7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003), Orlando, USA, July 27-30, 2003

The Effect of SNR and GCI Perturbation on Speech Synthesis with Harmonic plus Noise Model

Parveen K. Lehana and Prem C. Pandey
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai - 400076, India
E-mail: {lehana, pcpandey}@ee.iitb.ac.in

ABSTRACT

Harmonic plus noise model (HNM) divides the speech spectrum in two bands: harmonic and noise. As most of the non-periodic components are removed in harmonic part, it may be expected that HNM synthesis is less susceptible to additive noise in input speech. As HNM analysis/synthesis is performed at each glottal closure instant (GCI), errors in estimation of GCIs affect the quality of the synthesized speech. Effects of the amount of additive broadband noise in input speech and perturbation in GCIs on the synthesized speech quality with special reference to phoneme sets in Indian languages were studied. Synthesis results show that for SNR in the 2-10 dB range, the quality of synthesized speech is superior to that of the input speech. Investigations also show that the speech quality is very sensitive to positions of the GCIs. Perturbations above 8 % severely affect quality of the output speech.

Keywords: Speech Synthesis, Glottal Closure Instants (GCIs), Harmonic Plus Noise Model (HNM), Electroglottogram (EGG).

1. INTRODUCTION

Different approaches for speech synthesis can be broadly classified in two categories: waveform and model based. Waveform based approach uses prerecorded phrases of speech and plays a subset of these phrases in a particular sequence for generating the desired speech output. This method provides high quality and naturalness, but can be used for limited vocabulary and usually only one voice. Model based synthesis approach uses a model of the speech generation such as articulatory, source-filter, or acoustic, and a set of parameters. In articulatory and source-filter approaches, the extraction of parameters is very difficult and time consuming. A very flexible acoustic model for speech synthesis is based on sinusoidal modeling of speech – the waveform is represented as summation of a finite number of sinusoids with arbitrary amplitudes, frequencies, and phases [1]. Harmonic plus noise model (HNM), a variant of sinusoidal model, requires smaller number of parameters and is computationally efficient [2]. Time and frequency scaling is also relatively easy.

HNM divides the spectrum of the speech in two sub bands: one is modeled with harmonics of the fundamental and the other is simulated using random noise. The frequency, which partitions the speech spectrum in these two bands, is known as maximum voiced frequency (F_m) and varies from frame to frame. In natural speech, the noise part is contributed by non-periodic components as fricative or aspiration noise and period-to-period variation of the glottal excitation, etc. As most of the non-periodic components from the speech are removed in harmonic part, it may be expected that synthesis from harmonic part is less susceptible to noise mixed in the input speech. Our earlier investigations [3] have shown that all the syllables in Indian languages except /*a*[illegible]/ and /*asa*/ can be synthesized with perceptually acceptable quality using only the harmonic part.

In HNM synthesis, the analysis/synthesis is performed at each glottal closure instant (GCI) [2], and hence it is a pitch synchronous technique. This is well suited for high quality speech output [4], but we need to estimate the glottal closure instants (GCIs) precisely. GCIs may be directly estimated from the speech signal or they may be estimated using the pitch information. Errors in pitch values depend on the estimation methods [5],[6]. 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 [7]. 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 GCIs. If the speech is analyzed and synthesized using these GCIs, the quality of the synthesized speech is affected.

We investigated the effect of the amount of additive broadband noise in speech and perturbation in GCIs on the synthesized speech quality that is perceptually acceptable, investigations are carried out with special reference to phoneme sets in Hindi and some other Indian languages. These phoneme sets have five places of articulation for stops. Aspiration is a distinguishing feature for stops, and there are only unvoiced fricatives [8]. GCIs are obtained by processing the speech signal using Childers and Hu's algorithm [9] and from the

electroglottogram (EGG) signal obtained using an impedance glottograph [10],[11].

