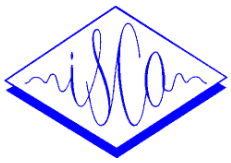

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SELF EXCITED THRESHOLD AUTO-REGRESSIVE MODELS OF THE GLOTTAL PULSE AND THE SPEECH SIGNAL

Jean Schoentgen *

Institute of Modern Languages and Phonetics, CP110, Université Libre de Bruxelles, Av. F. D. Roosevelt, 50, B-1050 Brussels, Belgium
(* National Fund for Scientific Research, Belgium)

ABSTRACT

We propose a composite signal model whose general form is valid for both the glottal pulse and the speech signal. The model consists of two linear autoregressive sub-models. The two sub-models are respectively fitted to the open and return phase components of the glottal pulse or speech signal. In the case of the glottal pulse, the orders of the two sub-models are equal to 2 and 1 respectively. In the case of the speech signal the orders are higher in order to take into account the effects of vocal tract resonance. The switch from one sub-model to the next occurs when the signal crosses a critical threshold. The advantage is that the number and positions of these thresholds are independent of the position and length of the analysis window. As a result, the optimal threshold position, i.e. the best possible segmentation into the open and return phase components, can be found automatically by means of a conventional optimizer. Results show that the proposed model enables the glottal pulse to be segmented automatically and the sub-models to be fitted from within an excitation-asynchronously positioned analysis window. Similarly, when applied to the speech signal, the model automatically provides glottis cycle lengths, the open and return phase components of the speech signal and the open and return phase formant frequencies inside an excitation-asynchronously positioned analysis window.

I. INTRODUCTION

A glottal pulse is an acoustic signal generated by the vibrating vocal folds. A self-excited threshold auto-regressive model (*SETAR*) [7] is a signal model that is made up of two linear auto-regressive sub-models. The switch from one sub-model to the next occurs when the signal amplitude crosses a critical threshold.

The objective of this article is to propose a signal model whose general form is valid for both the glottal pulse and the speech signal. The purpose is to wholly automate (*i*) the fit of glottal pulse models to glottal waveforms obtained by glottis inverse filtering and (*ii*) the excitation-synchronous extraction of formant frequencies of the speech signal.

The best-known glottal pulse model is the so-called Liljencrants-Fant model [2]. It consists of two curves; the first curve (1) is a raised cosine multiplied by an exponential and the second, (2) is a decaying exponential.

$$s_p(n) = A_1 K_1^n cos(\omega n) + C_1 \quad (1)$$

$$s_p(n) = A_2 K_2^n + C_2 \quad (2)$$

$s_p(n)$ is the glottal pulse, n the time index, C_1, C_2, K_1, K_2 and ω are curve parameters and A_1 and A_2 are amplitudes. The switch from curve (1) to curve (2) takes place at the beginning of the so-called return phase of the glottal cycle. The switch from curve (2) to curve (1) occurs at pulse onset following the return phase.

Problems raised by this model are the following:

(*a*) The glottal pulse must be segmented into two components before the individual curves (1) and (2) can be fitted. Within an excitation-asynchronously positioned analysis window, the search for the critical events when the switch between sub-models comes about is difficult to automate because the number of these events is a priori unknown. Also, it is not known whether the first sub-model inside the analysis window is curve (1) or (2).

(*b*) Curves (1) and (2) are nonlinear in parameters ω, K_1, K_2. Fitting curves (1) and (2) to segmented glottal pulses means that the values of these parameters must be iteratively determined by means of an optimization algorithm such as the steepest descent method. Such algorithms easily become bogged down in local minima when good initial estimates of parameters ω, K_1, K_2 cannot be obtained.

(*c*) Curves (1) and (2) are difficult to combine formally with linear models of the vocal tract transfer function to form speech signal models. The reason is that transfer function models are linear constant coefficient difference equations whereas curves (1) and (2) are solutions of such equations. Moreover, these curves are nonlinear in their definitional parameters. Combined source and vocal tract models are needed for excitation-synchronous formant extraction, for example, when the model must distinguish between the open and closed phases of the glottal cycle.

(*d*) The glottal pulse is generated by a self-sustaining oscillator in the shape of the vibrating vocal folds. By contrast, model (1) and (2) is physically interpreted in terms of a parametric oscillator, which is an oscillator that is externally driven by switching one or several of its parameters to and fro. Model (1) and (2) is therefore at odds with the physical principle that underlies voice production.

