ON THE DEVELOPMENT OF TEXT TO SPEECH SYSTEM FOR HINDI

Rajesh Verma, A.Sada Siva Sarma, Nisheeth Shrotriya, Anil Kumar Sharma and S.S.Agrawal

Speech Technology Group, Central Electronics Engineering Research Institute Centre, CSIR Complex, Hill Side Road, New Delhi - 110 012.

ABSTRACT

Recently, the feasibility of using KLSYN88 has been shown to synthesize Hindi speech sounds including voiced/unvoiced and aspirated/ unaspirated stop consonants and trills with good quality [1-2]. The synthesizer was subsequently implemented on a PC(AT) and slightly modified to suit synthesis of Hindi aspirated consonants. Syllables were used to generate words after framing the joining rules and sentences were made using these words. This paper describes the feasibility and approach to develop a PC based Text To Speech (TTS) system for Hindi.

INTRODUCTION

Synthesizing high quality speech or natural sounding speech by machines/ computers has been a frontier research area for several decades. Efforts have been made to develop source coding techniques and modelling of the vocal tract system to achieve the above goal. During the past few years, major advances have been made with two terminal analogue formant synthesizers, the Klatt synthesizer [3] and the Holmes synthesizer, and both of them have been adopted for commercial use.

SPECIFIC FEATURES OF HINDI SOUNDS

The Hindi consonants possess certain special features which are not so common to European languages and American English [2]. The most significant differences are in stops and affricates which use both voicing and aspiration, to distinguish them from other languages. For this reason, the aspiration source of KLSYN88 has been modified to pull down the energy to 300 Hz onwards, to generate more natural sounding voiced/unvoiced aspirated sounds. The trills /r/ and /l/ have large allophonic variations in different contexts. Consonant clusters like CCV, CCCV, CCVC etc. also occur frequently in Hindi speech. It is therefore necessary to study these specific characterstics of clusters as a separate category.

TEXT TO SPEECH CONVERSION SYSTEM

Syllables have been chosen as the basic units of Hindi speech to generate an unlimited vocabulary in the proposed TTS system. The most frequently occurring 29 consonants and 10 vowels (5 long and 5 short) in Hindi give rise to 290 CV and 290 VC syllables in all, that form major part of the **database**. Experiments have been conducted to generate short vowels out of corresponding long vowels, using some durational and frequency rules. Therefore number of syllables to be generated have been limited to 290 only. In addition, a special class of around 150 clusters have been included as part of the **database**. Therefore about 500 basic units would be adequate to generate unlimited vocabulary.

The basic building blocks of a text to speech conversion system are shown in *figure 1*. Text Input from the keyboard is fed to the **word parser** . Therefore

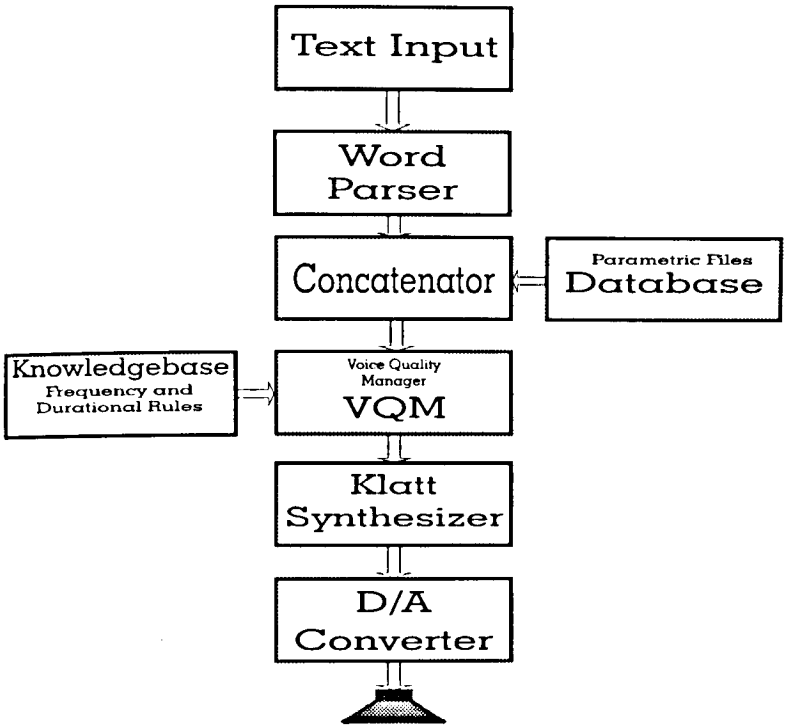

Figure 1. Block diagram of TTS system

word parser identifies the basic syllables, available in **database**, using which the words are generated. These file names are given to the **concatenator**, which picks up files from the **database** and merges them properly to make a new parametric file for the given input word.