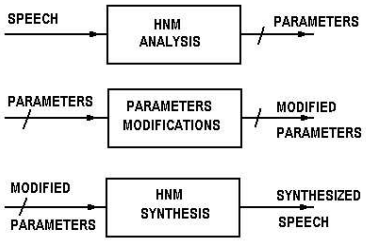


Fig. 1. Analysis/synthesis scheme, modified from [3].

2. HARMONIC PLUS NOISE MODEL

In HNM based synthesis, the speech signal is assumed to be composed of two parts: harmonic and noise

$$s(t) = s'(t) + n'(t) \tag{1}$$

The harmonic part accounts for the quasi-periodic components of the speech signal while the noise part is responsible for non-periodic components: 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 [2].

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

$$s'(t) = \mathrm{Re} \sum_{l=0}^{L(t)} a_l(t) \exp\{ j[\int_0^t l\omega_o(\sigma)d\sigma + \theta_l]\} \tag{2}$$

where $a_l(t)$ and $\theta_l(t)$ are the amplitude and phase at time t of the lth harmonic, ω_o is fundamental frequency and $L(t)$ is the time-varying number of harmonics included in the harmonic part. In the upper band, the signal is represented as noise $n'(t)$, modeled as a white Guassian noise $b(t)$ filtered by a time-varying all-pole normalized filter $h(\tau;t)$ and amplitude modulated by an envelope function $w(t)$

$$n'(t) = w(t)[h(\tau;t) * b(t)] \tag{3}$$

The analysis/synthesis scheme is shown in Fig. 1. Input speech is analyzed on frame-by-frame basis for calculation of maximum voiced frequency F_m and the other parameters such as voiced/unvoiced, glottal closure instants, pitch, amplitudes and phases of harmonics of fundamental frequency (pitch), and parameters of noise part. These parameters may be modified and output is generated using HNM synthesis.

Analysis scheme is shown in Fig. 2. Speech signal is applied to the voicing detector, which declares the frames either voiced or unvoiced. As the analysis and synthesis in HNM is pitch-synchronous, it is necessary to estimate glottal closure instants (GCIs) precisely. For each voiced frame the maximum voiced frequency (F_m) is calculated. The analysis frame is taken twice the local pitch period. This voiced frame of the speech is analyzed at each GCI for calculating amplitudes and phases of all the pitch harmonics up to F_m.

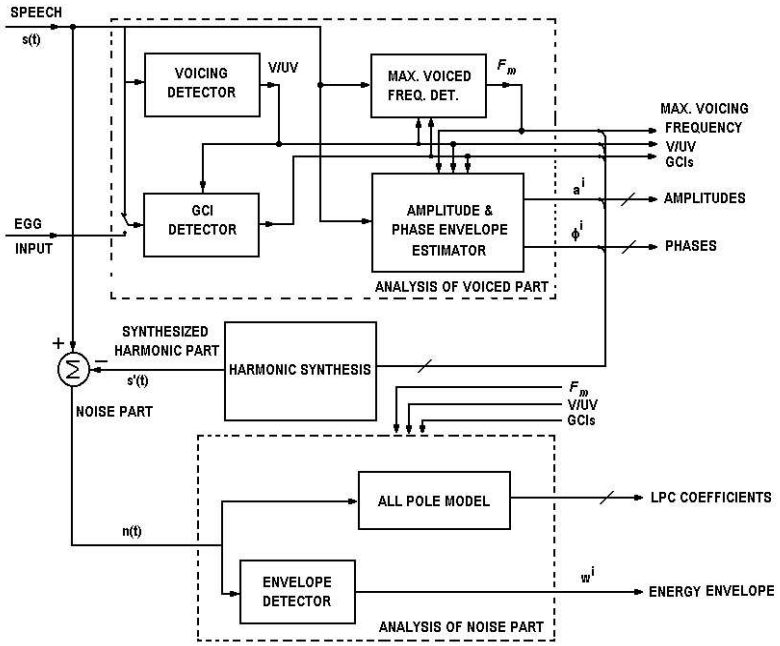


Fig. 2. Analysis of speech using HNM, modified from [3].

