PROPOSAL AND EVALUATION OF MODELS FOR THE GLOTTAL SOURCE WAVEFORM

Hiroya Fujisaki and Mats Ljungqvist
Faculty of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, Japan

ABSTRACT

Speech analysis for high quality speech synthesis or high accuracy speech recognition requires realistic models not only for the vocal tract but also for the voice source. In the present paper, we investigate models for the glottal volume velocity waveform. Previously proposed models are reviewed and classified according to their level of elaboration in expressing the glottal characteristics. A new model is then proposed which possesses all the important features of previously proposed models. A method is also described for simultaneously estimating the glottal source and vocal tract parameters. Using this method, evaluation of glottal model parameters is carried out on real speech by varying the number of parameters in the proposed model. The results indicate the importance of detailed modeling of the period of glottal closure for accurate analysis.

1 INTRODUCTION

Voice source studies are of obvious importance in basic speech research. Furthermore, it is well-known that accurate representation of the voice source is of great importance for naturally sounding speech synthesis. While the vocal tract characteristics can be fairly precisely modeled by all-pole or pole-zero models, we still lack an efficient model formulation for the voice source. The main reason is that we do not have enough knowledge of the phonatory behavior of the glottis, which has proved to be fairly complex. As a consequence there is a lack of speech analysis methods which take the voice source into consideration. Conventional LPC methods, for example, use an all-pole model for all the spectral shaping processes of speech production: voice source, vocal tract transfer function and radiation transfer impedance, assuming only an impulse train for the voice source.

Glottal models can be divided into two main categories: interactive and non-interactive. In interactive models, the glottal flow is calculated from glottal area [1-4] or conductance [5] functions by incorporating the various impedances of the acoustic system into the model. Structural modeling of the mechanical vibration of the vocal cords has also been attempted [6]. These models require detailed knowledge about the physical characteristics of the various parts of the glottis, which we do not usually possess. Furthermore, the perceptual significance of source-tract interaction is not well established. This makes, at present, non-interactive models of glottal flow the most attractive candidates for practical voice source modeling.

Non-interactive models directly parameterize the glottal flow or flow derivative function, thereby implicitly including some effects of source-tract interaction, such as inertive loading by the sub- and supraglottal acoustic systems which is a contributing factor to the skewing of the glottal pulse.

In the present paper we aim at classification and evaluation of various models for the glottal flow [7-11]. The models are classified according to their special features. A new model is then proposed which can approximate nearly all of the previously proposed models. A method is also presented for simultaneous estimation of both glottal and vocal tract models [12]. Using this method, comparison of several glottal models is carried out.

2 GLOTTAL MODEL FORMULATIONS

2.1 Previously Proposed Glottal Models

Glottal action can be studied by means of optical or electroglottographical methods, which measure the vocal cord movements, but do not give the glottal flow directly. Furthermore, it is desirable to be able to extract the essential features of the glottal source directly from the speech waveform. Thus, inverse filtering has become one of the most important methods for glottal source analysis. The adjustment of the inverse filter is, however, often based on subjective, qualitative judgements regarding the shape of the glottal wave. In order to establish objective, quantative criteria, modeling of the glottal wave has to be introduced. Thus, mathematical models capable of describing the most essential characteristics of the glottal waveform are of interest for the general study and classification of various modes of glottal excitation and for high quality speech synthesis. Many such models have been proposed. In Table I, most of the important glottal flow models to date are classified according to their level of complexity.