Once the doc files are merged and a sound file is generated, the quality of sound is not very good because of discontinuities in sounds at the boundaries of the syllables. Therefore, a **knowlegebase** of durational and frequency rules is provided to voice quality manager **VQM** which applies these rules to the parametric file given by the **concatenator** to smooth out the discontinuities at the syllable boundaries. Finally this parametric file is sent to the **Klatt Synthesizer** to generate the sound file. This sound file is then fed to a loud speaker through a **D/A converter** card. A blank character or a punctuation mark acts as a delimiter for a word. Hence sound output is given word by word.

PROCEDURE FOR ANALYSIS AND SYNTHESIS

The 29 consonants have been recorded by single male speaker having a standard Hindi speaking background and mother tongue, directly in to a PC/AT 386 computer which is equipped with a DSP56000 based SENSIMETRICS speech

station hardware and software.

The syllables were analysed using KAY Sonograph, Sensimetrics speech station and CD_SPEC[4] analysis program having facilities of displaying LPC/FFT Spectrum of windowed segment, pitch and energy etc.. All the important source and vocal tract parameters were extracted and tabulated for use in synthesis.

Based on the analysed data, a preliminary parametric (DOC) file was created and synthesis was done to obtain a starting synthetic file. A Set of 60 parameters [2] have been used for creating a synthesized document file. Then the spectrograms of natural and synthetic syllables were compared. A number of source and tract parameters were adjusted iteratively, in order to achieve a close imitation to natural CV/VC syllables.

RESULTS

All the twenty-nine frequently occurring consonants in combination with five long vowels have been synthesized to obtain 145 CV and 145 VC combinations. Spectrograms showing the comparison of original and synthetic sounds for syllable /t ʃ a/ are shown in *figure 2.*

Experiments have been done to generate words made of simple CVCV type syllables. The words were fed through keyboard using equivalent ASCII codes of the syllables. Durational and frequency rules have been applied at the syllable boundaries to smooth the discontinuities. *Figure 3* shows three spectrograms of the word /t ʃ at ʃ a/. In this figure the first spectrogram does not include any rule, the middle spectrogram shows the application of durational rule

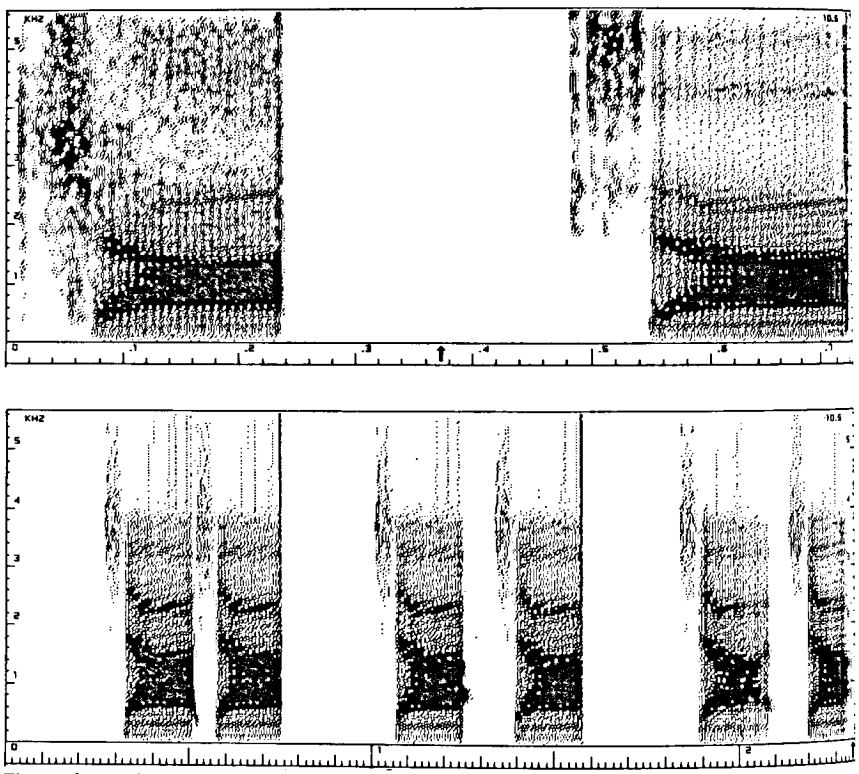

Figure 3. Application of durational and frequency rules to improve sound quality

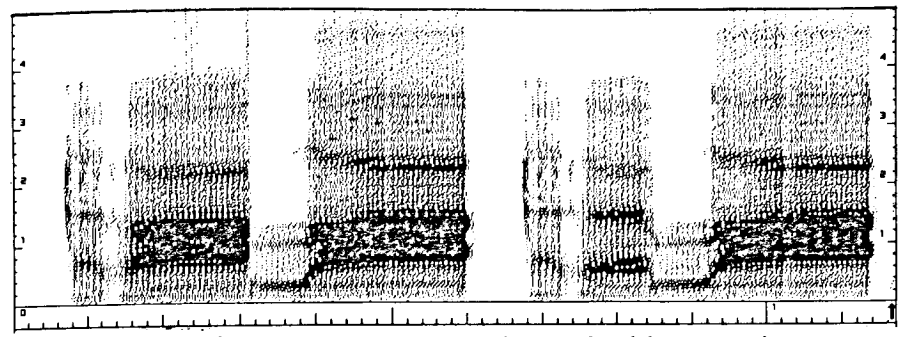

Figure 4. Coversion from /rama/ to /rʌma/ using durational and frequency rules

and the last spectrogram shows the effect of both durational as well as frequency rule to get a close imitation of the original sound.

Figure 4 shows the application of the rule for creating short vowel /ʌ/ from long vowel /a/. It shows the CVC type of word /rama/ and it is intended to make it /rʌma/. To achieve this two rules have to be applied. First, the vowel duration has to be reduced from 200 ms to 100 ms. Second, the first formant frequency is to be pulled down by 200 Hz and second formant frequency is to be pushed up by 200 Hz, in the pure vowel region. These rules lead to creation of vowel /ʌ/ from /a/. Different set of rules are being formed for different vowel contexts.

CONCLUSIONS

Feasibility of using KLATT synthesizer for close imitation of natural Hindi sounds have been shown. All the CV and VC syllables required for the **database** have been generated. Most frequently occurring clusters in Hindi have been selected. Work is under progress to synthesize those clusters as part of the **database. Word Parser** is currently working for words made of CV syllables only (maximum up to ten syllables) and is being updated. **Concatenator** block has been fully developed. **VQM** is currently working for few durational rules only. Some of the durational and frequency rules have been studied and further study is in progress for creation of **Knowledgebase.**

The study of the suprasegmental features, rules for intonation and stress patterns is in progress, for this system to act as a reading machine for blinds.

ACKNOWLEDGEMENTS

The authors are grateful to Prof R N Biswas, Director, CEERI, Pilani and DoE/UNDP for their encouragement and support. Award of SRF by CSIR to one of the author (NS) is highly acknowledged.

REFERENCES

[1] Agrawal S. S., and Stevens K., "Towards synthesis of Hindi consonants using KLSYN88," Proc. ICSLP-92, 177-180, 1992.

[2] Neal Pinto, et.al.,"Synthesis of Hindi CV Syllables in Three Vowel Contexts using the PC-KLATT Cascade/ Parallel Formant Synthesizer", Proc. 3rd ICAPRDT-93, ISI, Calcutta, India, Dec 28-31, 1993, p. 354.

[3] Klatt D H and Klatt L, "Analysis, synthesis and perception of voice quality variations among female and male talkers", JASA,87(2), 820-857,1990

[4] P K Dhanarajani, et.al.,"A PC based Graphic Tool for Analysis, Segmentation and Labelling of Speech Signals", Proc. 3rd ICAPRDT-93, ISI, Calcutta, India, Dec 28-31, 1993, p. 326.