Hereafter we show that turning driven model (1) and (2) into a self-oscillating signal model solves problems (*a*) to (*d*).

II. SETAR GLOTTAL PULSE MODEL

Sub-models (1) and (2) are solutions of linear second-order (3) and first-order (4) difference equations respectively.

$$s_p(n) = a_0 + a_1 s_p(n-1) + a_2 s_p(n-2), \quad (3)$$

$$s_p(n) = b_0 + b_1 s_p(n-1). \quad (4)$$

$s_p(n)$ is the glottis signal, n is the time index, and a_0, a_1, a_2, b_0 and b_1 are the model coefficients. Equations (3) and (4) can be collapsed into a single equation by letting coefficients a_i depend on time.

$$s_p(n) = a_0(n) + a_1(n) s_p(n-1) + a_2(n) s_p(n-2), \quad (5)$$

where $a_0(n) = a_0$, $a_1(n) = a_1$, $a_2(n) = a_2$ when $0 \leq n\Delta T < T_R$ and $a_0(n) = b_0$, $a_1(n) = b_1$, $a_2(n) = 0$ when $T_R \leq n\Delta T < T$. T is the duration of the glottal cycle, T_R the duration of the open phase and ΔT the sampling interval.

Curves (1) and (2) are solutions to equation (5). Formulating a glottal pulse model in terms of equations (3) and (4) or curves (1) and (2) is therefore mathematically equivalent. Two discrepancies exist, though. Firstly, curve (1) is not a general solution to equation (3). Equation (3) is therefore a more general model than curve (2). Secondly, equation (5) can only be used to synthesize waveforms when appropriate initial conditions $s_p(0)$ and $s_p(1)$ are

known. Therefore, when model (5) is used for synthesis purposes, a procedure for estimating the initial conditions must be provided for in addition to a method for estimating model coefficients a_i.

We propose to refashion model (5) so as to obtain an expression that solves problems (*a*) to (*d*).

(*i*) Reformulating model (1) & (2) in terms of its underlying dynamics (5) solves problem *b*, which is the nonlinear dependence of curves (1) and (2) on their parameters K_1, K_2 and ω. Indeed, these parameters, together with constants C_1 and C_2, can be directly calculated from coefficients a_i and b_i [5]. These coefficients can be estimated by conventional linear methods [4]. Amplitudes A_1 and A_2 are the integration constants. Their values cannot be arrived at from the values of coefficients a_i and b_i. They depend on the initial conditions of equation (5) instead.

(*ii*) The glottal pulse must be segmented into two components, to which sub-models (3) and (4) are fitted. The search for the juncture between these sub-models is difficult to automate because the number of these time points and the type of sub-model with which to start the search are unknown inside an excitation-asynchronously positioned analysis window (problem a). A solutions consists of replacing the search for critical time points by a search for critical thresholds. Indeed, a threshold cuts a convex signal (*i.e.* a signal that has a single maximum per cycle) into two segments. The advantage is that the number and positions of the thresholds are independent of the length and position of the analysis window. The search for the best threshold position can therefore be entrusted to a conventional optimizer.
The problem is that a single threshold cuts a signal into components which necessarily have the same amplitude at the beginning and the end. This is undesirable as far as the glottal pulse is concerned because the amplitudes at the beginning and end of the return phase, for example, are obviously not identical. The solution consists of introducing a delay d that postpones the action of signal s_p crossing the threshold. Combined threshold r and delay d enable any segment to be cut out of a convex signal: threshold r determines the length of the cut-out and delay d its position within the cycle. The search for threshold r and delay d can be automated because the number of delays and thresholds and their values are independent of the position and length of the analysis window. Figure 1 shows how a threshold and delay can be used to segment the glottal pulse into a return and an open phase.

(*iii*) Taking the previous arguments into account, model (3) & (4) is rewritten as follows.

$$\begin{aligned} s_p(n) &= a_0 + a_1 s_p(n-1) + a_2 s_p(n-2) \\ & \quad s_p(n-d) < r, \end{aligned} \tag{6}$$

$$\begin{aligned} s_p(n) &= b_0 + b_1 s_p(n-1) \\ & \quad s_p(n-d) \geq r. \end{aligned} \tag{7}$$

Model (6) & (7) can be collapsed into a single expression similar to (5). The difference is that the coefficients of expression (8) depend on signal amplitude and not on time.

$$s_p(n) = a_0(s_p) + a_1(s_p) s_p(n-1) + a_2(s_p) s_p(n-2). \tag{8}$$

Physically speaking, expression (8) describes a self-sustaining oscillator. Therefore, model (6) & (7) solves problem *d*, i.e. it belongs to the same general class of physical devices as the laryngeal oscillator.

