GLOTTAL FLOW DERIVATIVE MODELING WITH THE WAVELET SMOOTHED EXCITATION

Arthur P. Lobo[†]

Ericsson Mobile Communications (UK) Ltd.
The Keytech Centre, Ashwood Way, Basingstoke
Hampshire RG23 8BG, UK

ABSTRACT

This paper discusses a method for estimating glottal flow derivative model parameters using the wavelet-smoothed excitation. The excitation is first estimated using the Weighted Recursive Least Squares with Variable Forgetting Factor algorithm. The raw excitation is then smoothed by applying a Discrete Wavelet Transform (DWT) using Biorthogonal Quadrature filters, and a thresholding operation done on the DWT amplitude coefficients, followed by an inverse DWT. The pitch period and the instant of glottal closure (IGC) are estimated from the wavelet-smoothed excitation. A six-parameter glottal flow derivative model consisting of three amplitude and three timing parameters is aligned with the IGC and optimized by minimum square error fitting to the speech waveform. The optimization is done by the method of simulated annealing The model is then used to reestimate the vocal-tract filter parameters in an ARX procedure followed by further stages of voice source-vocal tract estimation. The results of analysis of speech utterances from the BK_TIMIT database will be presented.

1. INTRODUCTION

The glottal volume velocity waveform or the flow is the excitation component of the voiced speech production model [1]. It has been shown that the shape of the glottal pulse is important in preserving the naturalness of synthetic speech [2, 3, 4]. Further, the pulse can be parameterized and used as the excitation source in low bit rate speech coding [5] and for speech storage applications. In addition the glottal pulse parameters can be used to augment the vocal tract information for improved speaker identification. As a consequence, methods for estimating the parameters of the glottal flow from the speech signal are of significant interest.

Previously researchers have attempted to model the glottal flow from the inverse filtered speech waveform in which the filter is estimated from a LP analysis of the closed-glottal phase [6, 7, 8, 9, 10, 11]. In the proposed method, a vocal-tract ARMA filter is recursively estimated over a whole larynx cycle and the filter coefficients corresponding to the minimum in the wavelet-smoothed excitation are used in the optimization of the glottal flow derivative model. The optimization is done by minimum square error fitting to the speech waveform. The voice source model is then used as excitation in the reestimation of the ARMA parameters over the larynx cycle. A number of iterations of the source-tract estimation are carried out until the glottal flow derivative pulse shape converges. The glottal flow derivative model is allowed a non-zero value in the closed phase. Further the modeling of either a positive or negative going spike (corresponding to the instant of glottal closure) or both is allowed using the Fujisaki-Ljunqvist [12] model. The open phase is modeled using sine and cosine terms as done by Ananthapadmanabha [13].

The glottal flow derivative model parameters are optimized using the method of simulated annealing [14] in which the parameters are sampled from a bounded space. This method of combinatorial optimization has an advantage over nonlinear least-squares methods based on computing gradients [15, 16] when the error surface has a global minimum along with many local minima, as is the case for the glottal parameter optimization problem. Specifically in simulated annealing, solutions in the uphill direction are accepted in a stochastic manner dependent on the temperature (Metropolis' criterion).

2. METHODOLOGY

The speech signal is first preemphasized and the input excitation estimated by the adaptive Weighted Recursive Least Squares with Variable Forgetting Factor algorithm [17]. The raw excitation is smoothed by applying a Discrete Wavelet Transform (DWT) using 16-coefficient Biorthogonal Quadrature filters (symmetric low pass/ antisymmetric high pass) with three vanishing moments [18], and a thresholding operation done on the DWT

