International Conference on Universal Knowledge and Language - 2002, Goa, India, November 25-29, 2002, paper no. 14

SPEECH SYNTHESIS IN INDIAN LANGUAGES

P. K. Lehana and P. C. Pandey

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

Abstract: This paper presents the study of phonemes in the Indian languages for developing good quality speech synthesis. Harmonic plus noise model (HNM) which divides the speech signal in two sub bands: harmonic and noise, is implemented with the objective of studying its capabilities and to investigate the adaptation needed. Childers and Hu's algorithms are used for voicing and pitch detection. As the analysis and synthesis of the speech is pitch synchronous, so glottal closure instants (GCIs) should be accurately calculated. For comparison, impedance glottograph is also used to extract the GCIs. Investigations show that HNM is capable of synthesizing all syllables and passages spoken in different styles in Indian languages with natural sounding output. The quality of synthesized speech improves if the GCIs are taken from the glottal signal instead of being obtained by processing the speech signal. As HNM uses amplitudes and phases of the pitch harmonics, interpolating the spectrum at the desired pitch harmonics and synthesizing using these values can achieve pitch scaling.

1. INTRODUCTION

There are two approaches to generate synthetic speech: 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 [1],[2]. Model based synthesis approach uses a model of the speech generation such as source-filter, articulatory, or acoustic, and a set of parameters. For example, formant synthesizer uses the source-filter model of speech production. There are many variants of formant synthesizers such as cascade, parallel, and cascade/parallel formant synthesizer. Although formant synthesizer can synthesize infinite number of sounds with only small number of parameters but generation of fricatives, nasalized sounds, and plosive bursts are difficult to synthesize with acceptable quality [2]. In formant synthesis, the set of rules controlling the formant frequencies, amplitudes, and the characteristics of the excitation source is large. The parameter extraction is also very difficult and time consuming. Articulatory synthesis models the human speech production system directly, with a set of area functions between glottis and mouth. It allows accurate modeling of transients due to abrupt area changes. Although it can be used for high quality speech output, the collection of data and implementation of rules to drive the data correctly is very complex [2]. Acoustic model of speech is used in sinusoidal modeling [3]. Each frame of the speech is represented as summation of sine waves with different amplitudes, frequencies, and phases. 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. It gives high quality speech with less number of parameters. Also, the pitch and time scaling is relatively easy [4]. Because HNM seems to be a promising model for speech synthesis, it is implemented and its suitability for phoneme sets in Indian languages, particularly for features, which are peculiar to Indian languages, is investigated.

2. HARMONIC PLUS NOISE MODEL

HNM divides the speech signal into two parts: harmonic, and noise. The harmonic part is responsible for the quasi-periodic components of the speech signal and the noise part accounts for the non-periodic components such as fricative or aspiration noise, period-to-period variation of the glottal excitation, etc. These two bands are separated by a frequency called maximum voiced frequency F_m [4].

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

$$s(t) = \mathrm{Re} \sum_{l=o}^{L(t)} a_l(t) \exp\{ j[\int_0^t l\omega_0(\sigma) d\sigma + \theta_l]\} \quad (1)$$

$a_l(t)$ and $\theta_l(t)$ are the amplitude and phase at time of the lth harmonic, ω_0 is fundamental frequency and $L(t)$ is the number of harmonics. The noise part is modeled by an AR model and is 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 $w(t)$:

$$n(t) = w(t)[h(\tau;t) * b(t)] \quad (2)$$

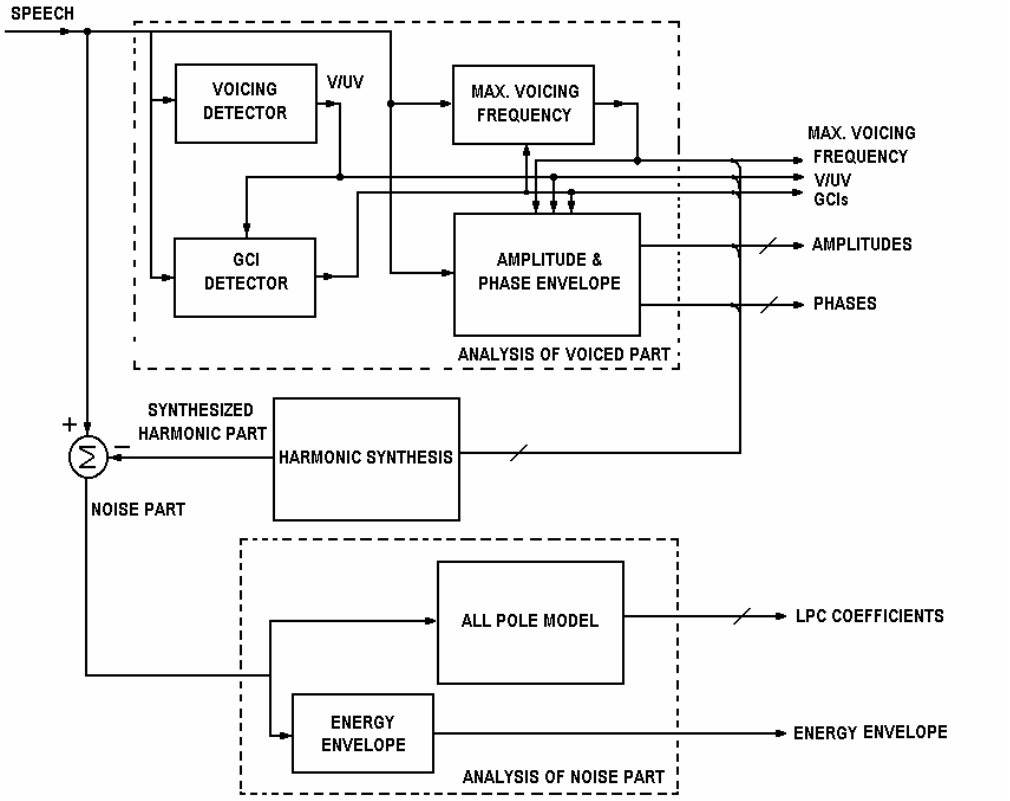


Fig. 1. Analysis of speech.

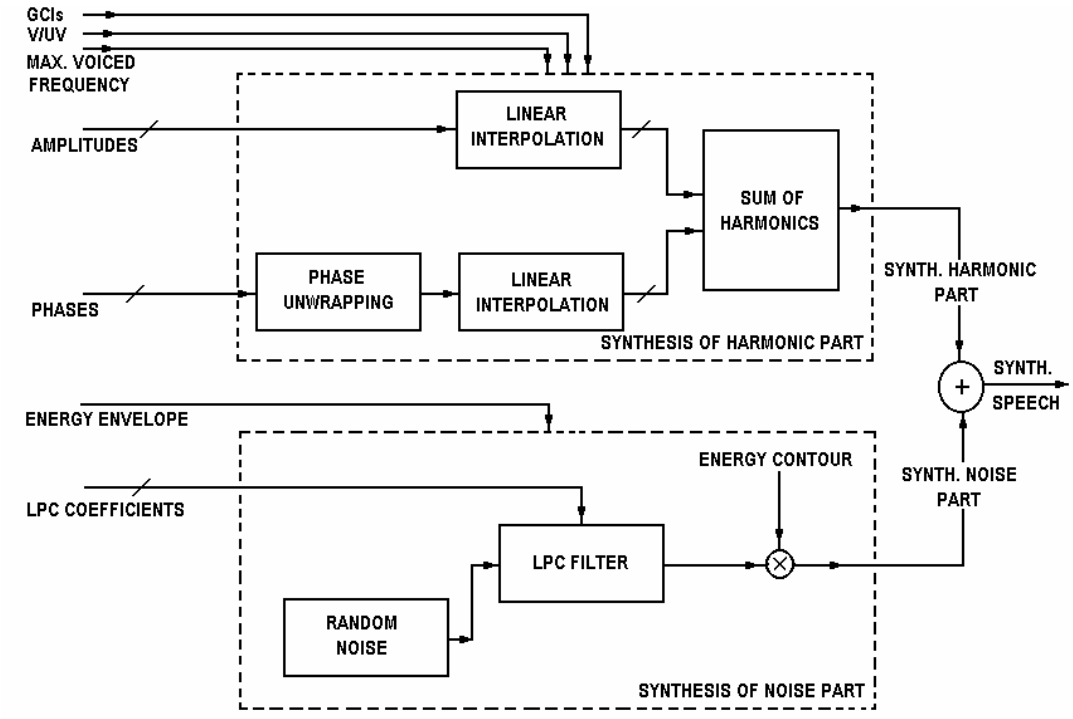


Fig. 2. Synthesis of speech.

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

The speech signal is applied to the voicing detector, which declares the frames either voiced or unvoiced using the method proposed by Childers [5]. As the analysis and synthesis in HNM are pitch synchronous, the glottal closure instants (GCIs) are calculated accurately. Childers and Hu's algorithm is used for these calculations [5],[6]. For comparison, we have also obtained GCIs from the glottal waveform from an impedance glottograph [7],[8]. 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 [7]. 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 envelope.

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.

Once the amplitudes and phases of the pitch harmonics are available, pitch scaling is easy to achieve. Scheme for pitch scaling is shown in Fig. 3. The parameters of the original speech (amplitudes and phases) are interpolated at the harmonics of the desired pitch. The number of harmonics to be interpolated depends on the maximum voiced frequency F_m. The last harmonic should be less than F_m. Then synthesis is performed with these interpolated parameters and pitch scaled speech is obtained.

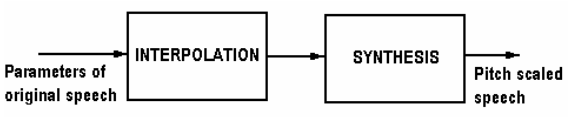
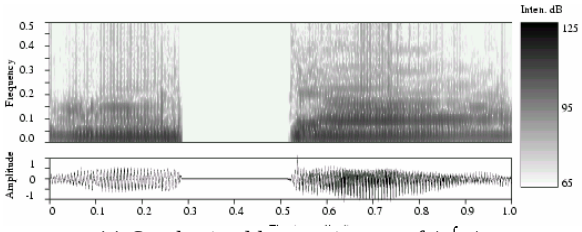


Fig. 3. Pitch scaling.

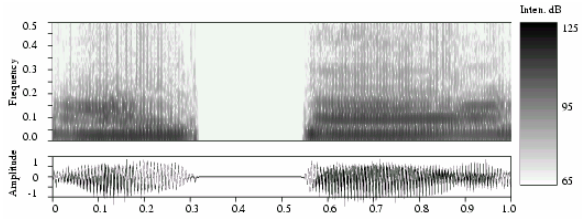
3. RESULTS

After implementing the HNM, sounds (vowels, syllables in Hindi with vowel taken as /*a*/, all-voiced passages, interrogative sentences spoken in three different styles: normal, high articulatory rate, showing anger) were recorded, analyzed and synthesized. First channel of line-in of the sound card was connected to impedance glottograph and second channel was used for recording of speech. The recording was done in an acoustic room at a sampling rate of 10 ksa/s with male and female speakers.

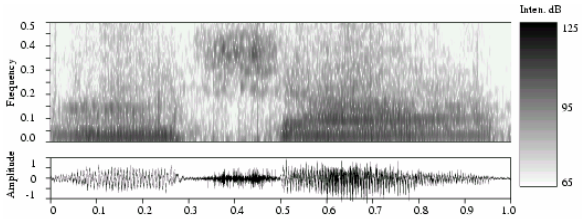
The quality of the synthesized vowels and all syllables with vowel taken as /*a*/ are intelligible with good quality if synthesized with only harmonic part except /*aʃa*/ and /*asa*/. Synthesized harmonic parts of these two sounds sound almost the same as confirmed by listening and spectrograms (Fig. 4). After adding synthesized noise part to these, sounds become well differentiable.



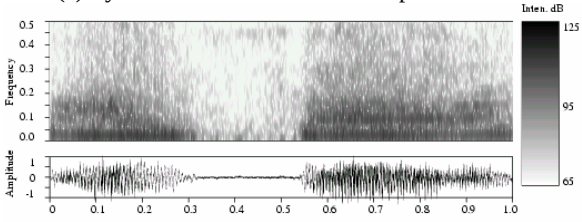
(a) Synthesized harmonic part of /*aʃa*/



(b) Synthesized harmonic part of /*asa*/



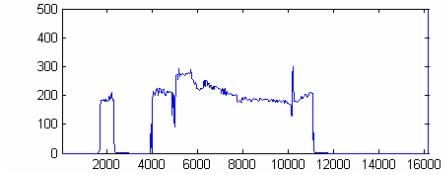
(c) Synthesized harmonic + noise part of /*aʃa*/



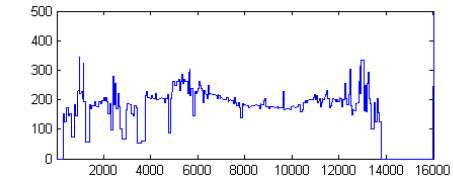
(d) Synthesized harmonic + noise part of /*asa*/

Fig. 4. Spectrograms for synthesized /*aʃa*/ and /*asa*/.

HNM could synthesize all the passages spoken in different styles and at different articulatory rate. For synthesizing these passages, two options were tried. In the first option the GCIs were extracted from the speech signal, and in the second the glottal signal was used to extract these GCIs. It is seen that the speech quality of the one synthesized with GCIs from glottal signal is better than using the GCIs from speech signal. Fig. 5 shows the pitch contours obtained from speech and glottal signal for the passage "/*ap kʌhan ja rʌhE hʌn*/". It is seen that the GCIs calculated from speech signal do deviate from those obtained from the glottal wave.



(a) Pitch contour obtained from glottal waveform



(b) Pitch contour obtained from speech waveform

Fig. 5. Pitch contours for the passage "/*ap kʌhan ja rʌhE hʌn*/". The units for x-axis are in samples and y-axis are Hz.

Experiments with pitch scaling gave satisfactory results. Pitch scaling factor α was varied in the range 0.5 to 1.5. Fig. 6 shows the original speech /*aka*/. Fig. 7 & 8 show the synthesized speech /*aka*/ without pitch scaling and with pitch scaling by α = 1.5 respectively. The pitch of the original speech was around 142 Hz. If the value of α is varied beyond this range, the quality of the speech is severely degraded.

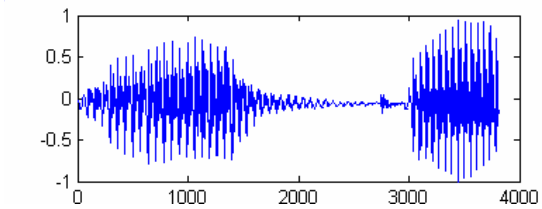


Fig. 6. Original speech /*aka*/.

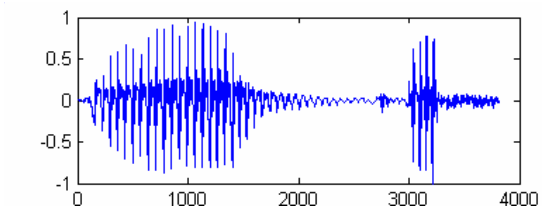


Fig. 7. Synthesized speech /*aka*/ without pitch scaling.

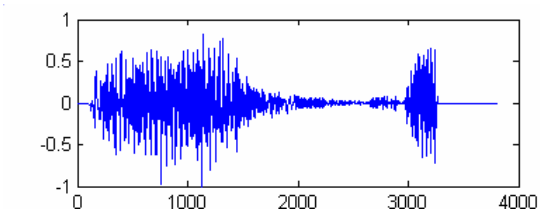


Fig. 8. Pitch scaled speech /*aka*/ by α = 1.5.

4. CONCLUSIONS

The quality of the synthesized sounds using HNM is satisfactory. It synthesizes all the vowels and syllables properly with good intelligibility and naturalness. All the syllables are intelligible if synthesized using only harmonic part except /*aʃa*/ and /*asa*/. HNM is able to synthesize the passages spoken with different styles and articulatory rates. The quality of the synthesized passages improves if we use GCIs obtained from the glottal signal instead of speech signal. Finally, the synthesis can be done with pitch scaling.

REFERENCES

[1] B. Mobius, J. Schroeter, J. Santen, R. Sprout, J. Olive, "Recent advances in multilingual text-to-speech synthesis," *Fortschritte der Akustik, DAGA-96*.

[2] D. O'Saughnessy, *Speech Communication - Human and Machine*. Hyderabad: Universities press, 2001.

[3] 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.

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

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

[6] 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.

[7] 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.

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