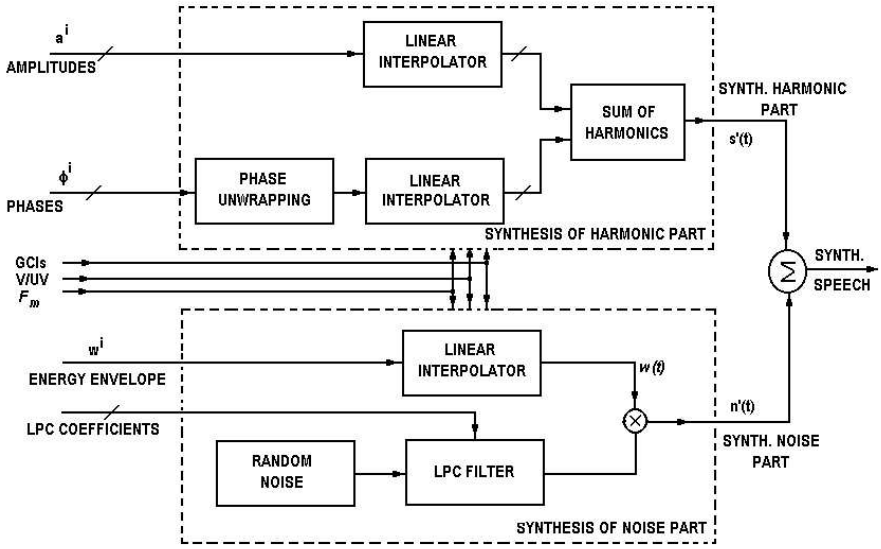


Fig. 3. Synthesis of speech using HNM, modified from [3].

For calculating the noise parameters, the synthesized voice part of speech is obtained using Eq. (2) and noise part is obtained by subtracting this from the original speech. Noise part is analyzed for obtaining LPC coefficients and energy envelope. 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.

Synthesis scheme is shown in Fig. 3. As the analysis parameters were obtained on frame-by-frame basis, they are interpolated before synthesis for obtaining the parameters values at each sample. Linear interpolation was used for this purpose. Phase unwrapping is carried out before interpolation of the phase values. 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 Fig. 2.

Voicing detector is implemented using the algorithm proposed by Childers [12] and GCIs are estimated using Childers and Hu' s algorithm [9],[12]. We have used electroglottogram (EGG) waveform from an impedance glottograph [10],[11] as a source for precise estimation of GCIs. The input of the GCI detector in Fig. 2 is either speech signal or EGG waveform.

3. METHODOLOGY

Speech segments corresponding to vowels and vowel-consonant-vowel in Hindi were recorded to be used as input for HNM based analysis/synthesis. The recording was done using B & K 4176 microphone connected to precision sound level meter B & K 2235 and its AC output was amplified and given to one channel of the line-in input of PC sound card. The second channel of the line-in was used for recording EGG waveform from an impedance glottograph. The recording was done in an acoustically treated room at a sampling rate of 10 k samples/s and 16 bit quantization. Speech from both male and female speakers was recorded.

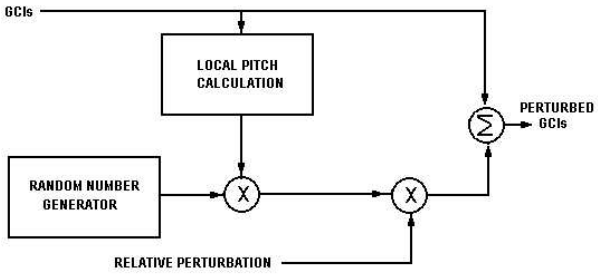
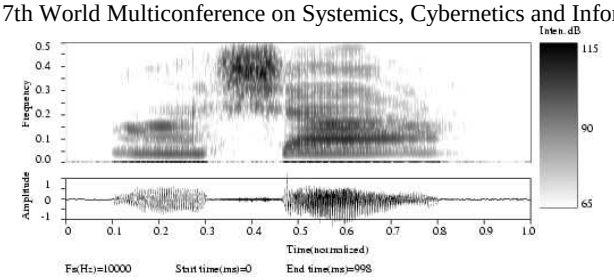


