# VOICED SPEECH SYNTHESIS USING PITCH ASYNCHRONOUS CODE EXCITED LINEAR FILTERS FOR THE GLOTTAL SOURCE

**Arun Kumar** [(1)], **Sandeep Kumar** [(2)] **and Indra Narayan Kar** [(3)]

*[(1)]Centre for Applied Research in Electronics, Indian Institute of Technology, Delhi, Hauz Khas, New Delhi – 110 016, India.* arunkm@care.iitd.ernet.in

*[(2)] As (1) above.*

*[(3)]Department of Electrical Engineering, Indian Institute of Technology, Delhi, Hauz Khas, New Delhi – 110 016, India.* ink@ee.iitd.ernet.in

## ABSTRACT

This paper proposes a model for natural quality voiced speech synthesis using code excited linear all-pole filter for modeling the glottal source signal. Classical glottal signal models are explicit-time functions which inhibit joint source-tract parameter estimation and require pitch synchronous estimation with precise segmentation of open and closed glottis phase. These problems are overcome in the proposed implicit-time glottal model. It is found that a switched, code excited linear all-pole filter for the glottal signal gives near natural quality voiced speech synthesis. Both objective and subjective performance results will be presented.

## INTRODUCTION

This paper presents objective and subjective performance analyses results of code excited linear all-pole filters in modeling the glottal source signal for natural quality voiced speech synthesis. The basis of speech synthesis is the universal speech production model comprising of a time-varying linear all-pole filter excited by a source signal [1]. It is generally considered that unvoiced speech of "equivalent" perceptual quality as voiced speech can be synthesized under the same bandwidth constraint. Thus, the research interest focuses on the problem of natural quality voiced speech synthesis. The efficient modeling of voiced speech is important in the context of natural quality speech synthesis, low bit rate speech coding, and several speech analyses problems [2], [3].

Classically, glottal signal models used in these problems are explicit-time functions [4]. They have the following generic limitations: (a) pitch synchronous parameter estimation is needed, which in turn requires precise segmentation of open and closed phases of the glottis, (b) there is nonlinear dependence of model parameters on the signal which has bearing on estimation complexity, and, (c) it is difficult to combine an explicit-time function voiced source model with a standard implicit-time function synthesis filter for more accurate joint estimation of source and tract parameters which may account for source-tract interactions in voiced speech production. We attempt to overcome these problems of classical glottal models with the design of pitch asynchronous code excited linear all-pole filters that are implicit-time functions.

## PRIOR ART

Several explicit-time glottal models have been proposed for diverse speech processing problems. In the context of voiced speech synthesis, glottal models are either non-interactive, or interactive according to the absence or presence of interaction between the glottal source and vocal tract, in the estimation of model parameters [4]. Some examples of non-interactive glottal models are Rosenberg's trigonometric and polynomial models and Liljiencrants and Fant model, while Ishizaka and Flanagan's "mechanical" model is an often used interactive glottal model.

Schoentgen [5] proposed the homogeneous switched affine glottal model for voiced speech synthesis. The explicit-time function glottal model due to Liljencrants and Fant [6], given by,

$$g_o[n] = A_1 K_1^n \cos(\omega n) + C_1 \quad (1)$$

$$g_c[n] = A_2 K_2^n + C_2 \quad (2)$$

is a solution of the homogeneous switched affine glottal model which is an implicit-time function model. Here, (1) and (2) give the glottal waveform models for the open and closed phase of the glottal cycle respectively, and $A_1, A_2, C_1, C_2, K_1, K_2, \omega$ are the model parameters. Schoentgen's model consists of two sub-models:

$$g[n] = a_0 + a_1 g[n-1] + a_2 g[n-2], \qquad g[n-d] < r \tag{3}$$

$$g[n] = b_0 + b_1 g[n-1]. \qquad g[n-d] \geq r \tag{4}$$

where, *g[n]* is the glottal signal, and $a_0, a_1, a_2, b_0, b_1, d, r$ are model coefficients which are estimated on a frame-by-frame basis. Within the frame, switching takes place between the two sub-models according to the given criterion. This model has the advantages of an implicit-time glottal model but it still does not produce natural quality voiced speech. Kumar and Gersho [7] generalized this model to an input driven switched affine model that significantly improves the voiced speech synthesis quality. In this paper, we further improve upon the performance of this model with the design of code excited linear all-pole filters for the glottal source signal.