[†] Email: Arthur.Lobo@eml.ericsson.se

0-7803-7041-4/01/$10.00 ©2001 IEEE

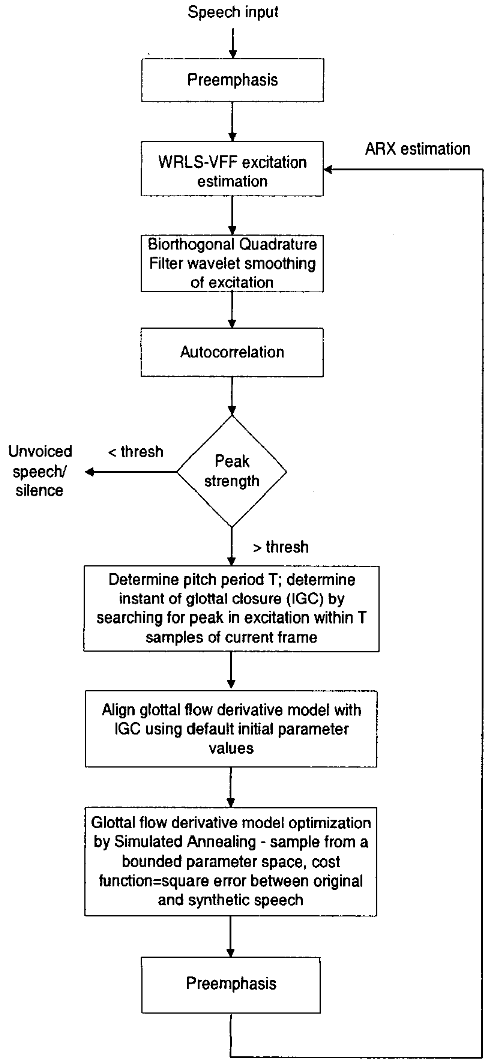
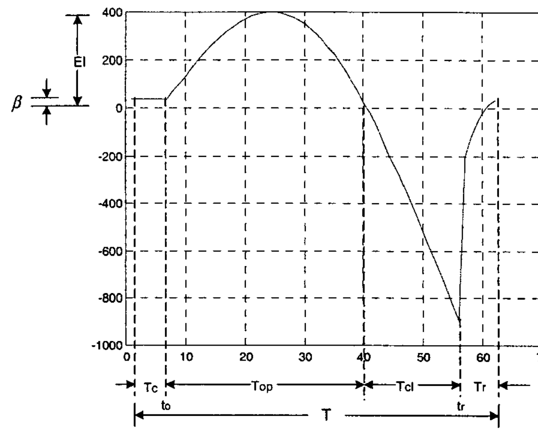


Fig. 1. Flow chart for estimation of Glottal Flow Derivative model parameters.

amplitude coefficients, followed by an inverse Discrete Wavelet Transform. In this way the aperiodic components in the input excitation for eg. due to aspiration noise and ambient noise are removed while at the same time preserving the slow and rapid variations in the underlying waveform. This is possible because of the compactness property of wavelets i.e. localization in time. The pitch period is determined by an autocorrelation of the wavelet-smoothed excitation. The instants of glottal closure (IGC) are determined next using the pitch information and searching for the peak in the excitation. A six-parameter glottal flow derivative model consisting of three amplitude and three timing parameters is preemphasized and aligned with the IGC. The IGC is also



T_c = closed phase
T_{op} = opening phase
T_{cl} = closing phase
T_r = return phase
β = slope during closure
EI = peak positive slope
EE = strength of excitation
C = slope after closure
T = pitch period

Fig. 2. Model of the Glottal Flow Derivative.

optimized within a range of [-*T*/4, *T*/4] where *T* is the pitch period. The initial amplitude parameter values are the default values for the first cycle at speech onset, or the previous cycles' optimized values for an intermediate cycle in a speech segment. The parameters are optimized to fit the preemphasized speech over the pitch period using the method of simulated annealing. The vocal-tract filter corresponding to the minimum in the wavelet-smoothed excitation is used to resynthesize speech. The vocal-tract filter parameters are then reestimated in an ARX (X-exogenous input) procedure with the excitation defined by the optimized glottal flow derivative model. This is followed by further stages of model optimization and vocal-tract estimation until the glottal source model converges. Fig. 1 illustrates one iteration of the estimation and optimization algorithm. Fig. 2 shows a model of the glottal flow derivative with time along the horizontal axis and amplitude along the vertical axis. The glottal flow derivative model is given by the following equations,

$$U_g'(t) = -\beta \qquad 0 < t \le t_0$$

$$U_g'(t) = -\beta + (EI + \beta) * \sin\left(\pi \frac{t - t_0}{T_{op}}\right) \qquad t_0 < t \le t_0 + \frac{T_{op}}{2}$$

$$U_g'(t) = (EI + EE) * \cos\left(\pi \frac{t - \left(t_o + T_{op}/2\right)}{4 * T_{cl}}\right) - EE$$

$$t_o + \frac{T_{op}}{2} < t \le t_r$$

$$U_g'(t) = -C + 2\frac{C-\beta}{T_r}(t - t_r) - \frac{C-\beta}{T_r^2}(t - t_r)^2 \qquad t_r < t \le T$$

where,

$$\beta = \frac{CT_r}{T_r - 3(T - T_{op} - T_{cl})},$$

t is the time variable and the prime denotes the first differential with respect to time. t_o and t_r are the opening and return instants respectively.

