Harmonic Plus Noise Model Based Speech Synthesis in Hindi and Pitch Modification

Parveen K. Lehana and Prem C. Pandey

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

Abstract

In harmonic plus noise model (HNM), each segment of speech is modeled as two bands: a lower "harmonic" part represented as amplitudes and phases of the harmonics of a fundamental and an upper noise part using an all-pole filter excited by random white noise, with dynamically varying band boundary. HNM based synthesis can be used for good quality output with relatively small number of parameters and it permits pitch and time scaling without explicit estimation of vocal tract parameters. We have investigated its use for synthesis in Hindi which has aspirated stops and lacks voiced fricatives. It was found that good quality synthesis could be carried out, including those of aspirated stops. The upper band of HNM was needed only for the palatal and alveolar fricatives. Sensitivity of output quality to the errors in glottal closure instants was studied and random perturbations exceeding 4% of the local pitch period resulted in noticeable degradation. Synthesis with pitch scaling showed that the frequency scale of the amplitudes and phases of the harmonics of the original signal needed to be modified by a speaker dependent warping function, obtained by studying the relationship between pitch frequency and formant frequencies for the three cardinal vowels spoken with different pitches.

1. Introduction

Sinusoidal modeling of speech provides a very flexible approach for speech synthesis. In this acoustic model, 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 speech spectrum in two sub bands: one modeled with harmonics of the fundamental and the other modeled 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, period-to-period variation of the glottal excitation, etc. Harmonic plus noise model permits pitch scaling without explicit estimation of vocal tract parameters.

In HNM synthesis, the analysis/synthesis is performed at each glottal closure instant (GCI) [2],[3], 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 period differs from the one or several pitch periods 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 and this further 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 HNM model for speech synthesis in Hindi and experiments were carried out to study the effect of the perturbations in GCIs on the synthesized speech quality that is perceptually acceptable. The phoneme set of Hindi has many interesting features. It has 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]. Synthesis with pitch modifications is also studied.

2. Harmonic plus noise model

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

$$s(t) = s'(t) + n'(t) \quad (1)$$

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

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

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 Gaussian noise $b(t)$ filtered by a time-varying all-pole normalized filter $h(\tau;t)$ and amplitude modulated by an envelope function $w(t)$

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

The analysis/synthesis scheme is shown in Fig. 1 and Fig. 2. 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 using Childers algorithm [12], 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.

As shown in Fig. 1, 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. 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 order of the LPC filter is fixed at 10. 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 is schematically shown in Fig. 2. 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.

We have used electroglottogram (EGG) waveform from an impedance glottograph [10],[11] as a source for precise estimation of GCIs.

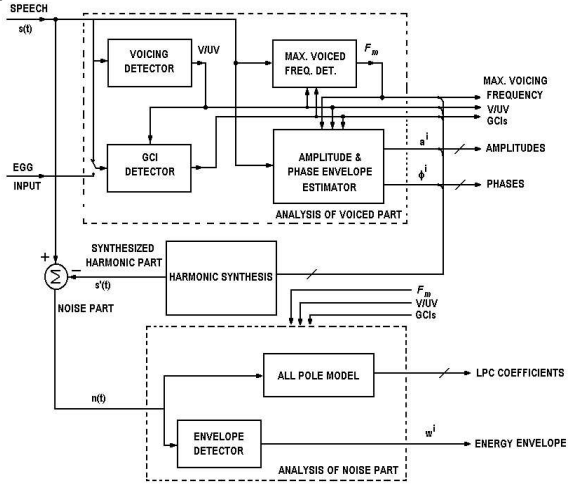


Figure 1: Analysis of speech using HNM [3].

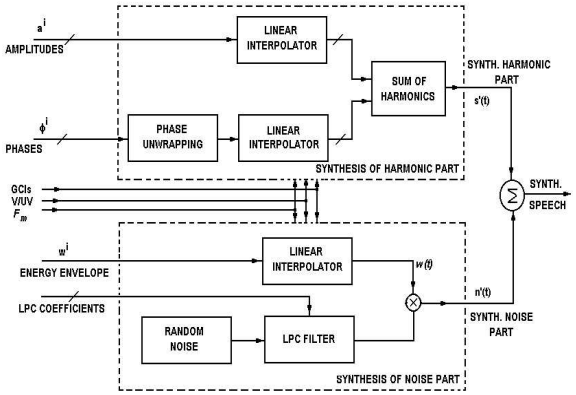


Figure 2: Synthesis of speech using HNM [3].

