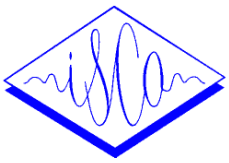

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

STUDIES OF GLOTTAL EXCITATION AND VOCAL TRACT PARAMETERS USING INVERSE FILTERING AND A PARAMETERIZED INPUT MODEL

J.P. LIU, G.BAUDOIN and G.CHOLLET

ESIEE department of signal and telecommunication
B.P.99, Noisy le Grand, 93162 Cedex France

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. This paper presents a comparison between two analysis methods for the calculation of the voice source and vocal tract parameters for voiced sounds. The first method computes once the vocal tract parameters and then ajusts a glottal parametric model on the residual by a quadratic optimization procedure. The second method iteratively computes the parameters of the vocal tract and glottal models. It was shown that first method, though a least complexity leads to signal-noise ratio slightly lower than the second method.

1 INTRODUCTION

There is clear evidence that the glottal source waveform has a significant impact on the voice quality of speech synthesizers. Differences between male and female voice [1], and between different phonation types (e.g., normal, breathy, pressed, etc) appear to be related to systematic differences in the characteristics of the glottal source waveform. To achieve natural sounding synthetic speech, it is essential that we understand and model these glottal source variations.

One of the major difficulties in using realistic glottal source models in synthesis is the lack of suitable analysis methods to estimate the glottal waveform parameters from natural speech. Assuming a linear model of speech production, the speech signal can be viewed as the convolution of the source signal with the impulse reponse 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 pratical to include also the radiation characteristic in the voice source model. Thus, we use the glottal flow derivative as a combined voice source and radiation model (Fig 1).

Two different categories of analysis techniques have been developed to extract the voice source and vocal tract parameters. Both of them use the same models for the voice source and the vocal tract. But the parameters are calculated differently. The first method is called PSIF: Pitch Synchronized Inverse Filtering and the second PSARX: Pitch Synchronized ARX. In the PSIF method the vocal tract parameters are calculated first by using a pitch synchronized LP covariance analysis [2] in a time interval where the glottis is supposed to be closed. Then the glottal model is fitted to the residual by a quadratic optimization procedure.

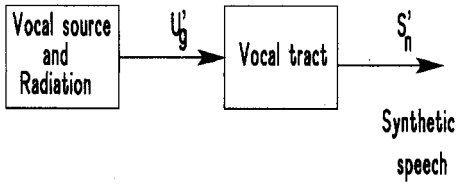

Figure 1: The model of synthesis.

In the PSARX method an analysis by synthesis is used in which the glottal parameters are iteratively estimated together with the vocal tract parameters by minimization of the difference between the prediction error and the glottal model [3] [4]. A quadratic optimization procedure is also used. The first method is of course suboptimal compared to the second method. But its complexity is smaller. The aim of this work is to compare the results of the two methods in terms of quality and complexity.

2 THE VOICE SOURCE AND VOCAL TRACT MODELS

2.1 The source model

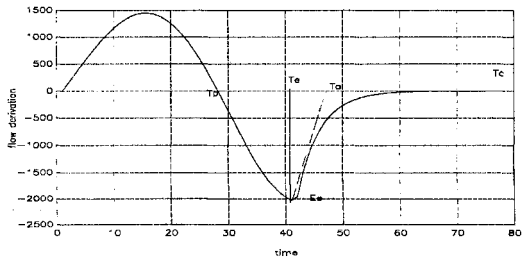

Figure 2: The Liljencrants-Fant model: the differential glottal model.

