Symp. Frontiers of Research on Speech and music (FRSM) - 2003, IIT Kanpur, Kanpur, India, Feb. 15-16, 2003, pp. 86-94

PERTURBATION IN GCI'S AND SPEECH QUALITY FOR PITCH SYNCHRONOUS SYNTHESIS

Parveen K. Lehana and Prem C. Pandey

Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai - 400076, India
{lehana, pcpandey}@ee.iitb.ac.in

ABSTRACT

In pitch synchronous speech synthesis the analysis/synthesis of the speech is done at each glottal closure instant (GCI). The errors in estimation of GCI's affect the quality of the synthesized speech. The effect of random perturbations in the GCI's, obtained from the speech and from glottal signal from an impedance electroglottograph using Childers and Hu's algorithm on the quality of speech synthesized using harmonic plus noise model (HNM), is investigated in this paper. Investigations show that the speech quality is very sensitive to positions of the GCI's. A small perturbation with maximum of 4 % of the local fundamental frequency considerably degrades the synthesized speech. Perturbations above 8 % severely affect quality of the out put speech. GCI's obtained from the glottal signal can afford slightly more perturbation as compared to the GCI's calculated from the speech signal.

1. INTRODUCTION

There are two methods of speech analysis/synthesis: pitch synchronous and pitch asynchronous. Pitch synchronous analysis/synthesis is well suited for high speech quality speech output. The problem of this method is that we need to estimate the glottal closure instants (GCI's) very accurately. GCI's can directly be calculated from the speech signal or they may be estimated using the pitch information. Different methods for the determination of pitch give different accuracies [1],[2].

Pitch perturbation is the cycle-to-cycle variability of the pitch or fundamental frequency. It is a measurement of how a given pitch value differs from the one or several pitch values that immediately precede or follow it [3]. Natural speech has a certain amount of perturbation or jitter. But errors in calculation of pitch also lead to perturbation in the pitch contour. Perturbation in the pitch leads to perturbations in the GCI's. If the speech is analyzed and synthesized using these inaccurate values of GCI's, the quality of the synthesized speech is affected. So it is necessary to employ a more accurate method for these calculations. Childers and Hu's algorithm can be used for GCI's estimation [4],[5].

For analysis/synthesis, harmonic plus noise model (HNM) is implemented. It is based on sinusoidal model of speech - the waveform is represented as summation of a finite number of sinusoids with arbitrary amplitudes, frequencies, and phases. There are many limitations of the sinusoidal modeling such as computation load and a big number of parameters [6]. Harmonic plus noise model (HNM) is a variant of sinusoidal modeling of speech and divides the spectrum of the speech into two sub bands, one is modeled with harmonics of the fundamental and the other is simulated using random noise [7]. HNM gives high quality speech with smaller number of parameters. Also, the pitch and time scaling is relatively easy [7].

The objective of this paper is to investigate the effect and the extent of perturbation in the GCI's on speech quality that is perceptually acceptable. GCI's are obtained by processing the speech signal using Childers and Hu's algorithm and from the glottal signal obtaining using an impedance electroglottograph [4],[5].

2. ANALYSIS / SYNTHESIS WITH HNM

HNM assumes the speech signal to be composed of two parts: harmonic, and noise. The harmonic part accounts for the quasi-periodic components of the speech signal while the noise part is responsible for non-periodic components (e.g., fricative or aspiration noise, period-to-period variation of the glottal excitation etc.). The frequency separating the two bands is called maximum voiced frequency F_m [7].

In the lower band, the signal is represented only by harmonically related sine waves with slowly varying amplitudes and frequencies:

$$s'(t) = \mathrm{Re} \sum_{l=0}^{L(t)} a_l(t) \exp\{ j[\int_0^t l\omega_o(\sigma)d\sigma + \theta_l]\} \quad (1)$$