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.

The scheme shown in Fig. 3 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. 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. 1 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. 2.

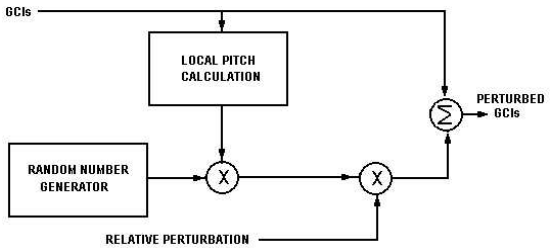


Figure 3: Calculation of perturbed GCIs.

Pitch scaling in HNM can be carried out by synthesizing the speech with interpolated original amplitudes and phases at the multiples of the scaled pitch frequency, experiments show that this results in an unnatural quality and for obtaining natural quality output frequency scale of the amplitudes and phases of the harmonics of the original signal are needed to be modified by a speaker dependent warping function. Hence, it is necessary to study the relation between the pitch frequency and the vocal tract parameters. The recordings of the three cardinal vowels (/*a*/, /*i*/, and /*u*/) from male and female speakers were analyzed. The speakers had a good training of classical music and they were able to utter the sustained vowels at different notes. The scheme for pitch scaling is shown in Fig. 4. First the parameters of the speech of the source speaker are calculated and then these parameters are modified for achieving the target pitch contour using a warping function. The warping function was obtained by studying the relationship between pitch frequency and formant frequencies for the three cardinal vowels each one spoken at several notes. Then the resynthesis is performed from the modified parameters.

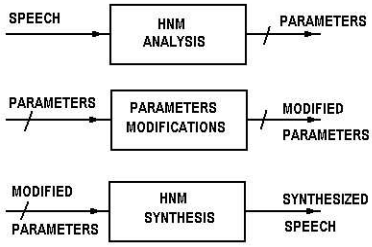
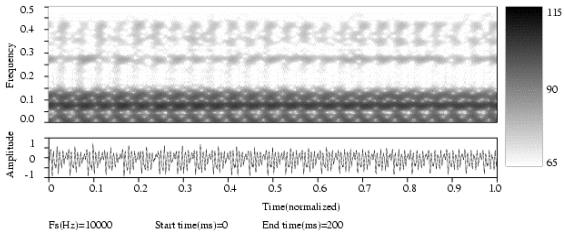


Figure 4: Analysis/synthesis scheme.

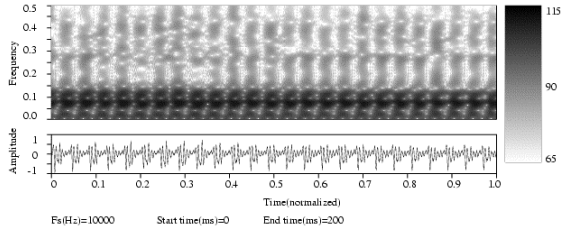
4. Results

All the synthesized syllables, including those with aspirated stops and affricates, were perceived correctly, even if synthesized from only the harmonic part, except for /*a*͡*ʃa*/ and /*asa*/. These two syllables, involving unvoiced palatal and alveolar fricatives, when synthesized with harmonic part only were perceived as /*ata*/. If the noise part was added to the harmonic part, they sounded natural.

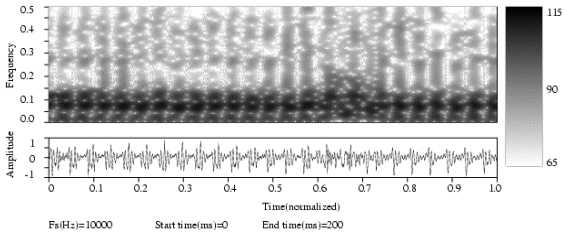
For investigating the effect of GCI perturbation, GCIs were obtained from the EGG waveform and speech signal. Using the scheme shown in Fig. 3, perturbation in 0 % to 20 % of the local pitch was introduced. Spectrograms of the vowel /*a*/ synthesized with different amount of perturbations, are shown in Fig. 5.