III. SETAR SPEECH SIGNAL MODEL

Since the vocal tract is considered to be a linear device, its signal processing behavior is conventionally described by means of a linear constant coefficient difference equation (9).

$$s(n) = q_0 + q_1 s(n-1) + q_2 s(n-2) + \ldots q_L s(n-L) + x(n). \tag{9}$$

$s(n)$ is the voiced speech signal outputted by a vocal tract without side cavities, q_i the equation coefficients, L the order of the model and $x(n)$ the source signal. Since the glottal pulse is made up of two sub-models that are linear and the concatenation of two linear models is itself a linear model [5], model (9) can be rewritten as follows (problem *c*).

$$\begin{aligned} s(n) &= c_0 + c_1 s(n-1) + c_2 s(n-2) + \ldots + c_N s(n-N) \\ & \quad w(n-d) \geq r, \end{aligned} \tag{10}$$

$$\begin{aligned} s(n) &= d_0 + d_1 s(n-1) + d_2 s(n-2) + \ldots + d_M s(n-M) \\ & \quad w(n-d) < r. \end{aligned} \tag{11}$$

$s(n)$ is the speech signal, c_i and d_i are the model coefficients and M and N the sub-model orders.

Formally, $w(n)$ is glottal pulse $s_p(n)$. But the glottal pulse is not available in most practical cases. Instead, the glottal pulse can be replaced by an auxiliary signal $w(n)$ that is convex and has the same period as the speech signal. Indeed, these two properties are the only properties of the glottal pulse that are made use of in the framework of model (10) and (11). Examples of auxiliary signals are the laryngogram and the smoothed speech signal. Auxiliary and speech signals need not be aligned. Automatically adjusted delay d compensates for any shifts between signals. Figure 2 shows how a convex auxiliary signal can be used to segment the speech cycle into two components.

The main application of model (10) and (11) is the excitation-synchronous extraction of the formant frequencies from within an excitation-asynchronously positioned analysis window. Threshold r, delay d and coefficients d_i and c_i are adjusted so as to minimize an overall modeling error. The closed-phase formant frequencies are calculated in a conventional manner by means of coefficients c_i [4].

IV. METHOD

When glottal pulse model (8) was employed for synthesis purposes, in addition to the values of c_i, d_i, r and d, estimates of the initial conditions of difference equations (6) and (7) were needed. Initial conditions were determined in the following manner. Firstly, the derivative of the glottal pulse was assumed to be zero at pulse onset [2]. This meant that the tangent at pulse onset must be zero, *i.e.* $s_p(P) = s_p(0)$. $s_p(P)$ was the last sample of return phase model (7) and $s_p(0)$ the first sample of open phase sub-model (6). $s_p(1)$ was arrived at by means of equation (6):

$$s_p(1) = a_0 + a_1 s_p(0) + a_2 s_p(P).$$

The initialization of return phase sub-model (7) was carried out in the following manner: $s_p(0) = b_0 + b_1 z$. Sample z was obtained as follows from the last sample of the open phase segment. Since the observed glottal pulses were noisy, the natural samples could not be used as such. Instead, z was obtained by projecting the triplet $[s_p(Q-2),\ s_p(Q-1),\ s_p(Q)]$ onto the plane defined by equation (6) in the space $[s_p(n-2),\ s_p(n-1),\ s_p(n)]$. Sample $s_p(Q)$ was the last natural sample of the open phase segment. The result of the projection was a synthetic triplet $[x,\ y,\ z]$ [1].

Hereafter we summarize two algorithms. The first was used to fit model (6) and (7) to glottal pulses obtained by glottis inverse filtering. The second was used to estimate the formant frequencies of the speech signal excitation-synchronously.

The steps for fitting model (8) to glottal waveforms were the following.

(*i*) Excitation-asynchronous positioning of the analysis window.

(*ii*) Initialization of the values of threshold r and delay d.