$a_l(t)$ and $\theta_l(t)$ are the amplitude and phase at time t of the lth harmonic, ω_0 is fundamental frequency and $L(t)$ is the time-varying number of harmonics included in the harmonic part. The upper band, which contains the noise part, is modeled by an AR model and modulated by a time-domain amplitude envelope. The noise part, $n'(t)$, is therefore supposed to have been obtained by filtering a white Gaussian noise $b(t)$ by a time varying, normalized all-pole filter $h(\tau;t)$ and multiplying the result by an energy envelope function $w(t)$:

$$n'(t) = w(t)[h(\tau;t) * b(t)] \quad (2)$$

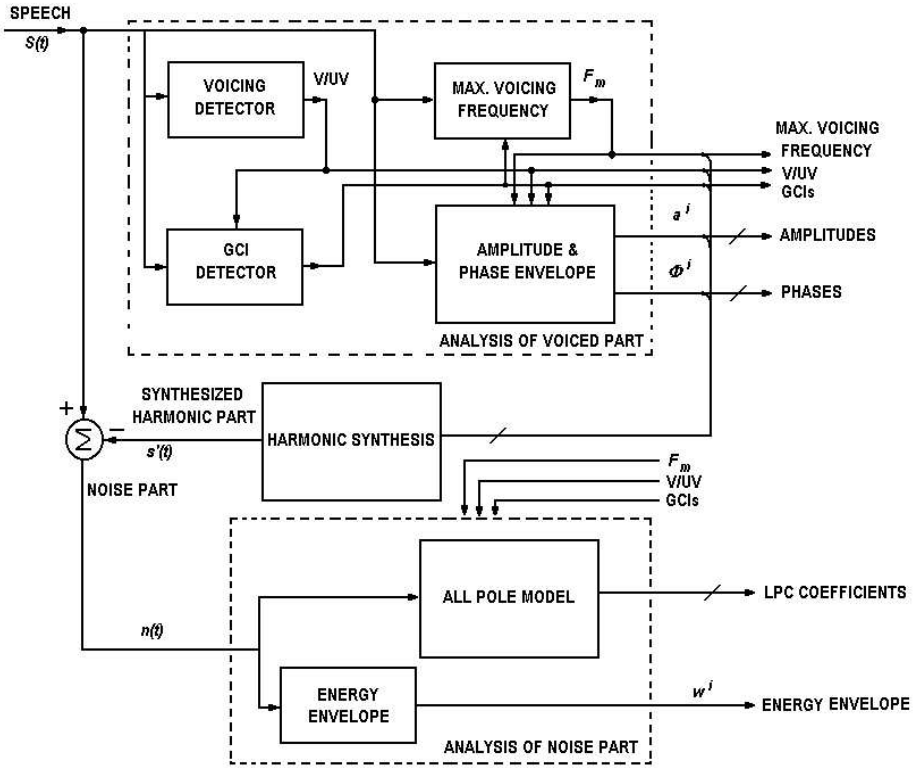


Figure 1. Analysis of speech using HNM, modified from [5].

Along with maximum voiced frequency F_m, the other parameters such as voiced/unvoiced, glottal closure instants, pitch, amplitudes, phases of harmonics of fundamental frequency (pitch), and parameters of noise part are calculated for each frame. Figure 1 shows a block diagram for analysis using HNM.