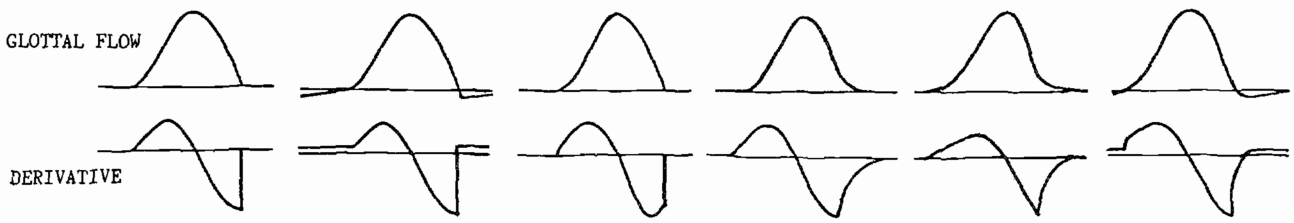
Rosenberg [7] proposed a number of models for the glottal flow with adjustable pulse amplitude, width and skew. One of them were composed of two trigonometric segments with a single slope discontinuity at glottal closure. This is usually referred to as the "Rosenberg model" (Fig. 1(a)). Hedelin [8] used the same model with the addition of low frequency drift in his glottal LPC- vocoder (Fig. 1(b)). Fant [9] has proposed a model having a facility for independent control of the flow derivative discontinuity (Fig. 1(c)). In certain cases, the glottal flow is characterized by a rounding at closure. This led to refinements in the models: [10] (Fig. 1(d)), and [11] (Fig. 1(e))

CH2243-4/86/0000-1605 $1.00 © 1986 IEEE

TABLE 1. Classification of glottal flow models.

	MODEL	SINGLE FLOW DERIVATIVE DISCONTINUITY	PROVISION FOR MULTIPLE FLOW DERIVATIVE DISCONTINUITIES	PROVISION FOR CONTINUOUS FLOW DERIVATIVE	WAVEFORM REALIZATION
AMPLITUDE, WIDTH AND SKEWING OF THE GLOTTAL PULSE.	(a)	YES	(YES)*	NO	SINUSOIDAL
	(b)	YES	NO	NO	SINUSOIDAL
INDEPENDENT CONTROL OF FLOW DERIVATIVE DISCONTINUITY.	(c)	YES	NO	YES	SINUSOIDAL
DETAILED MODELING OF THE GLOTTAL CLOSURE PERIOD.	(d)	YES	NO	YES	SIN+POLYN.
	(e)	YES	NO	YES	EXP*SIN
	(f)	YES	YES	YES	POLYNOMIAL

* ROSENBERG PROPOSED SEVERAL MODELS, SOME OF THEM PROVIDE FOR MULTIPLE DISCONTINUITIES.



(a) ROSENBERG[7] (b) HEDELIN[8] (c) FANT[9] (d) ANANTHAPADMANABHA[10] (e) LILJENCRANTS-FANT[11] (f) FUJISAKI-LJUNGQVIST

Fig. 1. Waveforms of the glottal models described in Table 1.

2.2 Proposal of a New Glottal Model

We propose a model for the glottal flow whose derivative is composed of polynomial segments as shown in Figs. 1(f) and 2. The choice of a polynomial model makes it easy to vary the number of parameters, and thereby the level of detail in the modeling, which is convenient when evaluating the relative importance of the various parameters. In its most elaborate form, it has three timing parameters controlling open phase duration (W), pulse skew (S) and the time interval from glottal closure to maximum negative flow (D), as well as three amplitude parameters controlling slope at glottal opening (A), slope prior to closure (B) and slope following closure (C).

Though the A-parameter is not common in other models, we have included it since a secondary excitation is often noted at glottal opening. The rounded closure, which is often noticed on glottal flow waveforms, is sometimes attributed to a gradual glottal closure leaving a small residual flow after the main excitation. We consider that there is also a component attributable to a period of negative flow due to a lowering of the vocal cords following glottal closure.

3 ESTIMATION OF MODEL PARAMETERS

Assuming a linear model of speech production, the speech signal can be viewed as the convolution of the source signal with the impulse responses of the vocal-tract and radiation filters.

By introducing realistic modeling of the voice source, we can separate the voice source from the rest of the acoustic system. It is practical to include also the radiation characteristics in the voice source model. Thus, we use the glottal flow derivative as a combined voice source and radiation model.

The estimation of the unknown excitation and vocal-tract transfer functions requires iterative procedures. We adopt an A-b-S (Analysis-by-Synthesis) approach in which the glottal model parameters are estimated together with the vocal tract transfer function, by minimization of the prediction error. The system is outlined in Fig. 3.