10.21437/ICSLP.1992-315

Fujisaki and Ljungqvist [6] have given an overview of parametric models that are used to describe the glottal waveform. We have chosen the Liljencrants-Fant (LF) model [7], because it is mathematically well developed and can generate a wide variety of realistic pulse shapes. The model contains a set of parameters believed to be important both from the viewpoints of speech perception and speech production [7]. The model is shown in figure 2, it requires four model parameters for a differential glottal wave: The parameter T_P indicates the moment of maximum flow and T_e corresponds to moment of glottal closure. The return-time T_a models the residual phase of progressing closure after the major discontinuity at T_e. The fourth parameter E_e represents the strength of the excitation at Te. The LF model is defined by

$$E(t) = \frac{\partial u_g(t)}{\partial t} = E_0 e^{\alpha t} \sin w_g t \quad 0 \leq t \leq T_e \qquad (1)$$

$$E(t) = \frac{E_e}{\varepsilon T_a}(e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)}) \quad T_e \leq t \leq T_c \qquad (2)$$

where E_e is $E(T_e)$, $w_g = \pi/T_p$ and T_c is the fundamental period.

The value of α can be derived by using the four basic parameters listed above and by solving the equation

$$\int_0^{T_c} E(t) = 0 \qquad (3)$$

similarly, the value of ε can be derived by solving the equation

$$\varepsilon T_a = 1 - e^{-\varepsilon(T_c-T_e)} \qquad (4)$$

2.2 The vocal tract model

A linear time-invariant, discrete-time all-pole model can be represented , in the time-domain, by

$$s_n + \sum_{i=1}^{p} a_i s_{n-i} = g_n \qquad (5)$$

where g is the input (glottal source), s is the output (speech) and p is the order of the model.

3 PSIF method

This method consists of three stages: in the first stage the fundamental period and instant of glottal closure are determined. In the second stage a pitch synchronized LP covariance is carried out in the time interval where the glottis is supposed to be closed, in order to determined the vocal tract parameters. In the third stage the glottal model is fitted on the residual by using a quadratic optimization method. A block diagram is shown in Fig. 3.

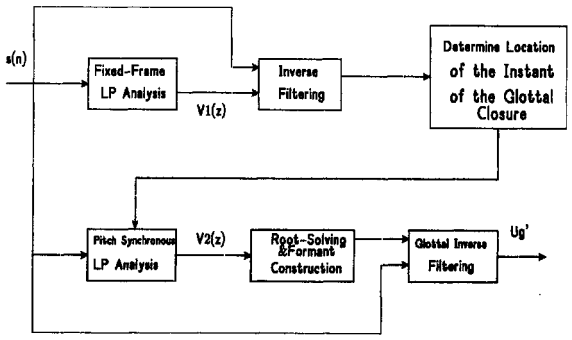

Figure 3: The pitch-synchronized inverse filtering.

3.1 Identification of the instant of the glottal closure

The first step in the analysis procedure is to locate the instant of glottal closure. For this purpose, a pitch-asynchronous (fixed frame) LP analysis is performed on the input speech signal $s(n)$. The estimated LP filter, $V_1(z)$, is used to derive the LP error signal, $q_1(n)$, by inverse filtering. The autocorrelation of the residual is computed to determine the average fundamental period within one frame, the method of epoch extraction [5] is then used to determine the moment of glottal closure. This method consists of computing the Hilbert envelope of the residual signal. The Hilbert transform is realized by using a frequency domain window function and FFT algorithms. The Hilbert envelope of the residual signal shows unambigous peak corresponding to the glottal closure. Figure 4 shows an example of this method.

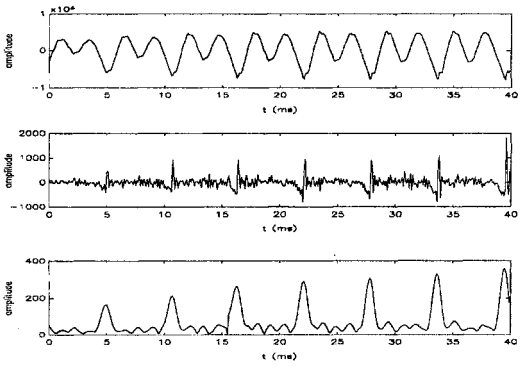

Figure 4: Speech signal of the vowel /e/, LPC residual and Hilbert envelope of the residual signal.