**VOICED SPEECH SYNTHESIS METHOD USING CODE EXCITED LINEAR GLOTTAL FILTER**

Figure 1 shows the block diagram of the proposed voiced speech synthesis method. Linear prediction analysis-synthesis is used for synthesizing voiced speech. For purposes of studying the glottal source models, only frames corresponding to voiced speech are processed for source signal modeling, while unvoiced frames are concatenated at the synthesized output to facilitate perceptual analysis. The voiced speech signal is analyzed pitch asynchronously on a 20 ms frame basis to produce the vocal tract and glottal model parameters. The glottal signal *g[n]* is obtained by inverse filtering the speech signal *s[n]*, followed by integration. A first order integrator with $a$=0.98 is used to obtain *g[n]*.

The integrated LP residual *g[n]* is modeled by code excited linear all-pole filter. The filter types that we propose, can be classified according to the update schemes which are: (i) block adaptive, and, (ii) threshold switched. In the first update scheme, the autocorrelation method is used to estimate the filter parameters once per frame.

The threshold switched linear all-pole model is given by:

$$g[n] = \sum_{i=1}^{P_1} a_i g[n-i] + e[n], \qquad \text{median}^{l}_{k=-l}\{g[n-d-k]\} > r \tag{5}$$

$$g[n] = \sum_{i=1}^{P_2} b_i g[n-i] + e[n]. \qquad \text{median}^{l}_{k=-l}\{g[n-d-k]\} \leq r \tag{6}$$

where, $P_1$ and $P_2$ are the respective sub-model orders, and the threshold parameter $r$ and delay $d$ determine the switching instants between the sub-models. The model parameters $a_1,\ldots,a_{P1}$, $b_1,\ldots,b_{P2}$, $r$ and $d$ are estimated by minimizing $E = \Sigma e^2[n]$ over the frame length. The stabilized covariance method is used for estimating the filter parameters. This ensures switching between two stable linear filters depending upon the switching parameters. To prevent random switching between the sub-models, a smoothed switching criterion based on (2$l$+1)-point median filtering, as given by (5) and (6), is used. The range of delay $d$ is restricted to integer values less than 0.3 times the pitch period. This is because a larger range of delay does not improve the subjective quality of synthesized speech. The threshold parameter $r$ is quantized with an adaptive codebook, where the adaptation is done according to the signal energy of the previous frame.

The glottal filter excitation signal is a gain scaled code vector. Two types of codebooks were studied, namely, sparse independent identically distributed (i.i.d.) Gaussian codebook and algebraic codebook [8]. The excitation parameters comprising of the code vector index and scalar gain factor are estimated using closed loop weighted minimum mean square error criterion, where the loop is closed either at the glottal signal domain as shown in figure 1, or at the speech signal domain. The excitation parameters are computed at a faster sub-frame rate. There are four sub-frames per frame. The modeled integrated LP residual signal is filtered through a cascade of a comb filter, a differentiator and all-pole synthesis filter (based on the interpolated LP polynomial) to obtain the synthesized voiced speech signal.