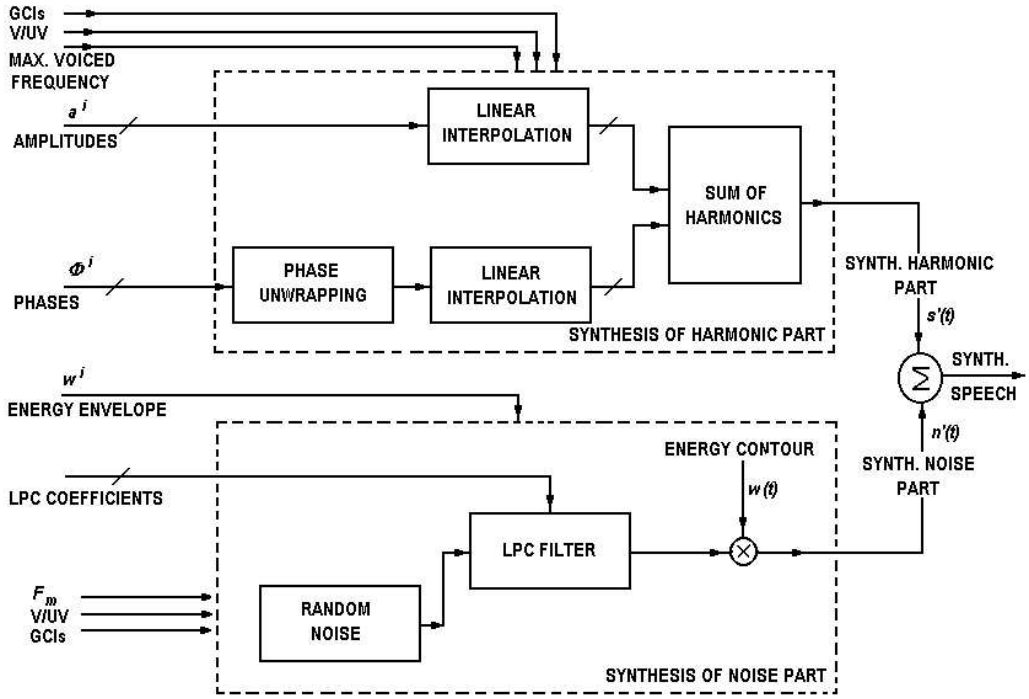


Figure 2. Synthesis of speech using HNM, modified from [5].

The speech signal is applied to the voicing detector, which declares the frames either voiced or unvoiced using the method proposed by Childers [8]. As the analysis and synthesis in HNM are pitch synchronous, it is necessary to calculate glottal closure instants (GCIs) accurately. Childers and Hu' s algorithm is used for these calculations [8], [4]. For comparison, we have also obtained GCIs from the glottal waveform recorded by using an impedance glottograph [9], [10]. First the glottal signal is bandpass filtered (100 Hz - 900 Hz) and given to the hysteresis comparator to convert the input into a rectangular waveform for calculating the period of the input signal. The thresholds of the hysteresis comparator are dynamically calculated using peaks, valleys, and average amplitude of the input [9]. Voiced part of the speech is obtained by retaining the voiced segments and inserting silence elsewhere. The analysis frame is taken twice the local pitch period and maximum voiced frequency is calculated for each frame. This voiced part of the speech is analyzed at each GCI for calculating amplitudes and phases of all the pitch harmonics up to the maximum voiced frequency.

For calculating the noise parameters, the synthesized voice part of speech is obtained using (1) and noise part is obtained by subtracting this from the original speech. Noise part is analyzed for obtaining LPC coefficients and energy envelop. The length of the analysis window for noise part is taken as two local pitch periods for both voiced and unvoiced frames. For voiced frames the local pitch is the pitch of the frame itself. For unvoiced frames the pitch of the last voiced frame is taken. Adding the synthesized harmonic and synthesized noise part gives the synthesized speech. It is to be noted that the harmonic synthesis is used in the analysis of the speech for obtaining the noise part as shown in Figure 2.

3. METHODOLOGY

The glottal closure instants (GCI' s) are first calculated for both glottal and speech waveforms using the Childers and Hu' s algorithm. Then random sequence with normal distribution in the range-[1, +1] is generated. This is multiplied by the local pitch and amount of desired relative perturbation in the GCI' s. Resulted value is added to the GCI under consideration. This process is repeated for all GCI' s. So each GCI is disturbed by an amount determined by the random number, local pitch, and relative of perturbation. Figure 3 shows the block diagram for this scheme.

After obtaining the perturbed GCI' s, the pitch synchronous analysis and synthesis is performed using HNM described in Figure 1 and Figure 2.

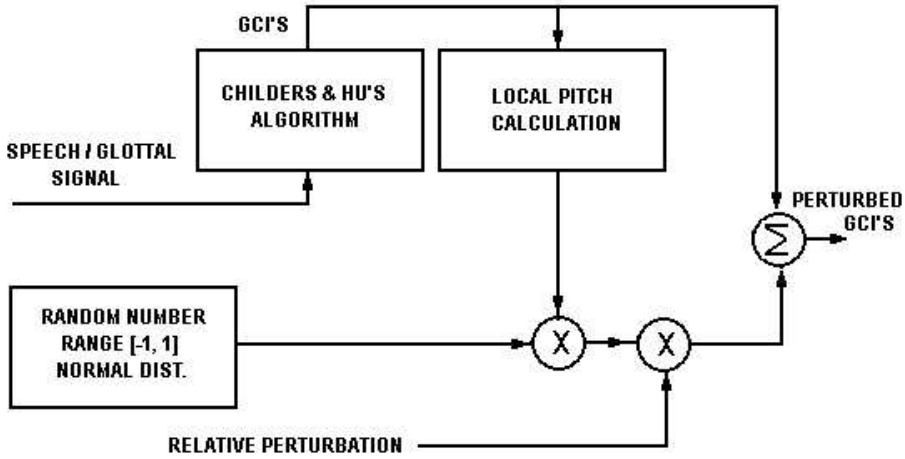


Figure 3. Calculation of perturbed GCI' s.