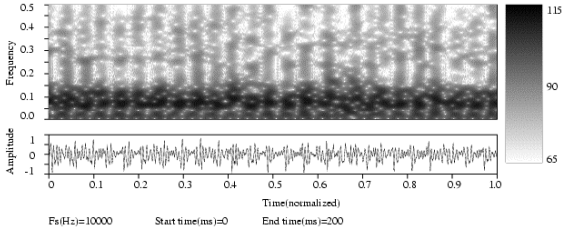
(a) Recorded



(b) Synthesized with 0 % perturbation



(c) Synthesized with 4 % perturbation



(d) Synthesized with 8 % perturbation

Figure 5: Vowel /*a*/ spoken by a male speaker, GCIs from speech signal.

Increase in the perturbation results in more noise in the synthesized speech. 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 spectrogram shows 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. 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. The speech remained acceptable for perturbation up to 6 %.

For pitch modification, analysis of the three cardinal vowels (/*a*/, /*i*/, /*u*/) recorded with different pitches showed that the first formant monotonically increased with the pitch for the three vowels uttered by three speakers (one female and two males). No consistent pattern in the changes for other formants could be seen. The relationship between pitch and formants was speaker dependent. Variation of formant frequencies with pitch for a male speaker is shown in Fig. 6. Synthesis of vowels by using Klatt' s formant synthesizer [13] after interchanging formants with those from a different pitch value resulted in unnatural output, indicating that an appropriate relationship between pitch and formants is necessary for natural sounding output. Next, a linear regression was fitted for each of the four formant frequencies and the pitch values for the recording of the three cardinal vowels spoken at different notes. Synthesis using the formant frequencies obtained from these relations resulted in good quality output.

For HNM synthesis with pitch modification, good quality vowels were produced if the frequency scale of the amplitudes and phases of the harmonics was scaled by a factor equal to that of pitch scaling. However, synthesized speech does not correspond to the speaker. Pitch scaling factor less than 1.5 resulted in good quality output. Scaling factor greater than 2 resulted in unacceptable quality. We are studying the use of a speaker dependent warping function for frequency of amplitudes and phases of harmonics in the harmonic part and LPC coefficients from the frequency warped spectrum in the noise part.

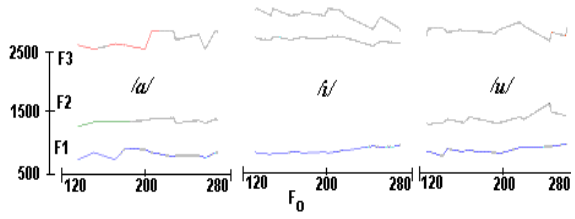


Figure 6: Variation of formant frequencies F1, F2, F3 with pitch F_o (all in Hz) for a male speaker.

5. Conclusions

HNM is capable of synthesizing the phoneme set in Hindi with good quality. 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. High quality pitch modification of speech required a speaker dependant warping function.

6. References

[1] McAulay, R. J., and Quatieri, T. F., "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. Acoust. Speech Signal Processing*, Vol. 34(4), 1986, pp. 744-754.

[2] Stylianou, Y., "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Trans. Speech Audio Processing*, Vol. 9(1), 2001, pp. 21-29.

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

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

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

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

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

[9] Childers, D. G., and Hu, T. H., "Speech synthesis by glottal excited linear prediction", *J. Acoust. Soc. Am.*, Vol. 96(4), 1994, pp. 2026-2036.

[10] Rothenberg, M., "A multichannel electro-glottograph", *J. Voice*, Vol. 6(1), 1992, pp. 36-43.

[11] Lehana, P. K., and Pandey, P. C., "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] Childers, D. G., *Speech Processing and Synthesis Toolboxes*, John Wiley, New York, 2000.

[13] Klatt, D. H., "Software for a cascade/paraller formant synthesizer", *J. Acoust. Soc. Am.*, Vol. 67(3), 1980, pp. 671-995.