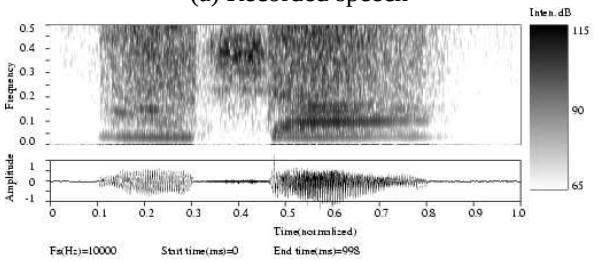
Fig. 4. Calculation of perturbed GCIs.

To study the effect of broadband noise in the input speech on the quality of the synthesized speech, broadband noise was added to each syllable at different SNRs. SNR value was maintained on the basis of signal and noise power over a pitch period. These were used for synthesis using HNM to study the limit of susceptibility of HNM for different syllables.

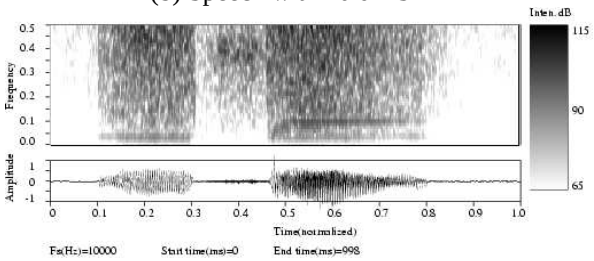
The scheme shown in Fig. 4 was used for introducing perturbations in GCIs. A random number, from a sequence with normal distribution in the range [-1, +1], multiplied by the local pitch and amount of desired relative perturbation is added to the GCI location.



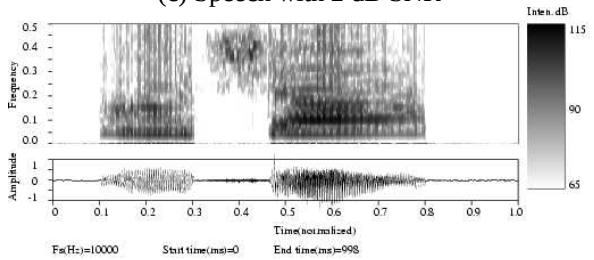
(a) Recorded speech



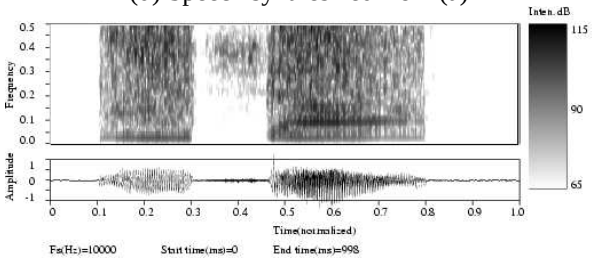
(b) Speech with 10 dB SNR



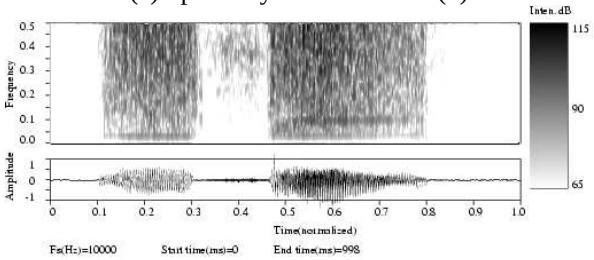
(c) Speech with 2 dB SNR



(d) Speech synthesized from (a)

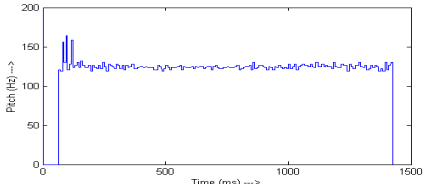


(e) Speech synthesized from (b)

