VOICED SPEECH SYNTHESIS WITH A NONLINEAR GLOTTAL MODEL*

Arun Kumar and Allen Gersho*

Department of Electrical & Computer Engineering
University of California, Santa Barbara, CA 93106 USA
gersho@ece.ucsb.edu; http://scl.ece.ucsb.edu

ABSTRACT

We propose a model for voiced speech synthesis using a threshold autoregressive (TAR) glottal flow model. It reconstructs the pitch periodicity and tracks small pitch period variations even at a low parameter update rate of 20 ms, without requiring explicit pitch information. The model overcomes several limitations of traditional glottal models and has potential for application to high quality speech synthesis and low bit rate speech coding.

1. INTRODUCTION

We have investigated a novel scheme for voiced speech synthesis using a code excited *implicit-time* glottal flow model. The pitch periodicity is reconstructed and tracked by the glottal model *without requiring explicit pitch information.* Previous studies have shown that shaping the excitation signal according to a glottal model in source-filter type of speech production/synthesis models helps increase the naturalness of synthesized speech, particularly when low bit rate coding is the target [1], [2]. Yet there has been only a limited study of glottal model based speech synthesis and coding schemes. This is because traditional noninteractive glottal flow models e.g., Rosenberg's model, Liljencrants - Fant (LF) model etc. [3] are explicit nonlinear time functions leading to (a) requirement of pitch synchronous parameter estimation and precise segmentation of the glottal signal into open and return phase of the glottis, (b) nonlinear dependence of the waveforms on model parameters, and (c) difficulty in combining explicit time function models of the glottal signal with linear synthesis filter for frame synchronous or joint parameter estimation.

It was shown by Schoentgen [4] that the Liljencrants-Fant model given by submodels:

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

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

for open and return glottis phase respectively (for each pitch period interval) are solutions of linear second order and first order difference equations respectively. He proposed a scheme for automatic switching between the two submodels. Based on this approach, we propose a threshold autoregressive (TAR) model to approximate the glottal signal obtained by inverse filtering and apply it to voiced speech synthesis.

2. PERFORMANCE OF TAR GLOTTAL MODEL

The glottal signal, $g(n)$, is obtained by inverse filtering the speech signal, $s(n)$, followed by an integration (see fig. 1). The LP and glottal flow model parameters are estimated once every frame of length 20 ms. We use a 1st-order integrator with $a = 0.98$ to obtain $g(n)$. The TAR glottal model is given by [4]:

$$g(n) = a_0 + \textstyle\sum_{i=1}^{P_1} a_i g(n-i) + e(n), \quad g(n-d) > r, \quad (3)$$

$$g(n) = b_0 + \textstyle\sum_{i=1}^{P_2} b_i g(n-i) + e(n), \quad g(n-d) \leq r \quad (4)$$

where P_1 and P_2 are orders of the submodel given by equations (3) and (4) respectively and the threshold parameter r and delay d determine the switching instant between the submodels. To make the switching robust to noise and prevent random switching we use a *smoothed* switching criterion by checking for r against $\sum_{i=-k}^{k} g(n-d-i)$, where $k = 1$ is typically adequate.

The model parameters $a_0, \ldots, a_{P_1}, b_0, \ldots, b_{P_2}, r$ and d are estimated by minimizing $\sum_{n=0}^{N-1} e^2(n)$ where $e(n)$ is the prediction error from equation (3) and (4), and $N = 160$, corresponding to 20 ms frame sampled at 8 kHz. The delay d is restricted to integer values between 1 and 20. The threshold parameter r is quantized with an adaptive codebook, where the adaptation is done according to the energy of previous frame's synthesized signal. For $P_1 = P_2 = 3$, the segmental prediction gain varies from 10.58 dB to 11.65 dB for 1 to 6 bit quantized representation of r. The segmental prediction gain is found to be 5.73 dB, 10.61 dB, 11.53 dB, 11.96 dB, 12.57 dB and 13.15 dB for $P_1 = P_2 = 0, 1, 3, 5, 8$ and 10 respectively. We use $P_1 = P_2 = 3$ and 4 bit quantized representation of r in the design of the voiced speech synthesis model.

*The first author is now with Center for Applied Research in Electronics (CARE), Indian Institute of Technology, Delhi, India. This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, ACT Networks, Advanced Computer Communications, Stratacom, DSP Group, DSP Software Engineering, Fujitsu, General Electric Company, Hughes Electronics, Intel, Moseley Associates, National Semiconductor, Nokia Mobile Phones, Qualcomm, Rockwell International, and Texas Instruments.

0-7803-4073-6/97/$10.00©1997IEEE.