4. RESULTS

After implementing the HNM, speech segments corresponding to vowels in Indian languages were recorded, analyzed and synthesized. The recording was done using a B & K 4176 microphone (set up is shown in Figure 4). The microphone is connected to precision sound level meter (B & K, Type 2235) and its AC output is amplified and given to second channel of the line-in input of sound card. The first channel is used for recording glottal signal using an impedance glottograph. The recording was done in an acoustic room at a sampling rate of 10 k samples/s. Speech was recorded from both male and female speakers.

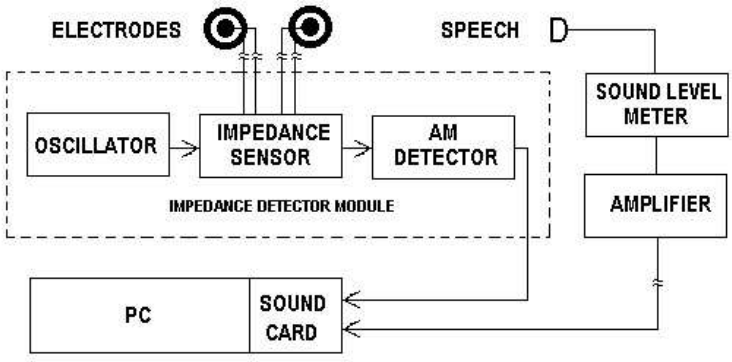
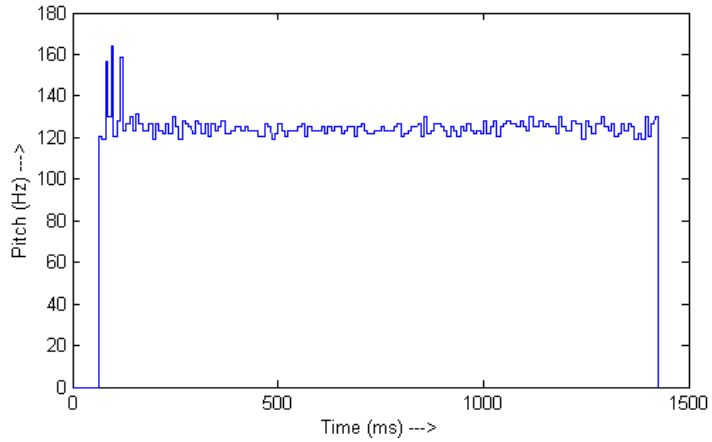


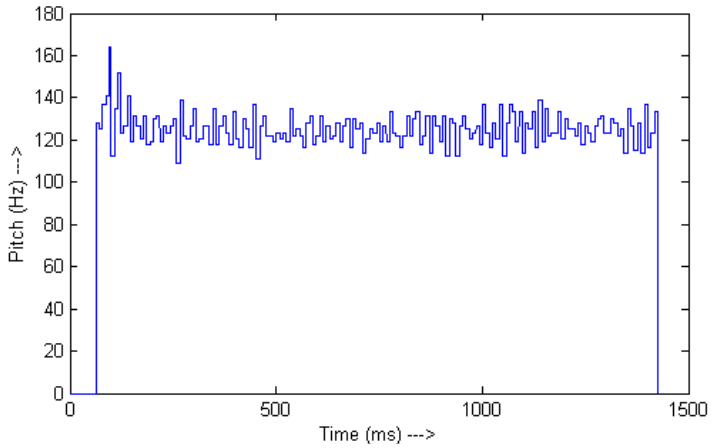
Figure 4. Set-up for recording, modified from [5].

GCI' s are obtained from the glottal and speech signals. Using the scheme shown in the Figure 3, perturbations in the GCI' s is added. The amount of perturbation is from 0 % to 20 % of the local pitch. Figures 5 and 6 show examples of pitch contours for different perturbations. Spectrograms of the vowel /*a*/ spoken by a male and female speaker, synthesized with different amount of perturbations, are shown in Figures 7 and 8 respectively.