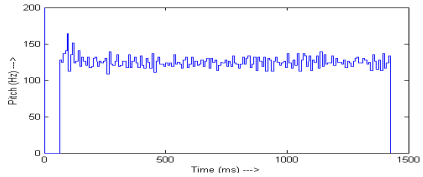


(f) Speech synthesized from (c)

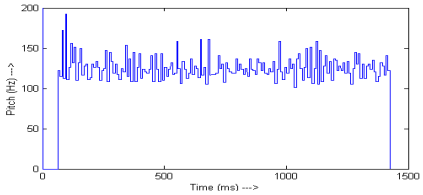
Fig. 5. Waveforms and spectrograms for /*a*/ for different amount of broadband noise and the corresponding synthesized outputs.



(a) 0 % perturbation

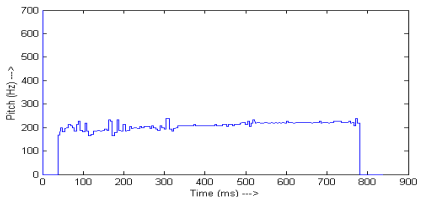


(b) 4 % perturbation

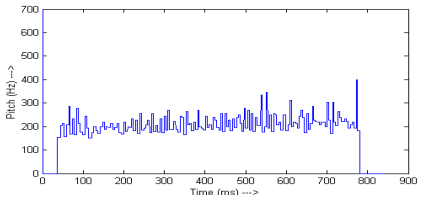


(c) 8 % perturbation

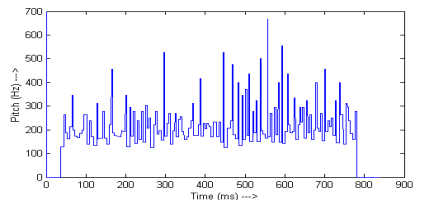
Fig. 6. Pitch contours for /*a*/ spoken by a male speaker, with perturbations to the GCIs obtained from the EGG waveform.



(a) 0 % perturbation

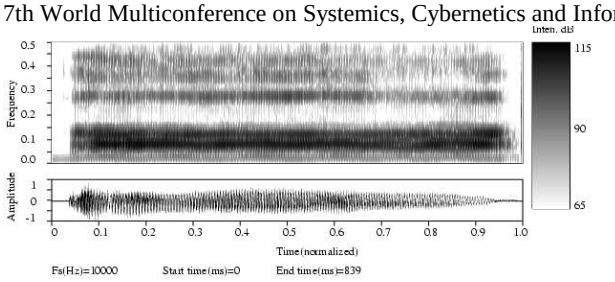


(b) 4 % perturbation

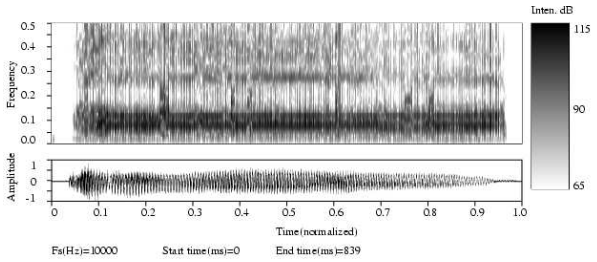


(c) 8 % perturbation

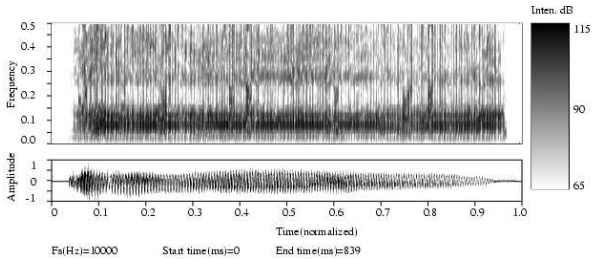
Fig. 7. Pitch contours for /*a*/ spoken by a female speaker with perturbations to GCIs obtained from the speech signal.