The length of the frame for the DWT is 256 and the number of decomposition levels is 7. At initialization, the bounds for the amplitude parameters *EI*, *EE*, and *C* are set as: $0 \le EI \le 32767$, $0 \le EE \le 32767$, $-32768 \le C \le 32767$. The bounds for the time parameters are initialized every pitch period to: $0 \le T_r \le \frac{T}{3}$, $0 \le T_{op} \le T$, $0 \le T_{cl} \le T$ where T is the pitch period. During the parameter sampling in the simulated annealing procedure, if $(T_{op} + T_{cl} + T_r) > T$, the parameters are resampled. The cost function is,

$$E_k = \left\| s_k - \hat{s}_k \right\|_2 \quad \text{where}$$

$$\hat{s}_n = -\sum_{i=1}^{p} a_i \hat{s}_{n-i} + \sum_{j=1}^{q} b_i u_{n-j} + u_n$$

E_k provides a measure of error between the speech signal, s_k and the synthetic speech signal, $\hat{s}_k$ generated using the glottal flow derivative model, u and the vocal-tract filter coefficients corresponding to the minimum in the wavelet-smoothed excitation within a larynx cycle. p and q are the pole and zero orders, and a_i and b_i are the AutoRegressive and Moving Average coefficients respectively of the vocal-tract filter. The index k refers to the k^{th} configuration of glottal parameters.

3. RESULTS

The results of analysis of a female speech utterance from the BK_TIMIT (Bruel & Kjaer secondary microphone recordings) database is shown in Fig. 3a. The utterance is "And their chr/on/iclers are not the dramatic poets but the prose novelists". The analyzed segment is given within //. The speech sampling rate is 8 kHz. The pole order for the WRLS-VFF algorithm is 10 and the zero order 2. 5 iterations of the estimation and optimization algorithm in Fig. 1 were executed per pitch period. The plots below (amplitude vs sample number) are for the speech segment, the wavelet-smoothed WRLS-VFF excitation at iteration 1, the optimized glottal flow derivative model at iteration 1, the WRLS-VFF excitation at the final iteration, the wavelet-smoothed WRLS-VFF excitation at the final iteration, and the optimized glottal flow derivative model at the final iteration.

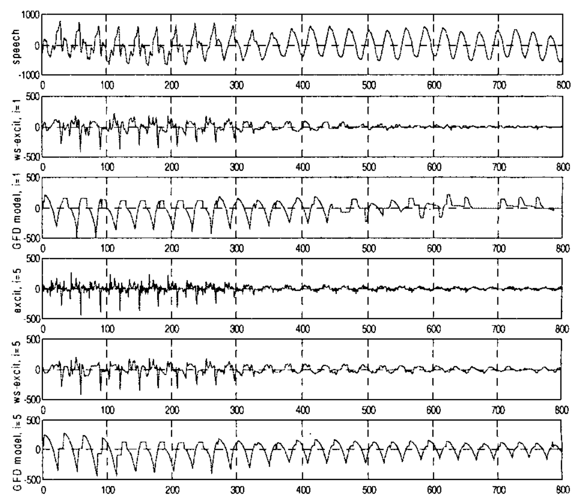


Fig. 3a. Analysis of the female speech utterance /aa n/.

The results of analysis of a male speech utterance from the BK_TIMIT database is shown in Fig. 3b. The utterance is "Then he fled, not /waiting/ to see if she minded him or took notice of his cry". The analyzed segment is given within //. The analysis parameters are the same as for the female utterance case.