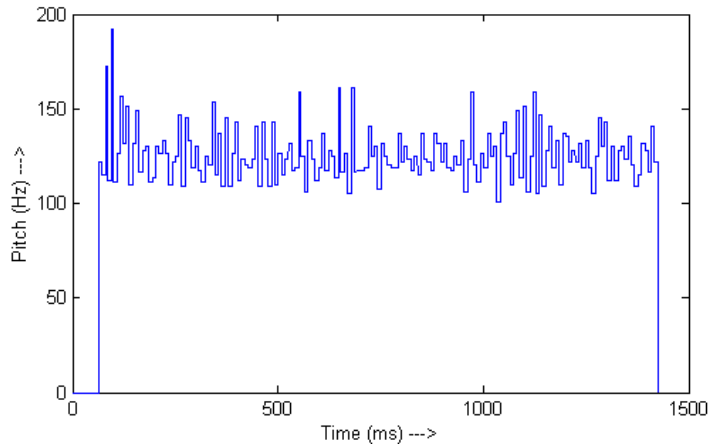
If the GCI' s are calculated from the speech signal and analysis/synthesis is performed with perturbation varying from 0 % to 4 % added to these GCI' s, the out put speech quality is not affected much as observed from the spectrograms (Figure 7c and 8c) and listening tests.



(a) 0 % perturbation

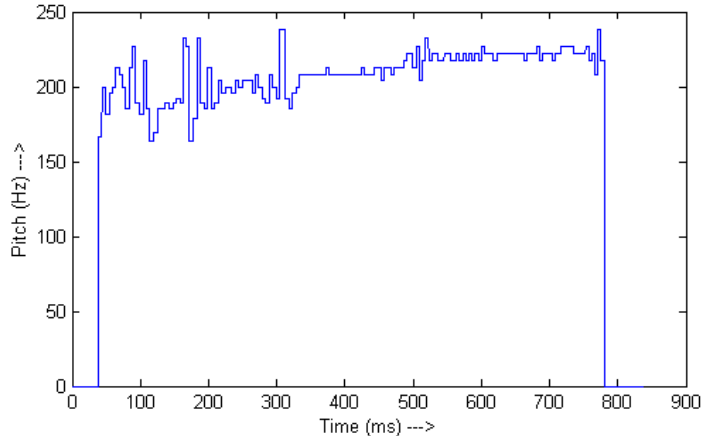


(b) 4 % perturbation

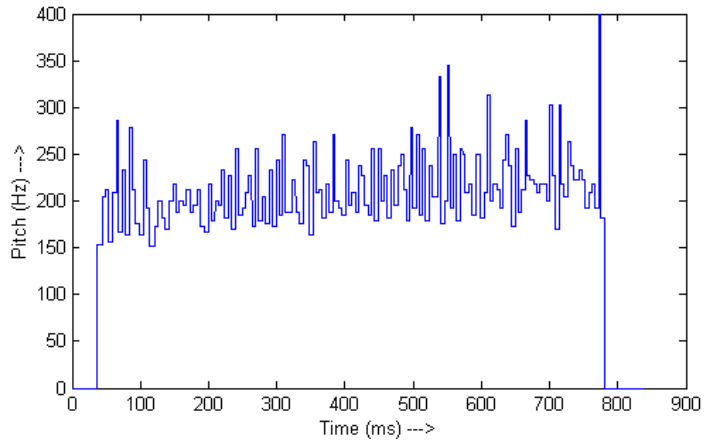


(c) 8 % perturbation

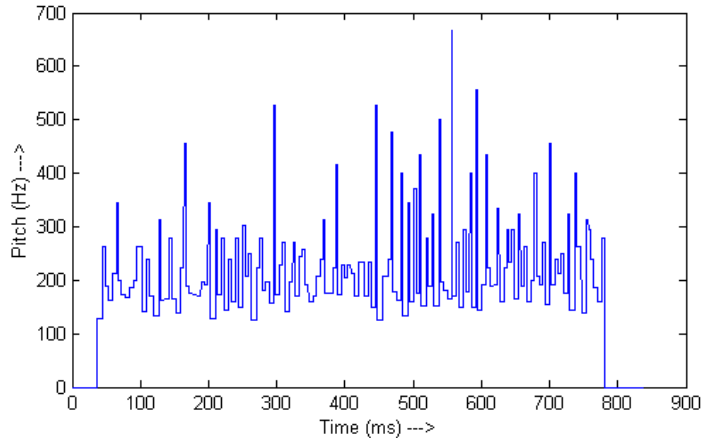
Figure 5. Pitch contours for /*a*/ spoken by a male speaker, with different amount of perturbations to the GCI' s obtained from the glottal signal.