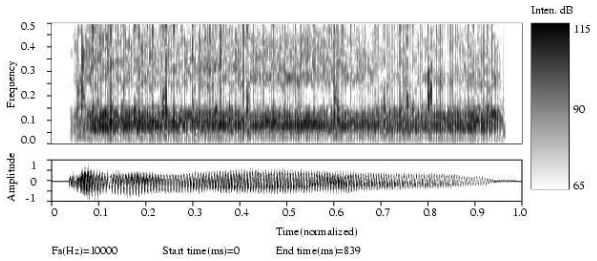
(a) Recorded



(b) Synthesized with 0 % perturbation

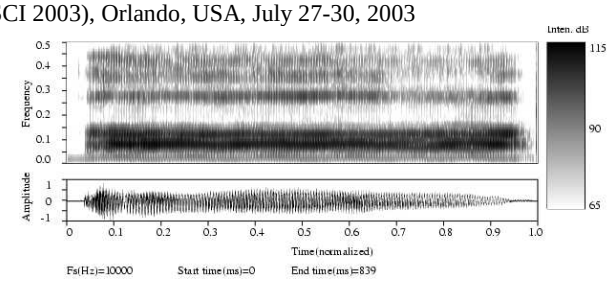


(c) Synthesized with 4 % perturbation

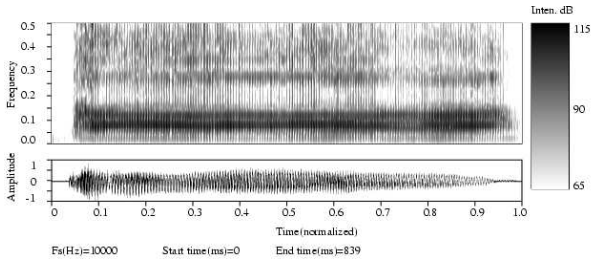


(d) Synthesized with 8 % perturbation

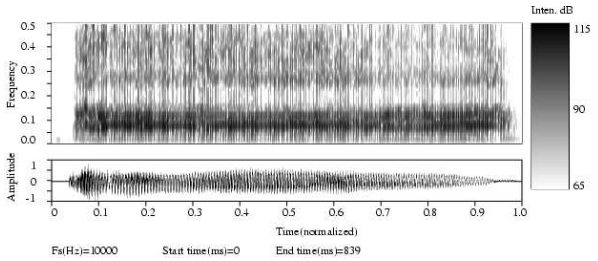
Fig. 8. Vowel /*a*/ spoken by a male speaker, GCIs from speech signal.



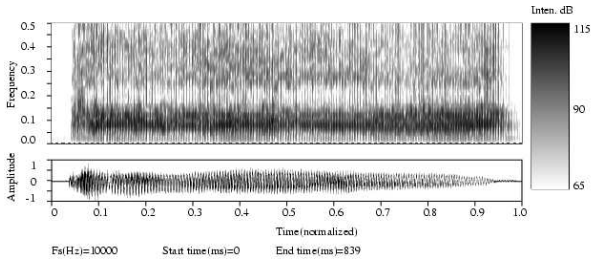
(a) Recorded



(b) Synthesized with 0 % perturbation



(c) Synthesized with 4 % perturbation



(d) Synthesized with 8 % perturbation

Fig. 9. Vowel /*a*/ spoken by a male speaker, GCIs from EGG waveform.

Thus each GCI is disturbed by an amount determined by a random number, local pitch, and relative perturbation. Harmonic synthesis block used in the calculation of noise parameters in Fig. 2 operated on the perturbed GCIs as input, in order to properly simulate the effect of random errors in the estimation of GCIs. The parameters of the noise part thus obtained along with those of the harmonic part are used for synthesis using the schematic of Fig. 3.

4. RESULTS

For SNRs higher than 10 dB, the noise added speech and the corresponding synthesized speech were fully intelligible and they were almost indistinguishable in quality. For SNR in the 2-10 dB range, all the syllables in both the types of speech remained fully intelligible. The quality of the synthesized speech was superior to that of the corresponding noise added speech. Intelligibility of both types of speech degraded for SNRs lower than 2 dB and the synthesized speech did not exhibit any improvement in intelligibility or quality as compared to the noise added speech.