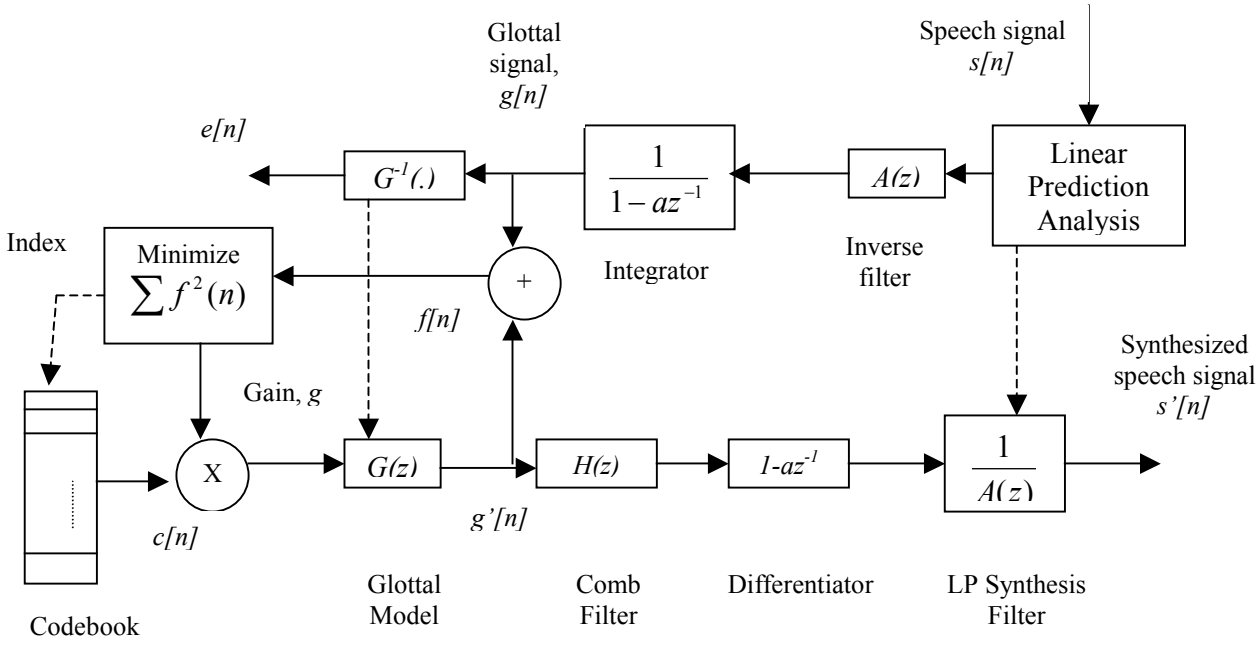


**Figure 1:** Block diagram representation of voiced speech synthesis method.

**RESULTS AND OBSERVATIONS**

The performance of code excited linear all-pole filters for glottal source was analyzed in three ways: (i) using objective measures, (ii) through informal listening tests of the synthesized speech, and (iii) by comparative listening tests with Rosenberg's trigonometric glottal model [4] based voiced speech synthesis. The experiments were performed using 10 minutes of speech from the TIMIT database. Figure 2 gives the signal-to-prediction error (SPER) ratio for the linear all-pole filter for glottal signal. It can be seen that the threshold switched parameter update method gives about 2 dB more SPER compared to the block adaptive update method. Figure 3 gives the SNR for different codebook sizes for the sparse independent identically distributed (i.i.d.) Gaussian random codebook.

The important observations of the performance analyses are as follows: (i) the code excited switched all-pole filter model gives near transparent quality synthesis; (ii) the constant terms of the switched affine model in (3), (4) are not important perceptually; (iii) third order polynomial sub-models in switched all-pole filters provide optimum combination of model order, perceptual fidelity and signal to prediction error ratio (SPER); (iv) the smoothed switching criterion in the switched all-pole filter model based on median filtering improves synthesis quality for filter order equal to three or more; (v) SPER of block adaptive all-pole filter saturates at approximately 2 dB below the SPER of switched all-pole filter as seen from figure 3; (vi) the perceptual quality of synthesized speech using block adaptive filter is "slightly inferior" to that obtained from the switched all-pole filter model; (vii) both the filter models perform significantly better than and are always preferred compared to Rosenberg's glottal model based synthesis; (viii) when closed-loop analysis is done at the speech signal domain, it becomes feasible to include auditory masking properties in the form of perceptually weighted mean square error minimization to enhance synthesized speech quality; (ix) in the threshold switched updated scheme, switching takes place between two stable sub-models which results in a time varying system. This does not guarantee the stability of overall time varying system. However, it is observed that the synthesized signal does not diverge over the entire database; (x) an algebraic type codebook with 4 pulses per 5 ms sub-frame gives better subjective performance compared to a sparse Gaussian distributed i.i.d. codebook of size 1024.

The proposed pitch asynchronous linear all-pole filters for glottal signal modeling overcome the limitations of explicit-time function glottal models while providing near transparent quality voiced speech synthesis. We propose the use of linear all-pole models for the glottal source because of perceptual redundancy of affine models. In doing so, we are able to incorporate great versatility in the analysis objectives. These factors contribute to significant potential for practical deployment in natural quality voiced speech synthesis, low bit rate coding and speech analysis tasks.

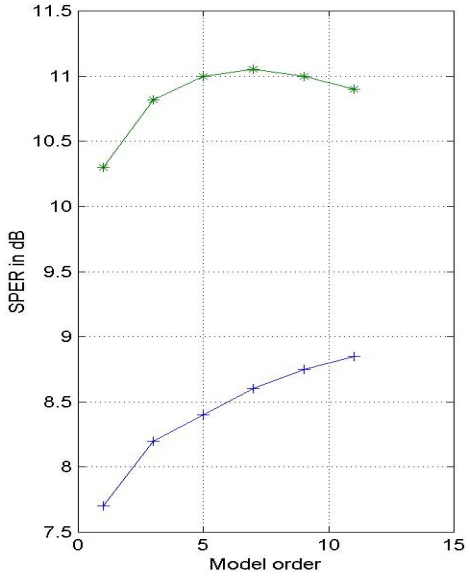


**Fig. 2:** Signal-to-prediction error ratio as a function of model order for linear all-pole model for the glottal signal. The '+' connected line is for block adaptive filter. The '*' connected line is for switched all-pole filter.

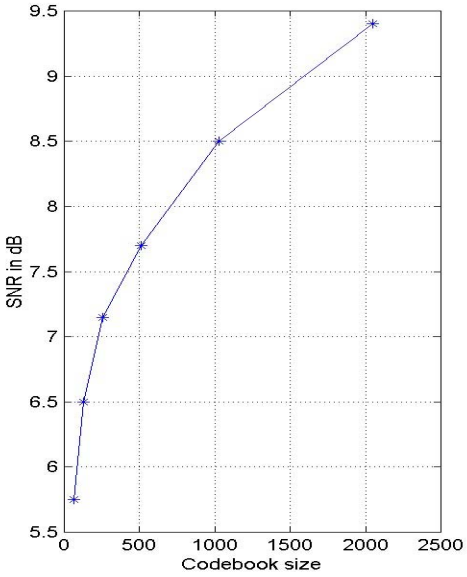


**Fig. 3:** Signal-to-noise ratio for the switched all-pole filter output as a function of the vector codebook size of the excitation. Here, $P_1=P_2=3$. See (5), (6).

The block adaptive all-pole source filter can be combined with the cascade of comb filter, differentiator and all-pole synthesis filter for joint parameter estimation that can model source-tract interactions as well. Also, by removing the constant terms in the model and making the source filter linear, it becomes possible to incorporate structured codebooks such as an algebraic codebook that can further improve coding efficiency. In the context of (ix) above, although the synthesized speech does not diverge, the theoretical proof of the stability of a switched time varying system remains an open problem.

## REFERENCES

[1] D. O'Shaughnessy, *Speech Communcations: Human and Machine*, 2nd ed., John Wiley, 2000.
[2] P. Hedelin, "High quality glottal LPC vocoding," *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, Tokyo, pp. 465-468, 1986.
[3] J. Linden, L. Skoglund and P. Hedelin, "Low rate speech coding using a glottal pulse codebook," *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, pp. 105-106, 1995.
[4] K. E. Cummings and M. A. Clements, "Glottal models for digital speech processing: A historical survey and new results," *Digital Signal Processing*, Vol. 5, no. 1, pp. 21-42, 1995.
[5] J. Schoentgen, "Self excited threshold autoregressive models of the glottal pulse and the speech signal," *Proc. Intl. Conf. on Spoken Language Processing*, pp. s19-9.1-s19-9.4, 1994.
[6] G. Fant, J. Liljencrants and Q. Lin, "A four parameter model of glottal flow," *Speech Transmiss. Lab. Q. Prog. Status Rep.*, pp. 1-14, 1985.
[7] A. Kumar and A. Gersho, "Voiced speech synthesis using threshold autoregressive glottal model," *Proc. IEEE Speech Coding Workshop*, Pennsylvania, 1997.
[8] J. –P. Adoul, P. Mabilleau, M. Delprat and S. Morissette, "Fast CELP coding based on algebraic codes," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1957-1960, April 1987.