(a) 0 % perturbation



(b) 4 % perturbation



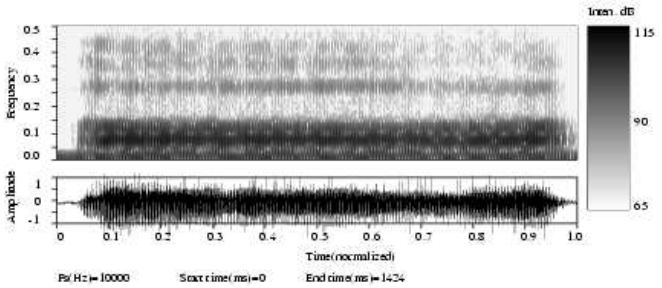
(c) 8 % perturbation

Figure 6. Pitch contours for /*a*/ spoken by a female after different amount of perturbations added to GCI' s obtained from the speech signal.

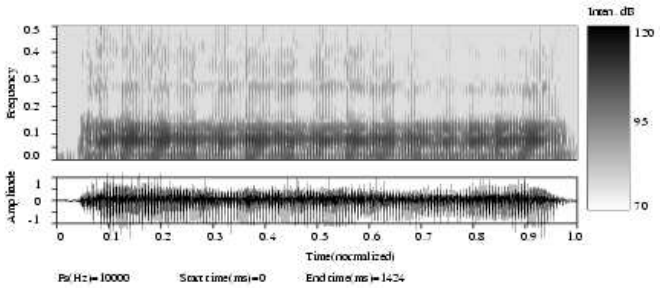
Although spectrograms show some noise at even 4 % perturbation but the output speech is perceptually satisfactory. When the perturbation is further increased from 5 % to 8 %, the speech

quality starts deteriorating and after 8 % it is severely affected (Figure 7d and 8d). These limits for perturbation are approximately the same for male and female speakers.

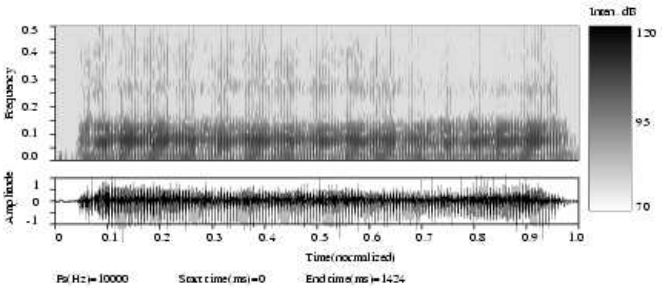
If the GCI's are obtained from glottal signal instead of obtaining from speech, the performance is slightly better. For example, the speech remains acceptable even for the values of perturbation as high as 6 %.