In studies of the detailed waveform of the glottal source, it is important to assure freedom from phase distortion. Ordinary speech recordings are often subject to serious low frequency phase distortion. This can be compensated for by compensation filtering, which we carry out in a two step procedure [13]. In the first step, distortion introduced by the tape recorder and amplifier is cancelled using a prerecorded calibration signal and time-reversed filtering. In the second step, the microphone characteristics is compensated for by inserting a variable all-pass filter after the glottal model. The filter parameters are then estimated in the error minimization procedure, and the filter is used for simulating the phase characteristics of the microphone.

3.1 The Vocal Tract Model

In conventional linear predictive analysis, the spectral envelope is modeled by an all-pole filter. The speech signal can then be expressed by the equation

$$s(n)=\sum_{i=1}^{p} a_i s(n-i)+e(n),$$

where s(n) is the speech signal, a_i's are the predictor coefficients, p is the predictor order and e(n) is the error signal.

When introducing a known input signal, the above equation can be modified into the following form

$$s(n)=\sum_{i=1}^{p} a_i s(n-i)+a_{p+1}g(n)+e(n),$$

Here $g(n)$ is the glottal model (combined with radiation) with its amplitude given by a_{p+1}. Minimization of the squared error:

$$E=\sum_{n=0}^{N-1}[s(n)-\sum_{i=1}^{p} a_i s(n-i)-a_{p+1}g(n)]^2,$$

results in a system matrix similar to the one in conventional covariance LPC, but with one additional row and column, which contain the crosscorrelation between the source and speech signals. The system can be solved efficiently by Cholesky decomposition.

3.2 Error Minimization

The routine for optimizing the glottal parameters is based on a "hill-climbing" search for the minimum prediction error as outlined in Fig. 3.

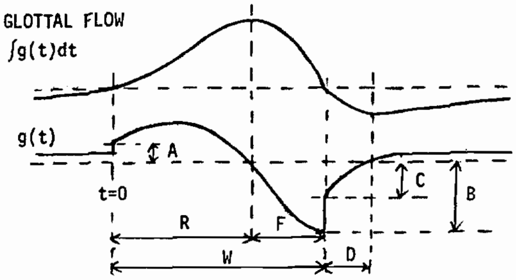
The algorithm can be divided into the following three steps:

(1) Generation of a glottal model wave.

(2) The generated glottal wave is used with the speech signal as input in the linear estimation of the vocal tract transfer function by minimization of the mean squared error. In this step the glottal amplitude is obtained.

(3) The prediction error is evaluated and the glottal parameters are modified accordingly.

The procedure is repeated until the prediction error reaches a minimum. In this way simultaneous estimation of both voice source and vocal tract parameters is achieved. The minimization scheme requires a large amount of computation which, however, can be reduced by taking advantage of the relative stationarity of the glottal waveform. During stationary voiced segments, most of the computation time is then consumed in the correct localization of the glottal pulse, which has to be fairly accurate. The algorithm allows for a pulse positioning accuracy of 1/10th of the sampling period. Presently the system runs at 50-100 times real time on a Hitachi M280H computer, depending on the number of glottal parameters that are to be estimated.



GLOTTAL PARAMETERS

W - PULSE WIDTH (R+F)	A - SLOPE AT GLOTTAL OPENING
S - PULSE SKEW (R+F)/(R-F)	B - SLOPE PRIOR TO CLOSURE
D - GLOTTAL CLOSURE TIMING	C - SLOPE FOLLOWING CLOSURE

$$g(t)=\begin{cases} A-\frac{2A+R\alpha}{R}t+\frac{A+R\alpha}{R}t^2, & 0<t\leq R,\\ \alpha(t-R)+\frac{3B-2F\alpha}{F^2}(t-R)^2-\frac{2B-F\alpha}{F^3}(t-R)^3, & R<t\leq W,\\ C-\frac{2(C-\beta)}{D}(t-W)+\frac{C-\beta}{D^2}(t-W)^2 & W<t\leq W+D,\\ \beta & W+D<t\leq T, \end{cases}$$

$$\text{where}\quad \alpha=\frac{4AR-6FB}{F^2-2R^2}\quad\text{and}\quad \beta=\frac{CD}{D-3(T-W)},$$

T = fundamental period.

Fig. 2. Parameters and formulas for the proposed glottal model (Fujisaki-Ljungqvist model).