3.2 Pitch Synchronized Inverse filtering

In the second stage, a pitch-synchronous covariance LP analysis with adaptive framelength and filter order is used to estimated an improved LP filter, $V_2(z)$. The formants resonances of the vocal tract are estimated by solving the roots of the LP polynomial. Minor adjustments are necessary to ensure that the inverse filter will only

remove the formant poles from the speech signal. They include: (1) discarding the roots with center frequencies below $250Hz$, (2) discarding the roots with bandwidths greater than $500Hz$, and (3) merging two adjacent roots. The refined formants resonances are, then, used in the second pass inverse filtering procedure. The direct output of the glottal inverse filtering operation is a differential glottal volume-velocity $u'_g(n)$, which represents the combined effect of the lip radiation and the glottal volume-velocity.

3.3 Optimization of glottal parameters

The glottal model is fitted on the residual by means of a nonlinear optimization method; this optimization is performed by a sequential quadratic programming (SQP) method [8] as offered by the MATLAB package.

4 THE PSARX METHOD

In this method the glottal source and vocal tract parameters are estimated together in an iterative analysis by synthesis method. The procedure is quite similar to the one described by Millnert and Jacobson[9].The analysis is carried out pitch synchronously. The voice source and the speech signal are preemphasized. The autocorrelation function of the residual of one frame is computed to determine the average fundamental period within one frame, the method of epoch extraction is used to identifiy estimates of the events in a glottal cycle and to determine the fundamental period by searching for two maxima. The frame-length is three glottal periods and the frame step is one glottal period. The frame length is sufficient to avoid stability problems in the calculation of the vocal tract filter.

The synthetic speech (Fig.1) can be written:

$$s'_n = g_n - \sum_{i=1}^{p} a_i s'_{n-i} \qquad (6)$$

Though it may look most natural to minimise the error criterion $v = \sum e_{1n}^2$ where the z-transform of e_{1n} is

$$E_1(z) = S(z) - S'(z) = S(z) - \frac{G(z)}{A(z)} \qquad (7)$$

this leads to a non-linear minimization problem. Instead we use a modified criterion $w = \sum e_{2n}^2$ with

$$E_2(z) = A(z)S(z) - G(z) \qquad (8)$$

The vocal tract and glottal parameters are estimated by minimizing the mean square error between the prediction residual and the glottal model. Afterwards we call this error e_{2n} "source error". The procedure is: 1.Generation of the voice source signal; 2.Estimation of the vocal tract transfer function using generated source signal and speech signal; 3.The source error is evaluated and the glottal parameters are modified using the method discussed in 3.3. The procedure is repeated as the glottal parameter space is searched for a combination that gives the best description of the input signal in term of minimum source error.

5 RESULTS

Studies were carried out on french oral vowels (/a/, /o/, /i/, /e/, /u/). 10 fundamental periods were analyzed for each vowels. Signals were recorded with special precautions on the linearity of the phase of the acquisition system. The results were evaluated in terms of the ratio in decibels of the energy in the original speech to the resynthesis error resulting from the resynthesizing of the speech wave. Thus, higher the value of this ratio, better the quality of the resynthesized speech.

Figures 5, 6 and 8 illustrate and compare the results of analysis of a segment of the vowel /a/ (male voice) using the two methods. Figures 5 shows the glottal wave derived from the speech samples by using the two methods. It can be seen in Figure 6 that there is a significant reduction of the source error for the PSARX method. Fig. 7 illustres the average number of iterations for arriving the local minimum with the same conditions initial, it indicates that the complexity of the method PSIF is smaller. Figure 9 shows the average original signal to resynthesis error ratio for the analysis of 5 vowels, it indicates that the method PSARX leads to better signal to noise ratio than PSIF.

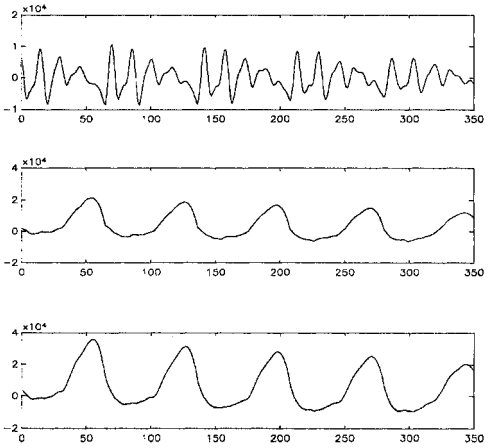

Figure 5: Speech wave /a/, the glottal flow using the method PSARX and PSIF (from top to bottom).