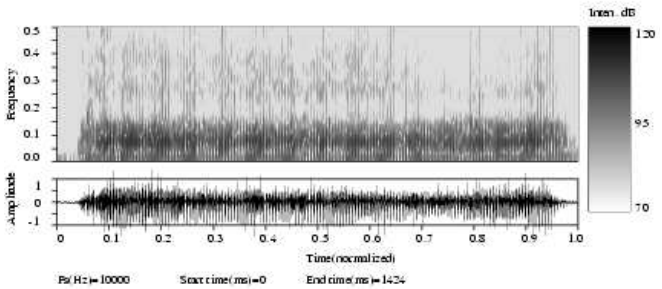
(a) Recorded



(b) 0 % perturbation

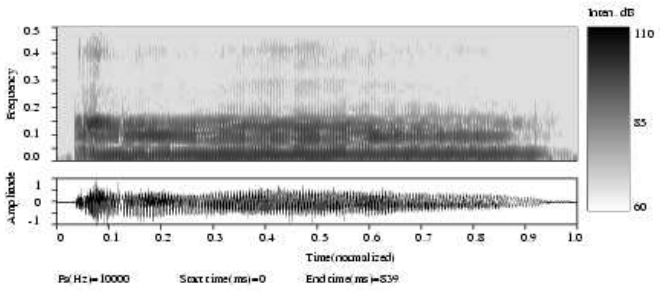


(c) 4 % perturbation

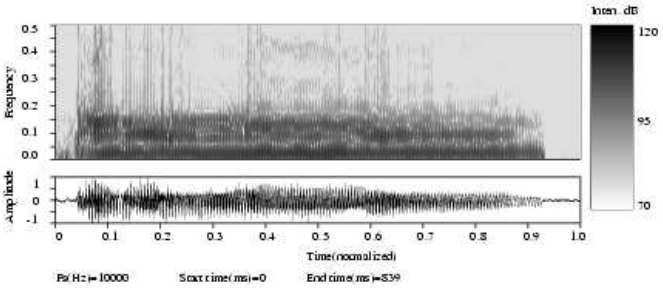


(d) 8 % perturbation

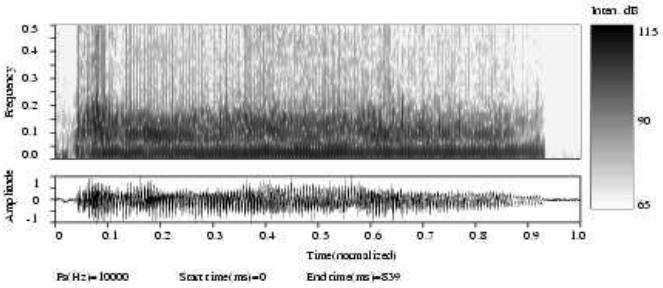
Figure 7. Spectrograms for /*a*/ spoken by a male speaker for different amount of perturbations added to GCI's obtained from glottal signal.



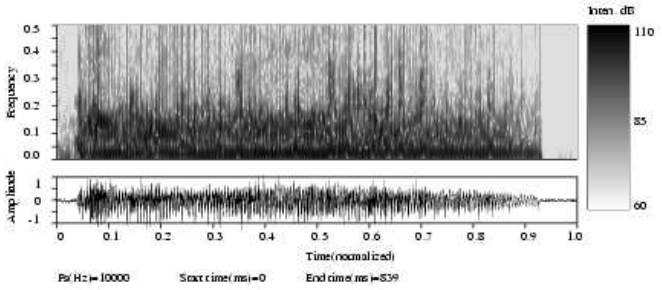
(a) Recorded



(b) Synthesized with 0 % perturbation



(c) Synthesized with 4 % perturbation



(d) Synthesized with 8 % perturbation

Figure 8. Spectrograms for /*a*/ spoken by a female speaker for different amount of perturbations added to the GCI' s obtained from speech signal.

5. CONCLUSIONS

Investigations with the pitch synchronous synthesis of vowels with different amount of perturbations in GCI' s, show that perturbations beyond 5 % result in quality degradation. Hence accurate

determination of GCI' s is necessary for high quality synthesis of speech, and a simultaneous recording of glottal signal from an impedance electroglottograph can be used for this purpose.

REFERENCES

[1] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. New York: John Wiley, 2002.

[2] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonic-to-noise ratio of a sampled sound," in *Proc. IFA*, University of Amsterdam, vol. 17, 1993, pp. 97-110.

[3] H. Kasuya, Y. Kobayashi, T. Kobayashi, and S. Ebihara, "Characteristics of pitch period and amplitude perturbations in pathologic voice," in *Proc. ICASSP*, 1983, pp. 1372-1375.

[4] D. G. Childers and T. H. Hu, "Speech synthesis by glottal excited linear prediction," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2026-2036, 1994.

[5] P. K. Lehana and P. C. Pandey, "Speech synthesis in Indian languages," in Proc. *Int. Conf. on Universal Knowledge and Languages-2002*, Goa, India, Nov. 25-29, 2002, paper: pk1510.

[6] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 4, pp. 744-754, 1986.

[7] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 21-29, Jan. 2001.

[8] D. G. Childers, *Speech Processing and Synthesis Toolboxes.* New York: John Wiley, 2000.

[9] P. K. Lehana and P. C. Pandey, "A low cost impedance glottograph and glottal pitch analyzer," in *Proc. Bio Vision 2001 Int. Conf. Biomed. Eng.*, Bangalore, India, Dec. 21-24, 2001, pp. 33-37, session 1B.

[10] M. Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, no. 1, pp. 36-43, 1992.