work such as a frequency-based fitting algorithm [8] which shows robustness to low-frequency phase distortion.

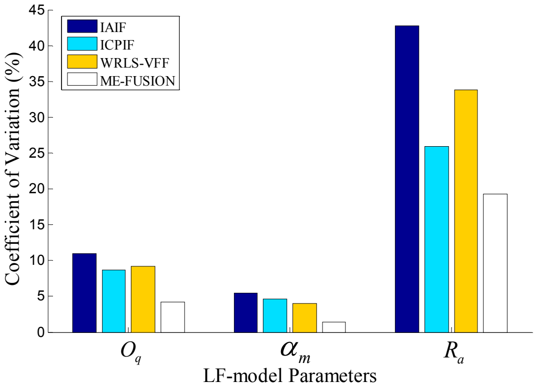

Figure 3. Coefficient of variation (%) for the three LF-model parameters in male speech segments

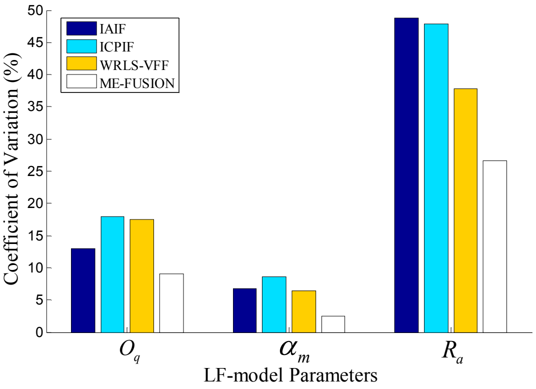

Figure 4. Coefficient of variation (%) for the three LF-model parameters in female speech segments

Also because of the flexibility of the fusion algorithm many different glottal source estimation methods can be applied. For example, although for the fusion framework the source-vocal tract separation and glottal source model fitting procedures are in two levels, joint source-tract estimation [9] can be used to directly extract the source parameters. In addition, other effective multi-sensor data fusion techniques (information filter, data association, etc.) will be utilised to test the significance of the proposed algorithm for a variety of speech signals.

To test the effectiveness of the fusion algorithm for real speech, more comprehensive evaluation methods would be applied. For example, the estimated LF-model parameters can be used to remove the glottal effect from the original speech signal to yield only the vocal tract component. Subsequently, spectral analysis techniques such as discrete all-pole modeling (DAP) [19] can be applied to measure the goodness-of-fit of the DAP model to the extracted vocal tract component. In addition, perceptual listening tests on speech re-synthesised from estimated voice source parameters could be used to measure the effectiveness of the fusion method.

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] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis", *ISCA SSW6*, 2007.

[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] H. Strik, L. Boves, "On the relation between voice source parameters and prosodic features in connected speech", *Speech Communication*, Vol. 11, pp. 167-174, 1992.

[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] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", *Speech Communication*, vol. 11, no. 23, pp. 109-118, 1992.

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

[7] 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.

[8] J. Kane, M. Kane, and C. Gobl, "A spectral LF model based approach to voice source parameterisation", in *11th Annual Conference of the International Speech Communication Association*, 2010.

[9] 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.

[10] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform", *IEEE Trans. on Acoustics, speech and signal processing*, pp. 350-355, 1979.

[11] T. Drugman, T. Dubuisson, N. D'Alessandro, A. Moinet, and T. Dutoit, "Voice source parameters estimation by fitting the glottal formant and the inverse filtering open phase", in *Proceedings of 16th European Signal Processing Conference,* 2008.

[12] D. G. Childers, J. C. Principe, Y.T. Ting, "Adaptive WRLS-VFF for speech analysis", *IEEE Trans. Speech and Audio Processing*, vol.3, no.3, pp.209-213, 1995.

[13] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation", in *Proc. Interspeech*, pp. 116-119, 2009.

[14] H. Strik and L. Boves, "On the relationship between voice source parameters and prosodic features in connected speech," in *Speech Communication*, vol. 11, pp. 167-174, 1992.

[15] V. Shin, Y. Lee, and T. Choi, "Generalized millman's formula and its application for estimation problems," *Signal Processing*, vol. 86, pp. 257-266, 2006.

[16] M. Tooher and J. G. McKenna, "Variation of glottal LF parameters across F0, vowels, and phonetic environment", in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[17] 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.

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

[19] A. El-Jaroudi, J. Makhoul, "Discrete all-pole modeling", *IEEE Trans. on Signal Processing*, vol.39, no.2, pp.411-423, Feb 1 991.