6 CONCLUSIONS

We have presented a comparison of two methods for the determination of the glottal source and vocal tract parameters. Analysis experiments show that the method of PSARX gives improved performance over the method of PSIF in terms of the smaller source error and the smaller resynthesis error, the method of PSIF gives the smaller complexity.

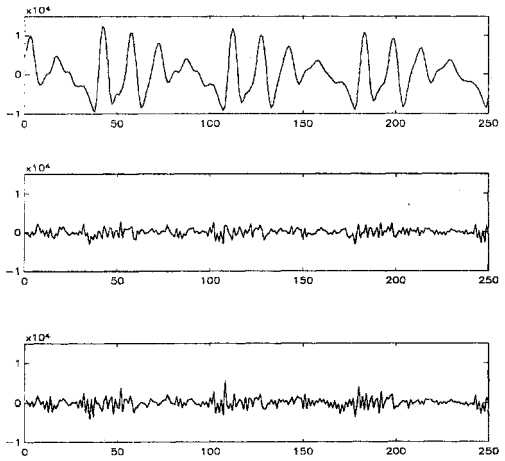

Figure 6: From top to bottom: Speech wave /a/, the source error using the method PSARX and PSIF (amplified 4 times as compared to speech wave).

Methods	Number of iterations
PSIF	76
PSARX	97

Figure 7: The average number of iterations for both methods.

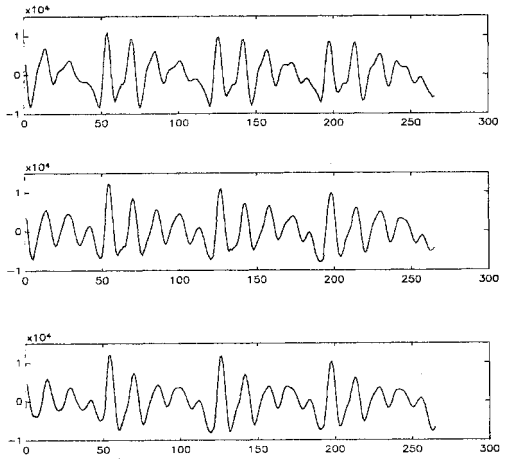

Figure 8: Speech wave /a/ and the resynthesized speech using the method PSARX and PSIF (from top to bottom).

Average original signal to resynthesis error ratio (DB)						
	/a/	/i/	/o/	/e/	/u/	Average
PSIF	16.0608	15.5143	17.0532	15.231	11.7242	15.206
PSARX	17.5703	16.6025	18.0699	16.823	13.1088	16.455

Figure 9: The average original signal to resynthesis error ratio.

REFERENCES

1. D.H.Klatt and L.C.Klatt, "Analysis, Synthesis and Perception of voice quality variations among male and female talkers." J. Acoust. Soc. Amer., vol.87, pp.820-857, Feb. 1990

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

3. H.Fujisaki and M.Ljungqvist, "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

4. P.Hedelin, "High Quality Glottal LPC-Vocoding." Proc. IEEE ICASSP, pp.9.9.1-9.9.4, 1986

5. T.V.Ananthapadmanabha and B.Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval." IEEE Trans. on ASSP, vol. ASSP-27, Aug, 1979

6. H.Fujisaki and M.ljungqvist, "Proposal and Evaluation of models for the glottal source waveform." Proc. IEEE ICASSP, pp.31.2.1-31.2.4, 1986

7. G.Fant, J.Liljencrant, and Q.Lin, "A four parameter model of glottal flow." STL-QPSR 4/1985, pp.1-13

8. P.E.Gill, W.Murray and M.H.Wright, "Practical Optimization." Academic Press, London, 1981

9. A.ISAKSSON, M.MILLNERT, "Inverse glottal filtering using a parameterized input model." Signal Processing, vol.18, No.4, pp. 435-445, Dec. 1989.