(*iii*) Segmentation of the glottal waveform according to expressions (6) and (7).

(*iv*) Fitting of sub-models (6) and (7) by means of singular-value-decomposition [6].

(*v*) Calculation of initial condition $s_p(0)$ of sub-model (7) by

means of a projection of a triplet of natural samples onto plane (6).

(*vi*) Synthesis of the glottal pulse by means of model (6) & (7) and initial condition $s_p(0)$ of sub-model (7).

(*vii*) Calculation of the total modeling error by means of a sample-to-sample comparison of the synthetic and natural pulses.

(*viii*) Choice by an optimizer of new values for threshold r and delay d on the basis of the modeling error. Move to (*iii*) when the error is not at a minimum, otherwise move to (*ix*).

(*ix*) Re-synthesis of the glottal pulse by means of model (6) & (7) and initial condition $s_p(0)$ of model (7).

The steps required for estimating formant frequencies excitation-synchronously were the following.

(*i*) Excitation-asynchronous positioning of the analysis window.

(*ii*) Smoothing of the speech signal inside the analysis window to obtain auxiliary signal $w(n)$.

(*iii*) Initialization of the values of threshold r and delay d.

(*iv*) Segmentation of the speech signal by means of auxiliary signal $w(n)$ according to relations (10) & (11).

(*v*) Fitting by means of singular-value-decomposition [6] of sub-models (10) and (11) to the segments obtained during the previous step.

(*vi*) Computation of the normalized total prediction error.

(*vii*) On the basis of this error, choice by an optimizer of new values for threshold r and delay d. Move to (*iv*) when the error is not at a minimum, otherwise move to (*viii*).

(*viii*) Re-estimation of the values of coefficients c_i by means of an *LPC* multi-interval covariance method.

(*ix*) Calculation of the formant frequency values from the roots of polynomial $\sum_{i=0}^{N} c_i z$.

V. RESULTS

The lower part of Figure 3 shows a voice source signal (dotted line) obtained by glottis inverse filtering from a sustained vowel [a] and the corresponding synthetic signal (continuous line) arrived at by means of model (8) proposed in this article. The upper part of Figure 3 shows the segments of the glottal pulse obtained in the final analysis and to which models (6) and (7) were fitted. Segmentation and fitting were carried out from within an excitation-asynchronously positioned analysis window without any preprocessing or intervention from the outside. The glottis signal displayed in Figure 3 was obtained by glottis inverse filtering by means of formant frequency and bandwidth estimates arrived at by the excitation-synchronous formant extraction method proposed in this article.

Figure 4 shows the speech signal and formant trajectories of the sentence [(ɑlaˈ)ɑtlǝmetejeãsilɛsɛrekɔ(lt)]. Formants displayed were obtained during the return phases arrived at by means of signal model (10) & (11). Formants the bandwidths of which were greater than 500 *Hz* were omitted from the display. No other kind of post-processing was carried out. It is seen that the formant trajectories obtained were plausible and that omission and insertions were infrequent.

VI. DISCUSSION AND CONCLUSION

(*i*) The analysis methods based on *SETAR* models of the glottal pulse and the speech signal were totally automatic and operated within excitation-asynchronously positioned analysis windows.

(*ii*) Glottal pulse model (6) & (7) and speech signal model (10) & (11) is formally identical. This is satisfying insofar that the speech signal is obtained by linearly filtering a source signal. Indeed, a linear system does not add frequency components to or subtract frequency components from the source signal and does not alter the number or the type of solutions of the difference equations that generate the source signal.

(*iii*) Results obtained so far show that the quality of the formant frequency estimates arrived at by means of the *SETAR* speech signal model was at least similar to the quality of the results obtained by other methods [3].

(*iv*) Two different linear methods were used for fitting linear sub-models: singular-value-decomposition and an *LPC* multi-interval covariance method. The former managed exceptions (*i.e* determinants close to zero, a feeble number of data, etc.) more effectively, and the latter obtained more robust formant estimates.

References

[1] N. Efimov. *Elements de géometrie analytique.* Mir, Moscow, 1966.

[2] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of the glottal flow. *STL-QPSR*, 4:1–13, 1985.

[3] A.K. Krishnamurthy and D.G. Childers. Two-channel speech analysis. *IEEE Trans. acoust. speech and sign. proc.*, ASSP-34(4):730–743, 1986.