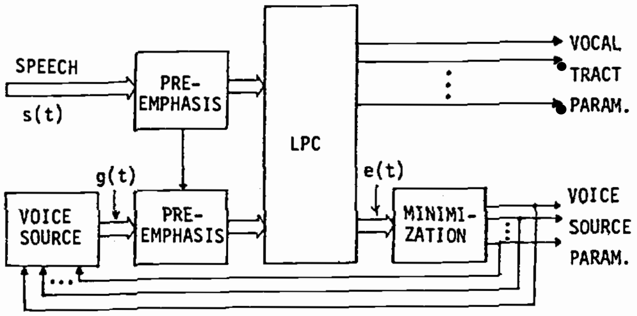


Fig. 3. Block diagram of the model estimation.

4 EXPERIMENTS ON MODEL COMPARISON

The above method was used for comparison of various glottal waveform models. We used four different versions of the Fujisaki-Ljungqvist (FL) model and compared their performances with those of Rosenberg (ROS), Hedelin (HED) and Liljencrants-Fant (LF) models. The four FL models were defined as: FL-1 with A, C and D constrained to be equal to zero and the other parameters being variable, FL-2 with C and D constrained to be equal to zero, FL-3 with C constrained to be equal to B, and FL-4 with all six parameters being variable. Natural vowels of various vocal efforts were analyzed and the result was evaluated in terms of the logarithm of the prediction error improvement ratio, i.e. the reduction of prediction residual power as compared to standard LPC for the same speech frame ($10\text{X}\log_{10}(E_{lpc}/E_{glpc})$).

Figures 4 and 5 illustrate and compare the results of analysis of a segment of the vowel /a/ (male voice), using the three models FL-1, FL-4 and LF. It can be seen in Fig. 4, that there is a significant reduction of the prediction error when more refined models are used (LF and FL-4). Figure 5 shows the estimated transfer functions and the FFT-spectrum for the segment. This example and analysis of other speech samples indicate that more accurate voice source modeling results in better separation of the voice source and the vocal tract characteristics, which is manifested as a reduction of the prediction error. Figure 6 shows the average prediction error reduction, as compared to conventional LPC, from analysis of 30 vowel samples. It can be seen that even the simpler models give about 3 dB reduction of the prediction error, while the model FL-4, which allows the most detailed modeling of the glottal closure phase, gives about 4.2 dB reduction.

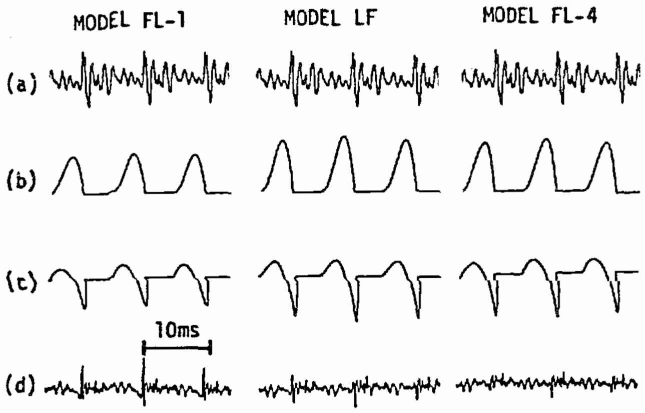


Fig. 4. Results of analysis of a segment of the vowel /a/ (male voice, pitch synchronous analysis, predictor order 10). (a) speech wave, (b) glottal flow, (c) flow derivative, (d) prediction error (amplified 3 times as compared to (a)).

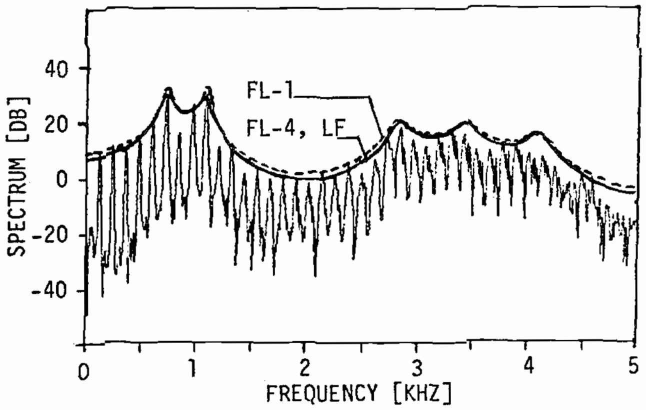


Fig. 5. Vocal tract transfer functions estimated in the analysis of Fig. 4 and FFT-spectrum of the same speech segment.

