IMPROVING QUALITY OF 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

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 for improving the quality of speech synthesis in Indian languages. Investigations show that HNM is capable of synthesizing all vowels and syllables with good quality. All the syllables are intelligible if synthesized using only harmonic part except /*aʃa*/ and /*asa*/. This fact can reduce the size of the database. For pitch synchronous analysis and synthesis glottal closure instants (GCIs) should be accurately calculated. The quality of synthesized speech improves if these instants are obtained from the glottal signal (output of an impedance glottograph) instead of these being obtained by processing the speech signal. Further the noise part is synthesized pitch synchronously for voiced frames. A database of HNM parameters for VCV syllables is developed for Indian languages. The number of parameters for each frame is comparable to that of formant synthesizer but the quality of synthesized speech is much better.

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 articulatory, source-filter, or acoustic, and a set of parameters.

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].

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. Formant synthesizer can synthesize infinite number of sounds with a 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. Extraction of parameter tracts is also 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. There are many limitations of the sinusoidal modeling such as computation load and a big number of parameters [3]. 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 [4].

Selection of a model depends on many factors such as quality of synthesized speech, ease of parameter extraction, modification of parameters, number of parameters, and computation load. HNM gives high quality speech with less number of parameters. Also, the pitch and time scaling is relatively easy [4], [5]. 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 that are peculiar to Indian languages, is investigated.

2. IMPLEMENTATION OF 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 [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 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 $w(t)$:

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

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 using the method proposed by Childers [6]. 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 [6], [7]. For comparison, we have also obtained GCIs from the glottal waveform recorded by using an impedance glottograph [8], [9]. 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 [8]. 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. CONCATENATION OF SEGMENTS

The purpose of the concatenation of segments is to generate longer units of speech from smaller units. So we can generate unlimited words from a small database of phonetic units. Most of the synthesizers use phonemes, morphemes, diphones, triphones, or syllables as basic acoustic units. Combination of these units is also used in some synthesizers [2]. But we have employed VCV (vowel-consonant-vowel) syllables for creating the database of Indian

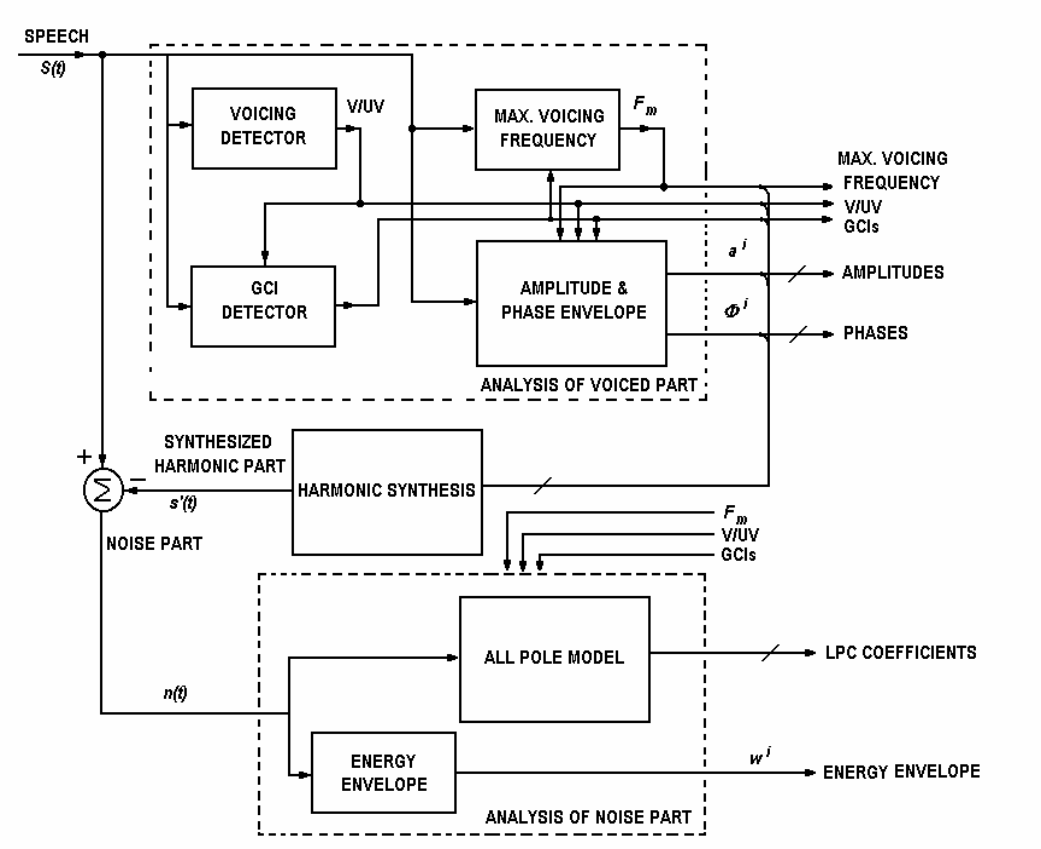


Figure 1. *Analysis of speech using* HNM, *modified from* [5].

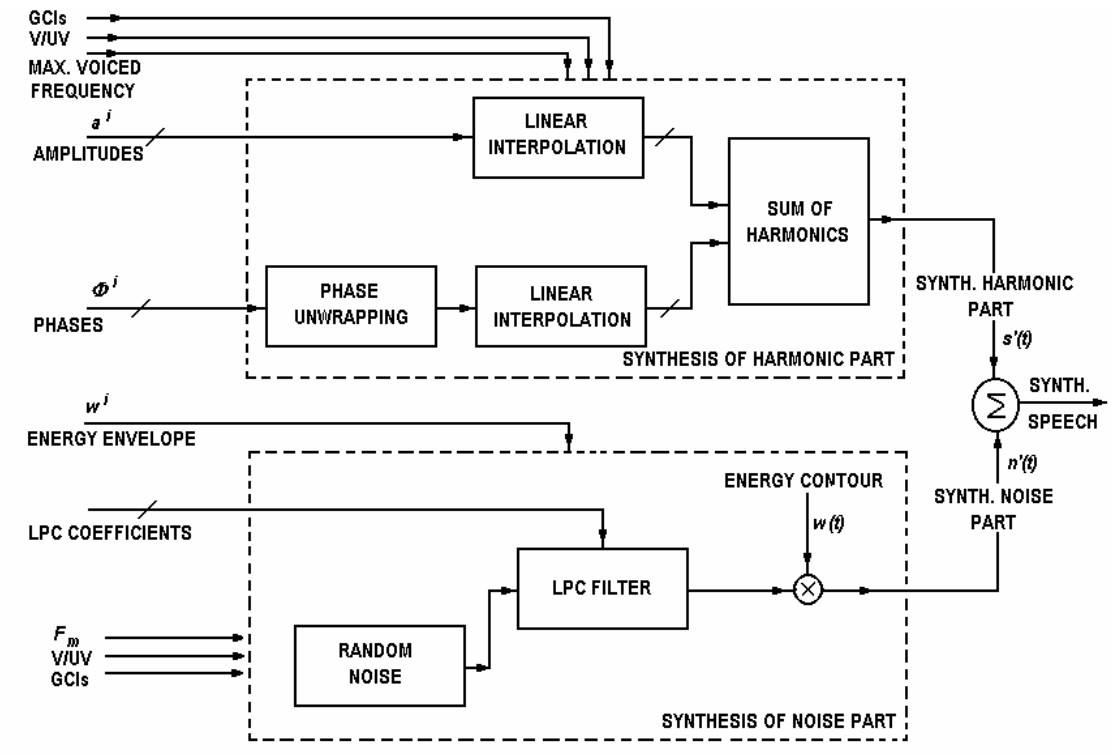


Figure 2. *Synthesis of speech using* HNM, *modified from* [5].

languages. As the full natural transitions across the consonants are preserved, the quality of the speech is not affected [10], [11].

After recording and labeling the VCV syllables, analysis is done to extract the HNM parameters. These parameters are stored in a database for future reference. The average number of parameters for each voiced frame is around 40 and for unvoiced frame is only 11. It should be noted that the number of parameters are comparable to that of a formant synthesizer. A phrase may be synthesized by parsing its phonetic transcript, fetching the parameters of the required units from the database, and concatenation of parameter tracts after pitch contour smoothening and energy normalization.

As the parameters are defined only at the analysis instants (one set for each frame), the parameters are interpolated for the remaining instants in-between the frame boundaries. Then using (1) & (2) the speech can be synthesized. Interpolation in the parametric domain minimizes the spectral distortion at the joints.

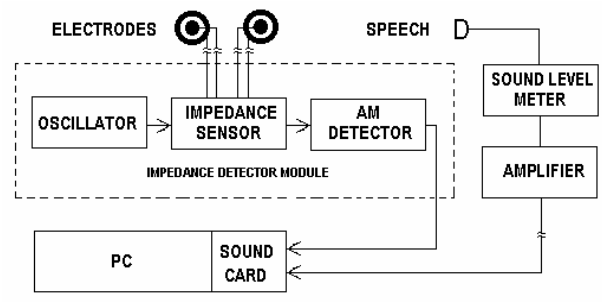
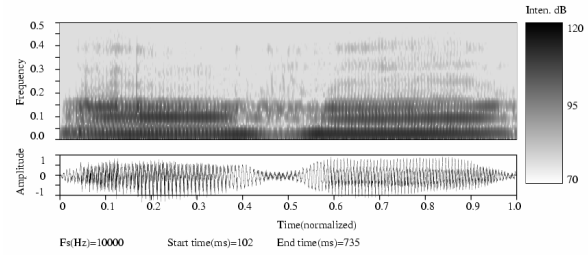


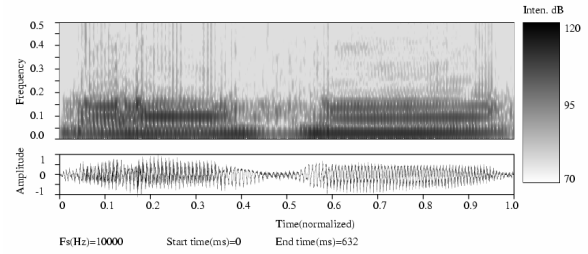
Figure 3. *Hardware setup for recording, modified from* [8].

4. RESULTS

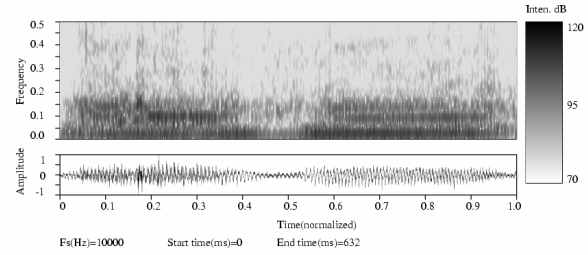
After implementing the HNM, speech segments including vowels, syllables in Hindi, all-voiced passages, interrogative sentences spoken in three different styles: normal, high articulatory rate, showing anger were recorded, analyzed and synthesized. The recording was done using a B & K 4176 microphone (set up is shown in Figure 3). 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 signals at the input of the sound card must be in the range ± 500 mV. 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.



(a) *Recorded /aha/*



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

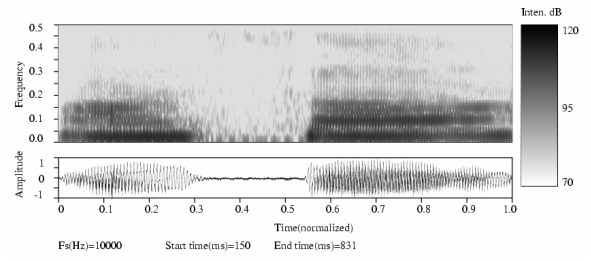


(c) *Synthesized harmonic + noise part of /aha/*

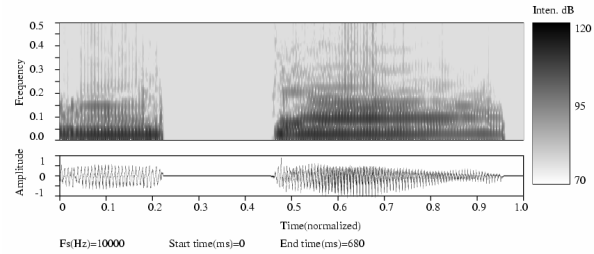
Figure 4. *Spectrograms for synthesized /aha/.*

The synthesis of the noise part was done using 10th order LPC filter. Both pitch synchronous and asynchronous methods of synthesizing the noise part were investigated. The results can be summarized as follows.

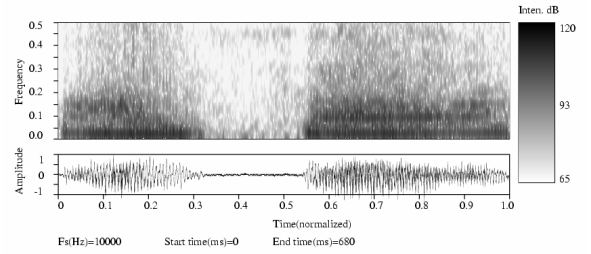
- For unvoiced stops, difference between aspirated and unaspirated is the voice onset time (VOT). VOT for aspirated stop is greater than that for unaspirated stop. For example in /*aka*/ and /*ak*^{h}*a*/, the VOTs are 49 ms and 60 ms respectively. When the VOT of /*aka*/ is increased by adding silence, /*ak*^{h}*a*/ is perceived.
- For voiced stops, the distinguishing feature between the aspirated and unaspirated is the different from the above. Although the duration of the release phase is greater for aspirated stop



(a) *Recorded /asa/*



(b) Synthesized harmonic part of /asa/

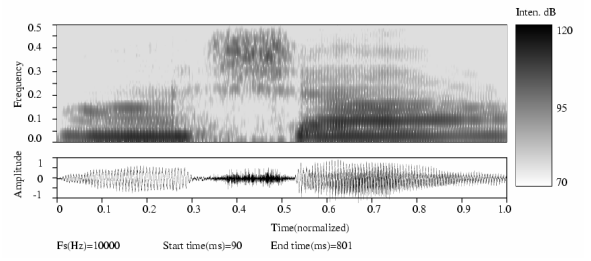


(c) *Synthesized harmonic + noise part of /asa/*

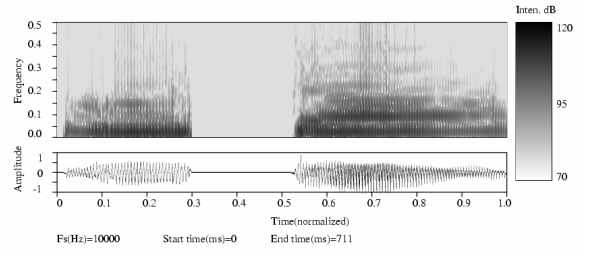
Figure 5. *Spectrograms for synthesized /asa/.*

than the corresponding unaspirated stop, but after adding noise, silence, or repeating the burst in the release phase does not produce the aspirated stop. On the contrary, when the release duration of the aspirated stop is decreased, the corresponding unaspirated stop is perceived.

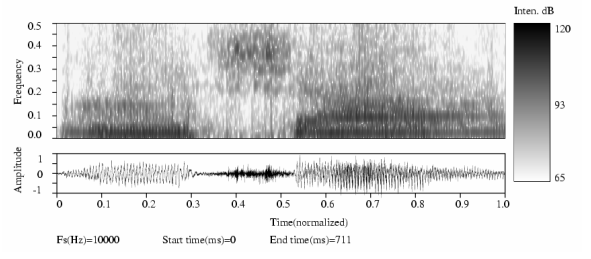
- The quality of the synthesized vowels and syllables is intelligible and natural. Harmonic part is sufficient for good quality and intelligibility except for /*aʃa*/ and /*asa*/. Synthesized harmonic parts of these two syllables sound almost the same as confirmed by listening and spectrograms (Figure 4, Figure 5, and Figure 6). After adding synthesized noise part, sounds become well differentiable. So it is not necessary to store the parameters of noise for all syllables. This fact is very useful for reducing the size of the database.
- It was observed that the HNM could synthesize all the passages spoken in different styles and at different articulatory rate.



(a) *Recorded /aʃa/*

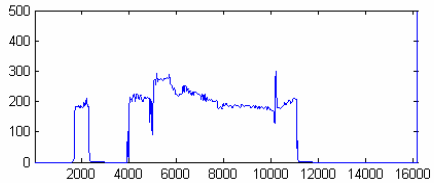


(b) *Synthesized harmonic part of /aʃa/*

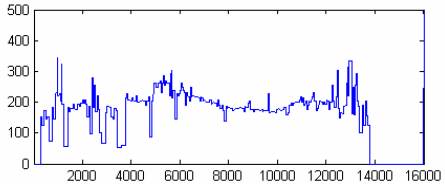


(c) *Synthesized harmonic + noise part of /aʃa/*

Figure 6. *Spectrograms for synthesized /aʃa/.*



(a) *Pitch contour obtained from glottal waveform*



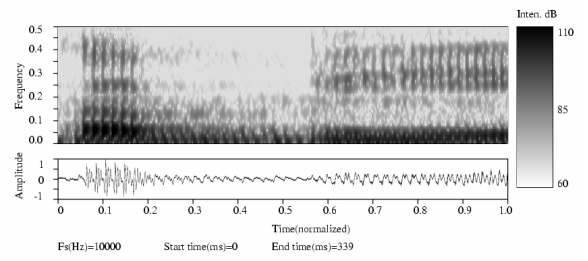
(b) *Pitch contour obtained from speech waveform*

Figure 7. *Pitch contours for the passage "/ap kʌhan ja rʌhE hæn/"*

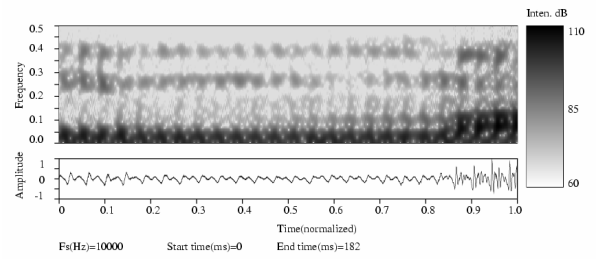
For synthesizing these passages, two options were tried. In the first option, the GCIs were extracted from the speech signal. In the second the glottal signal was used to extract these GCIs. The speech quality with GCIs from glottal signal

is better than with the GCIs from speech signal. Figure 7 shows the pitch contours obtained from speech and glottal signal for a sentence.

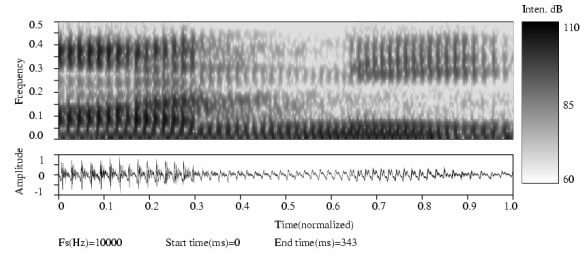
Figure 7 also shows mismatching at certain instants between the pitch contour obtained from the glottal signal and the one obtained from the speech. This mismatching is due to the limitation of the algorithm employed for determination of pitch.



(a) /ʌb^{h}I/



(b) /*Ima*/

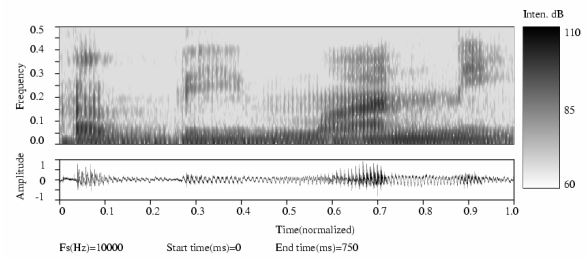


(c) /*ani*/

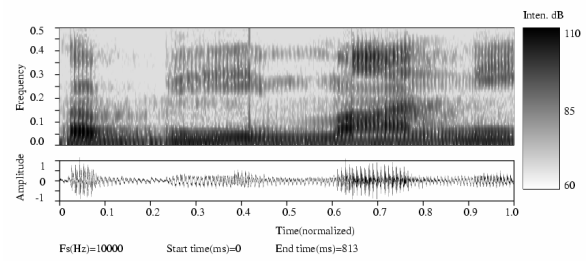
Figure 8. *Spectrograms for isolated syllables used in the synthesis of the word /ʌb^{h}Imani/.*

• When larger speech units are constructed by concatenating the parameter tracts of smaller units, the quality is not affected much. For example /ʌb^h*Imani*/ can be generated using three VCV syllables /ʌb^{h}I/, /*Ima*/, and /*ani*/ spectrograms for these three syllables are shown in Figure 8. After adjusting the parameters for pitch contour smoothening and energy normalization, the synthesis is done. For comparison, two methods are tried for concatenation of the three units. In the first, individual units are synthesized and concatenation is done in time-domain. Second method used concatenation in parametric domain followed by synthesis.

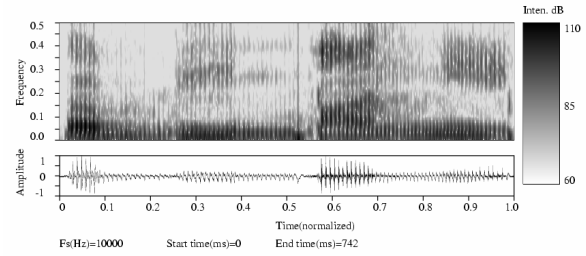
For comparison, /ʌb^h*Imani*/ is also recorded as a single word. The spectrograms for the recorded word and the two synthesized versions are shown in Figure 9. Listening tests and spectrograms show that concatenation in time-



(a) *Recorded*



(b) *Concatenation in time-domain*



(c) *Concatenation in parametric domain*

Figure 9. *Spectrograms for the recorded and synthesized word /ʌb^{h}Imani/.*

domain degrades the quality slightly due to discontinuity at the joints.

5. CONCLUSIONS

Investigations with some peculiar sounds of Indian languages showed that one difference between aspirated and unaspirated stops is the release duration. Unvoiced aspirated stops can be synthesized using the corresponding unaspirated stops by increasing the release duration. Voiced aspirated and unaspirated stops behave in a

different manner. Voiced unaspirated stops can be synthesized by reducing the release duration of corresponding aspirated stops.

The quality of the synthesized sound using HNM is satisfactory. It can synthesize 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*/. This fact can reduce the number of parameters needed for these sounds. 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. Larger speech units can be constructed from smaller units without loss of appreciable quality, naturalness, and intelligibility.

Further work is being carried out on pitch modification, generation of parameter tracts involving allophones, modification of articulatory rate, and other speaker characteristics.

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

[7] 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.

[8] 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.

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

[10] E. Moulines and J. Laroche, "Techniques for pitch-scale and time-scale transformation of speech. Part I. Non parametric methods," *Speech Communication*, vol. 16, 1995.

[11] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453-467, 1990.