It was observed that all the synthesized syllables, were perceived correctly, even if synthesized from only the harmonic part, except for /*a*ʃ*a*/ and /*asa*/. These two unvoiced fricative syllables when synthesized with harmonic part only were perceived as /*ata*/. If the noise

part was added to the harmonic part, they sounded natural. Fig. 5 shows the speech waveforms and spectrograms for speech with added broadband noise and corresponding speech synthesized using both the harmonic and noise parts. Spectrograms also show the reduction of noise in the synthesized syllables.

It may be inferred that in 2-10 dB range, the additive noise does not affect the calculation of parameters of the harmonic part, and it gets almost rejected during synthesis, and hence the synthesized output is enhanced in quality. Below SNR of 2 dB, the parameters of HNM are not correctly calculated and hence the synthesized output is not enhanced.

For investigating the effect of GCI perturbation, GCIs were obtained from the EGG waveform and speech signal. Using the scheme shown in Fig. 4, perturbation in 0 % to 20 % of the local pitch was introduced. Fig. 6 and 7 show examples of pitch contours for different perturbations. Spectrograms of the vowel /*a*/ spoken by a male speaker, synthesized with different amount of perturbations, are shown in Fig. 8 and 9 for the two methods of GCI estimation.

When the GCIs were estimated from the speech signal and analysis/synthesis was performed with GCI perturbation varying from 0 to 4 %, the output speech quality was not affected much as confirmed with listening tests. Although spectrograms show some noise at 4 % perturbation, the output speech is perceptually satisfactory. When the perturbation was further increased from 5 to 8 %, the speech quality started deteriorating and after 8 % it was severely affected as seen in Fig. 8d. These limits for perturbation were approximately the same for male and female speakers.

When the GCIs were obtained from EGG waveform, the speech quality was slightly better. For example, the speech remained acceptable for perturbation up to 6 %. We see that the spectrograms for synthesized speech with GCIs from EGG in Fig. 9 show less disturbance as compared to those with GCIs obtained from the speech signal itself in Fig. 8.

5. CONCLUSIONS

Investigations showed that for SNR in the 2-10 dB range, the quality of synthesized speech was superior to that of the speech mixed with noise, indicating that additive noise did not affect the estimation of parameters of the harmonic part.

Experiments with the pitch synchronous synthesis of vowels with different amount of perturbations in GCIs showed that perturbations beyond 4 % result in quality degradation. Hence precise determination of GCIs is necessary for high quality synthesis of speech, and a simultaneous recording of EGG waveform from an impedance glottograph can be used for this purpose.

REFERENCES

[1] 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, 1986, pp. 744-754.

[2] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis”, **IEEE Trans. Speech Audio Processing**, Vol. 9, No. 1, 2001, pp. 21-29.

[3] P. K. Lehana and P. C. Pandey, “Speech synthesis in Indian languages”, **Proc. Int. Conf. on Universal Knowledge and Languages-2002**, Goa, India, Nov. 25-29, 2002, paper no. pk1510.

[4] S. Miki, K. Mano, T. Moriya, K. Ogychi, and H. Ohmuro, "A pitch synchronous innovation CELP (PSI-CELP) coder for 2-4 kbits/s", **Proc. ICASSP**, 1994, pp. II.113 - II.116.

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

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

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

[8] P. Ladefoged, **A Course in Phonetics**, New York: Harcourt Brace Jovanovich, 1975.

[9] D. G. Childers and T. H. Hu, “Speech synthesis by glottal excited linear prediction”, **J. Acoust. Soc. Am.**,Vol. 96, No. 4, 1994, pp. 2026-2036.

[10] M. Rothenberg, “A multichannel electroglottograph”, **Journal of Voice**, Vol. 6, No. 1, 1992, pp. 36-43.

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

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