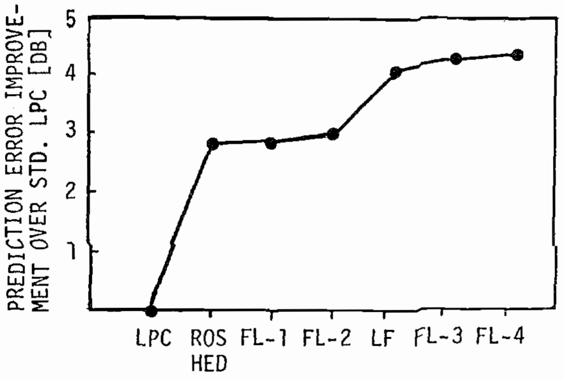


Fig. 6. Model evaluation in terms of average prediction error improvement over conventional LPC.

5 SUMMARY

We have presented a comparative study of several waveform models for the glottal flow. For this purpose, we proposed a new glottal model as well as a method for simultaneous estimation of glottal and vocal tract models based on LPC and iterative search.

It was shown that the introduction of a glottal model reduces the prediction error with about 3-4 dB as compared to conventional LPC. It was also shown that a model which provides for detailed modeling of the glottal closure period performs the best, while provision for glottal opening discontinuity seems to be of less importance.

Work is in progress on a speech analysis-synthesis system based on the glottal modeling approach described in this paper.

ACKNOWLEDGEMENT

This work was supported by a Grant-in-Aid from the Sound Technology Promotion Foundation.

REFERENCES

[1] Guérin, B., Mryati, M., and R. Carré, "A Voice Source Taking Account of Coupling with the Supraglottal Cavities," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp.47-50, 1976.

[2] Ananthapadmanabha, T.V. and G. Fant, "Calculation of True Glottal Flow and its Components," STL-QPSR, KTH, No 1, pp.1-30, 1982.

[3] Allen, D.R., W. J. Strong, "A Model for the Synthesis, of Natural Sounding Vowels," J. Acoust. Soc. Am. Vol. 78, No. 1, 1985.

[4] Titze, I.R., "Parameterization of the Glottal Area, Glottal Flow, and Vocal Fold Contact Area," J. Acoust. Soc. Am., Vol. 75, No. 2, 1984.

[5] Rothenberg, M., "An Interactive Model for the Voice Source," STL- QPSR, KTH, No. 4, pp.1-17, 1981.

[6] Ishizaka, K., and J.L. Flanagan, "Synthesis of Voiced Sounds from a Two Mass Model of the Vocal Cords," Bell Syst. Tech. J. Vol. 50, pp. 1233-1268, 1972.

[7] Rosenberg, A.E., "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," J. Acoust. Soc. Am., Vol. 49, No.2(part 2), 1971.

[8] Hedelin, P., "A Glottal LPC-vocoder," Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 1.6.1-1.6.4, 1984.

[9] Fant, G., "Glottal Source and Excitation Analysis," STL-QPSR, KTH, No. 1, pp. 85-70, 1979.

[10] Ananthapadmanabha, T.V., "Acoustic Analysis of Voice Source Dynamics," STL-QPSR, KTH, No. 2-3, pp. 1-24, 1984.

[11] Fant, G., Liljencrants, J. and Q. Lin, "A Four-Parameter Model of Glottal Flow," Presented at French-Swedish Symposium, Grenoble, France, April 1985.

[12] Ljungqvist, M. and H. Fujisaki, "A Method for Simultaneous Estimation of Voice Source and Vocal Tract Parameters Based on Linear Predictive Analysis," Trans. Committee on Speech Research, Acoust. Soc. Japan, No. S85-21, 1985.

[13] Ljungqvist, M. and H. Fujisaki, "Correction of Low Frequency Distortion in Speech Recordings and its Effect on the Glottal Wave Shape," Proc. Spring Meeting of Acoust. Soc. Japan, pp. 161- 162, 1985.