[4] J.D. Markel and A.H. Gray. *Linear Prediction of Speech.* Springer-Verlag, New-York, 1976.

[5] A.V. Oppenheim and A.S. Willsky. *Signals and Systems.* Prentice Hall, Englewood Cliffs, 1983.

[6] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes - The Art of Scientific Computing.* Cambridge University Press, New York, 1987.

[7] H. Tong and K.S. Lim. Threshold autoregression, limit cycles and cyclical data. *J. Roy. Statist. Soc.*, B(42):245–292, 1980.

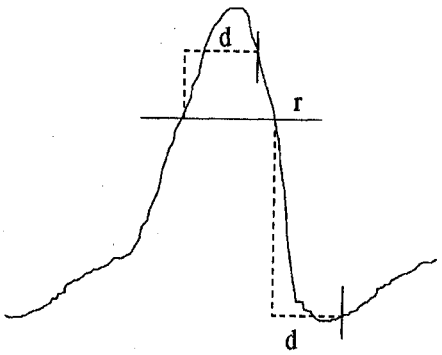

Figure 1:

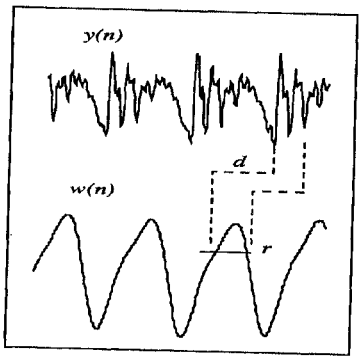

Figure 2:

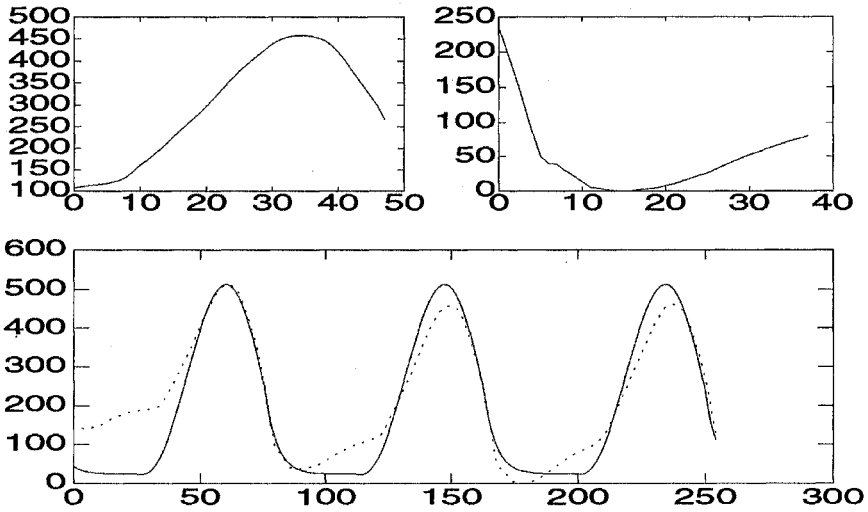

Figure 3:

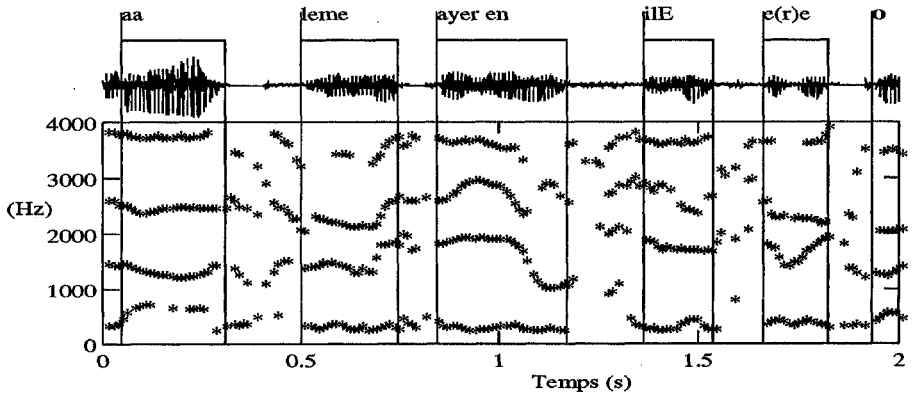

Figure 4: