Effect of GCI Perturbation on Speech Quality in Indian Languages

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

Abstract— Harmonic plus noise model (HNM) divides the speech spectrum into two bands: harmonic and noise. One is modeled with harmonics of the fundamental and the other is simulated using random noise. As HNM analysis/synthesis is performed at each glottal closure instant (GCI), errors in estimation of GCIs affect the quality of the synthesized speech. The objective of this paper is to investigate the effect of the perturbation in GCIs on the synthesized speech quality. Childers and Hu's algorithm is used for GCIs calculations. Investigations show that the speech quality is very sensitive to positions of the GCIs. A small perturbation with maximum of 4 % of the local pitch period considerably degrades the synthesized speech. Perturbations above 8 % severely affect quality of the output speech.

1. INTRODUCTION

Different approaches for speech synthesis can be broadly classified in two categories: waveform and model based. Waveform based approach uses prerecorded passages of speech and plays a subset of these passages in a particular sequence for generating the desired speech output. This method provides high quality and naturalness, but has a 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 most of these approaches the extraction of parameters is very difficult and time consuming. A very flexible acoustic model for speech synthesis 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 [1]. Same limitations of computation load and a big number of parameters are also applicable to this approach but harmonic plus noise model (HNM), which is a variant of sinusoidal modeling of speech 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 into 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. Also, analysis/synthesis is pitch synchronous in HNM [3],[2]. In pitch synchronous speech synthesis, the analysis/synthesis of the speech is performed at each glottal closure instant (GCI).

Although pitch synchronous analysis/synthesis is well suited for high speech quality speech output, but problem of this method is that we need to estimate the glottal closure instants (GCIs) very accurately. GCIs 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 [4],[5].

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 [6]. 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 inaccurate values of GCIs, the quality of the synthesized speech is affected. So it is necessary to employ a more accurate method for these calculations.

The objective of this paper is to investigate the effect of the perturbation in GCIs on the synthesized speech quality that is perceptually acceptable for both vowels and syllables. 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 [7]. GCIs are obtained by processing the speech signal using Childers and Hu's algorithm and from the glottal signal obtaining using an impedance electroglottograph [8],[3].

0-7803-7651-X/03/$17.00 © 2003 IEEE

2. HARMONIC PLUS NOISE MODEL

In HNM based synthesis, the speech signal is assumed 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 [2].

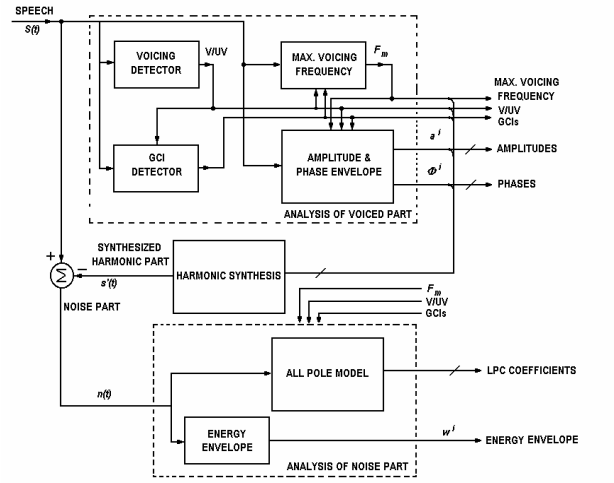


Figure 1. Analysis of speech using HNM, modified from [3].

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

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

$a_l(t)$ 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. 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 $w(t)$:

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

Along with maximum voiced frequency F_m, 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 are calculated for each frame.

Figure 1 shows a block diagram for analysis using HNM. The 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 calculate glottal closure instants (GCIs) accurately. GCIs can be calculated either using the speech signal or electroglottogram (EGG). So the input of GCI detector (Figure 1) is either speech signal or EGG. 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.

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.

Voicing detector is implemented using the algorithm proposed by Childers [9] and GCIs are calculated using Childers and Hu's algorithm [9],[8].

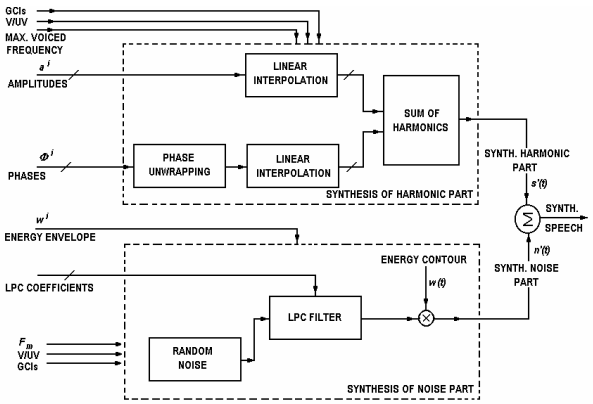


Figure 2. Synthesis of speech using HNM, modified from [3].

As the GCIs obtained by using EGG are more accurate [7], for comparison we have also obtained GCIs from the glottal waveform recorded by using an impedance glottograph [10],[11]. 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 [10].

3. METHODOLOGY

After implementing the HNM, speech segments corresponding to vowels and vowel-consonant-vowel in Indian languages were recorded, analyzed and synthesized. The recording was done using a B & K 4176 microphone. 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.

To study the effects of perturbation in GCIs, glottal closure instants (GCIs) are first calculated for both glottal and speech waveforms using the Childers and Hu's algorithm.

The scheme for adding perturbation to the GCIs is shown in Figure 3. Then random number generator generates a sequence with normal distribution in the range [-1, +1]. This is multiplied by the local pitch and amount of desired relative perturbation in the GCIs. Resulted value is added to the GCI under consideration. This process is repeated for all GCIs. So each GCI is disturbed by an amount determined by the random number, local pitch, and relative perturbation.

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

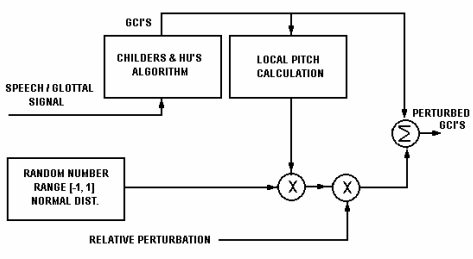


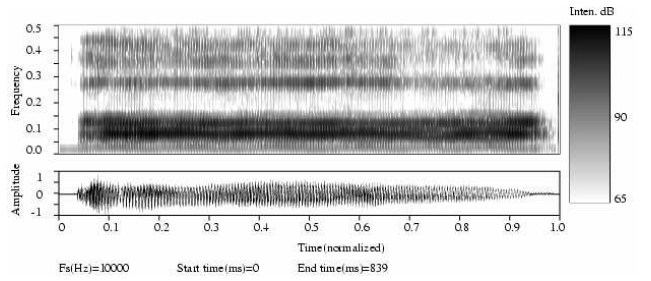
Figure 3. Calculation of perturbed GCIs.

4. RESULTS

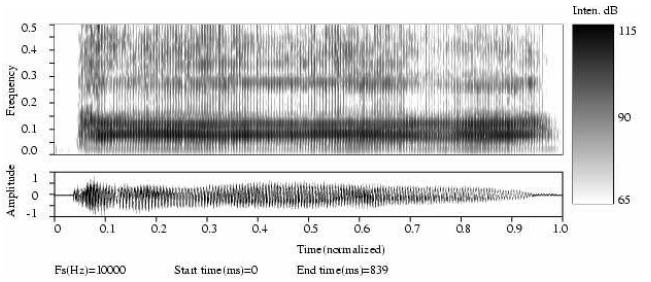
GCIs are obtained from the glottal and speech signals. Using the scheme shown in the Figure 3, perturbations in the GCIs is added. The amount of perturbation is from 0 % to 20 % of the local pitch. Spectrograms of the vowel /a/ spoken by a male and female speaker, synthesized with different amount of perturbations, are shown in Figures 4 and 5 respectively.

If the GCIs are calculated from the speech signal and analysis/synthesis is performed with perturbation varying from 0 % to 4 % added to these GCIs, the output speech quality is not affected much as confirmed and listening tests.

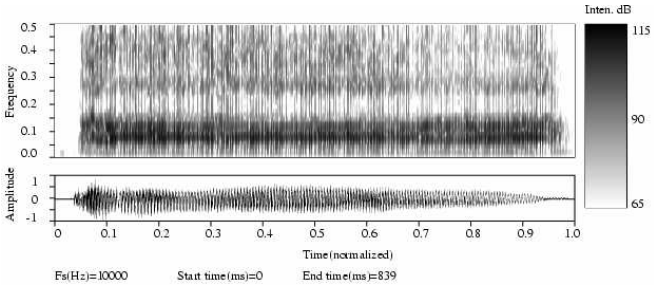
Although spectrograms show some noise at even 4 % perturbation but the output speech is perceptually satisfactory. When the perturbation is further increased from 5 % to 8 %, the speech quality starts deteriorating and after 8 % it is severely affected (Figure 4d and 5d). These limits for perturbation are approximately the same for male and female speakers.



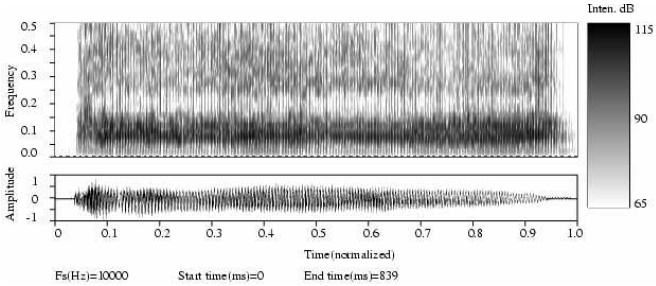
(a) Recorded



(b) 0 % perturbation



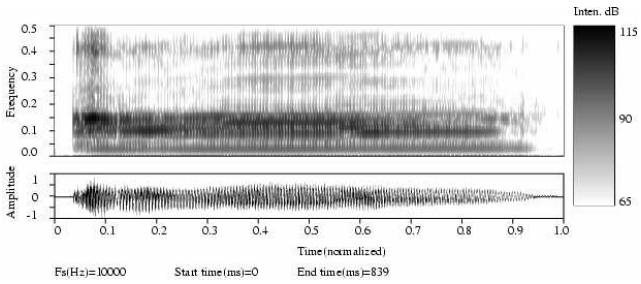
(c) 4 % perturbation



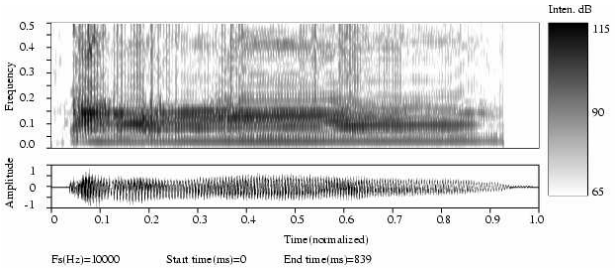
(d) 8 % perturbation

Figure 4. Spectrograms for /a/ spoken by a male speaker for different amount of perturbations added to GCIs obtained from glottal signal.

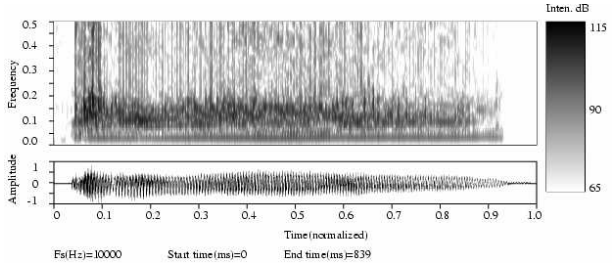
If the GCIs are obtained from glottal signal instead of obtaining from speech, the performance is slightly better. For example, the speech remains acceptable even for the values of perturbation as high as 6 %. That is the reason that the spectrograms (Figure 4) for synthesized speech by taking GCIs from EGG are clearer as compared to those (Figure 5) for which GCIs are obtained from the speech signal itself. Figure 6 & 7 show the pitch contours obtained before and after adding perturbation in the GCIs.



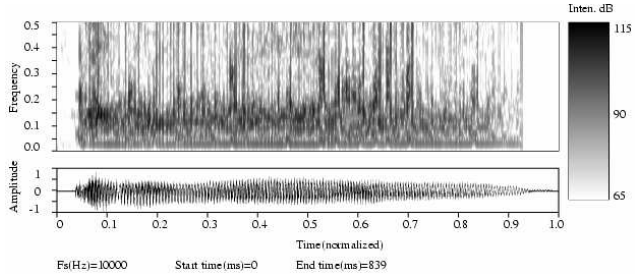
(a) Recorded



(b) Synthesized with 0 % perturbation

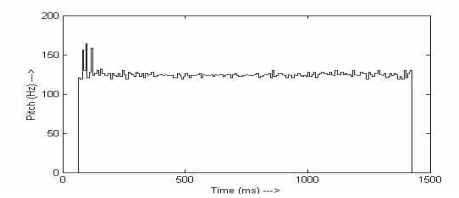


(c) Synthesized with 4 % perturbation

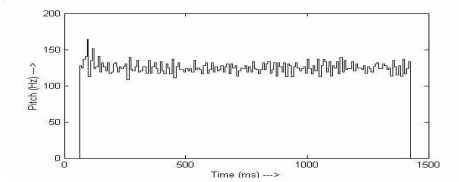


(d) Synthesized with 8 % perturbation

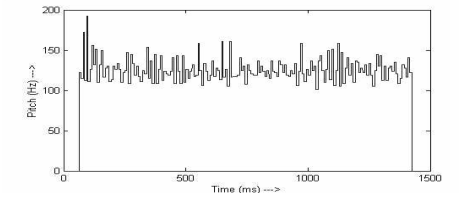
Figure 5. Spectrograms for /a/ spoken by a female speaker for different amount of perturbations added to the GCIs obtained from speech signal.



(a) 0 % perturbation

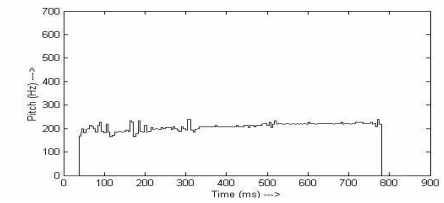


(b) 4 % perturbation

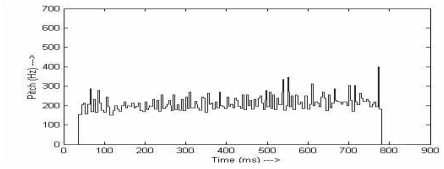


(c) 8 % perturbation

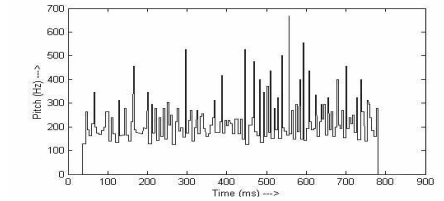
Figure 6. Pitch contours for /*a*/ spoken by a male speaker, with perturbations to the GCIs obtained from the EGG waveform.



(a) 0 % perturbation



(b) 4 % perturbation



(c) 8 % perturbation

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

5. CONCLUSIONS

Experimentations with the pitch synchronous synthesis of with different amount of perturbations in GCIs show that perturbations beyond 5 % result in quality degradation. Hence accurate determination of GCIs 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] 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, Jan. 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: pk1510.

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

[5] 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.

[6] 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.

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