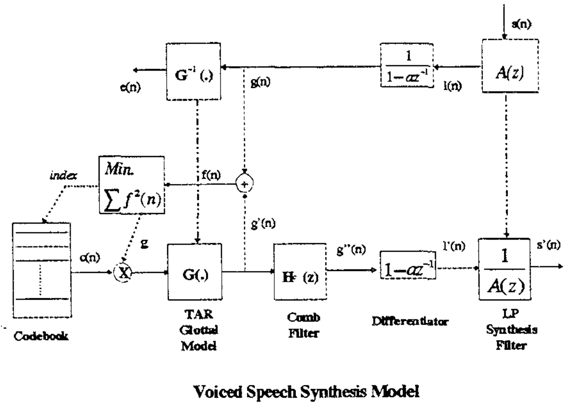

Figure 1: Voiced speech synthesis using TAR glottal model.

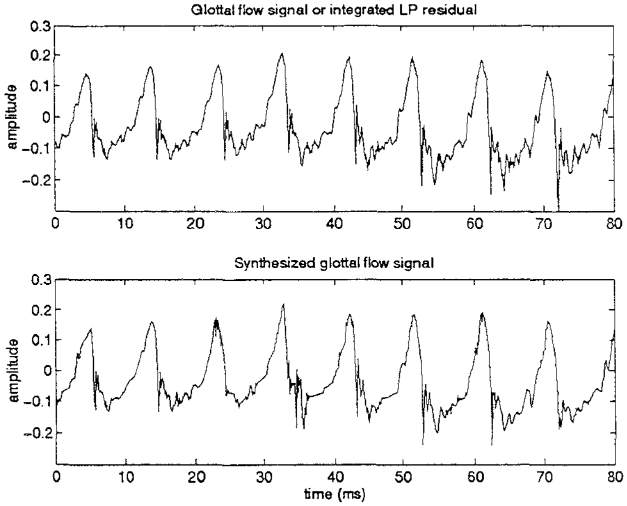

Figure 2: Glottal signal and its synthesis.

3. VOICED SPEECH SYNTHESIS MODEL

Figure 1 gives the block diagram description of the TAR glottal model based voiced speech synthesis scheme. The LP analysis and glottal model parameter estimation are done open loop once every 20 ms. The codebook index and gain g are estimated every subframe of 2.5 ms, by minimizing the m.s.e. between $g(n)$ and $g'(n)$, where

$$g'(n) = a_0 + \sum_{i=1}^{P_1} a_i g'(n-i) + gc(n), \quad g'(n-d) > r \qquad (5)$$

$$g'(n) = b_0 + \sum_{i=1}^{P_2} b_i g'(n-i) + gc(n), \quad g'(n-d) \leq r. \qquad (6)$$

The closed loop analysis is done in the "glottal" domain. We use an 8-10 bit center-clipped i.i.d. uniform noise codebook with 70% sparsity. Incremental improvement in terms of SNR in the glottal domain and subjective speech quality is obtained successively with i.i.d gaussian noise, sparse i.i.d. gaussian noise, i.i.d. uniform noise and sparse i.i.d. uniform noise codebooks. Further, there is no perceptible degradation if overlapped codebook (with overlap delay = 2) is used instead of the respective non-overlapped codebook. For each codebook index, an initial gain estimate is obtained which is then refined using few iterations of gradient descent algorithm.

Figure 2 shows the glottal signal, $g(n)$, and its synthesized approximation, $g'(n)$, using a third order TAR model and 10-bit center-clipped i.i.d. uniform noise codebook with 70% sparsity. A noteworthy property is the ability of the synthesized glottal waveform, $g'(n)$, to follow small variations in pitch period within a framelength although the glottal model is estimated once per frame. Also, note that the pitch periodicity in the synthesized glottal and speech signals is reconstructed without explicit knowledge of the pitch value.

Another feature of the semi-closed loop index and gain estimation procedure is the use of weighted minimization. The indices corresponding to impulses in the LP residual signal, l(n), are weighted by a greater factor than the other indices. (Note that the weighted m.s.e. minimization is done in the glottal domain.) This weighting improves the energy match between the synthesized and original speech.

4. RESULTS AND CONCLUSION

The synthesized speech is close to natural quality and it appears promising that with further work, the proposed model for voiced speech synthesis will provide imperceptible distortion. There are no perceptible artifacts that can be attributed to pitch periodicity mismatch. Furthermore, the TAR model is capable of accurately reconstructing the pitch periodicity and tracking variations even though the model parameters are updated every 20 ms and no explicit pitch analysis is performed. While further research is needed, the model has potential for becoming a useful technique for speech coding.

5. REFERENCES

[1] P. Hedelin, "High quality glottal LPC vocoding," *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Tokyo, pp. 465–468, 1986.

[2] 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.

[3] K. E. Cummings and M. A. Clements, "Glottal models for digital speech processing: a historical survey and new results," *Digital Signal Processing*, vol. 5, pp. 21–42, 1995.

[4] J. Schoentgen, "Self excited threshold auto-regressive models of the glottal pulse and the speech signal," *Proc. Int. Conf. on Spoken Language Process.*, Yokohama, pp. S19-9.1 – S19-9.4, 1994.