In both cases the glottal parameters are seen to vary more smoothly in the final iteration than in iteration 1. In the female utterance case, the pitch periods from sample 500 onwards, corresponding to the nasal /n/, are not distinct in the WRLS-VFF excitation, whereas they are clearly visible in the wavelet-smoothed excitation (iteration 5). The pitch periods are more distinct in iteration 5 than in iteration 1. In the male utterance case, a secondary excitation which appears at sample 617 in the raw excitation in iteration 4 has been diminished in strength in the wavelet-smoothed excitation, and the same is observed for a secondary excitation at sample 1299. Thus the wavelet-smoothing of the excitation has made it possible to fit glottal flow derivative models in

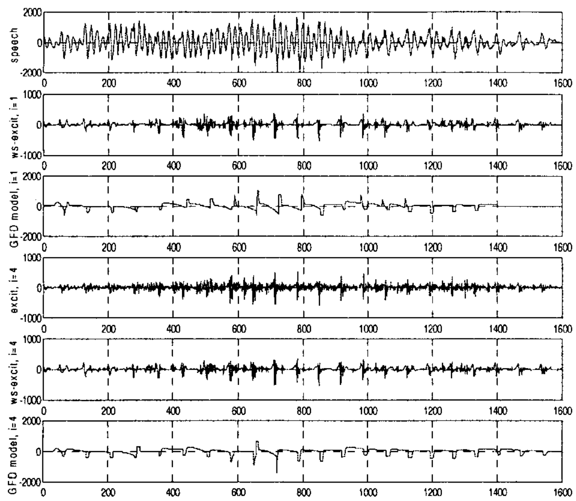


Fig. 3b. Analysis of the male speech utterance /w ey dx ih ng/.

the low signal-to-noise ratio regions. Secondly, extraneous IGCs in the raw excitation have been filtered out in the wavelet-smoothed excitation thus providing a robust estimate of the pitch period.

4. CONCLUSION

A single-channel method for estimating the parameters of a glottal flow derivative model from the speech signal has been implemented. The input excitation is estimated by the adaptive WRLS-VFF algorithm. The excitation is smoothed by applying a wavelet analysis and synthesis transform with thresholding of the wavelet transform amplitude coefficients. A glottal flow derivative model is optimized to fit the speech waveform pitch-synchronously using the vocal-tract filter coefficients corresponding to the minimum in the wavelet-smoothed excitation. The optimization is done by the method of simulated annealing. The optimized voice source model is then used to reestimate the vocal-tract filter parameters in an ARX procedure followed by further stages of model optimization and vocal-tract estimation. The method has been tested on female and male speech utterances.

REFERENCES

[1] G. Fant, "Acoustic theory of speech production," s-Gravenhage:Mouton, The Netherlands, 1960.

[2] A.E. Rosenberg, "Effects of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, vol. 49, pp. 583-590, 1971.

[3] J. N. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. on Audio and Electroacoustics*, AU-21, No. 3, pp. 298-305, 1973.

[4] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, 87, pp. 820-857, 1990.

[5] P. Hedelin, "A glottal LPC-vocoder," *Proc. IEEE ICASSP*, San Diego, CA, pp. 1.6.1-4, 1986.

[6] M.G. Beirouti, D.G. Childers, and A. Paige, "Glottal area versus glottal volume velocity," *Proc. IEEE ICASSP*, Hartford, CT, pp. 33-36, 1977.

[7] D.Y. Wong, J.D. Markel, and A.H. Gray, "Least-squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. on ASSP*, ASSP-27, pp. 350-355, 1979.

[8] H.W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Am.*, 56 No. 5, pp. 1625-1629, 1974.

[9] T.V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of the closed glottis interval," *IEEE Trans. on ASSP*, vol. 27, pp. 309-319, 1979.

[10] A.K. Krishnamurthy and D.G. Childers, "Two channel speech analysis," *IEEE Trans. on ASSP*, vol. 34, pp. 730-743, 1986.

[11] H. Strik, "Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses," *J. Acoust. Soc. Am.*, 103(5), pp. 2659-2669, 1998.

[12] H. Fujisaki and M. Ljunqvist, "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," *Proc. IEEE ICASSP*, pp. 637-640, 1987.

[13] T.V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics," *STL-QPSR*, 2-3, pp. 1-24, 1984.

[14] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671-680, 1983.

[15] A. Isaakson and M. Millnert, "Inverse glottal filtering using a parameterized input model," *Speech Communication*, 18, pp. 435-445, 1989.

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

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

[18] I. Daubechies, "Ten lectures on wavelets," *CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 61, SIAM Press, Philadelphia, Pennsylvania, 1992.