4.1 Building speech corpora for unit selection based concatenative text to speech system for Indian Languages

Natural Language Processing Lab Centre for Development of Advanced Computing
'Anusandhan Bhawan', C 56/1 Sector 62, Noida – 201 307, India

Background Information

Development and selection of speech corpora would essentially depend on the following two important criteria

1 The technology used for the development of the speech synthesis system

2 The application in mind

Unit concatenation speech synthesis systems have gained popularity, due to their more natural sounding output. For this different sized speech units can be used by different speech synthesis system developers. To get these units the phonetically rich text material needs to be selected and then get recorded. The recorded speech to be annotated and time aligned in terms of the speech units (SU)or phonetic data units(PDU)

In CDAC Noida type of speech units that have been used for the development of speech synthesis system are multiform units. These multiform units primerily consists of syllables & phonemes. However, in addition, most frequent words, vocabulary of monetary & temporal expressions have been used.

Text Corpora consisting of about 21 lakh words containing 77000 unique words have been used for selection and annotation of speech units.

A tool 'Vishleshika' has been used for analysis and selection of text Corpora for most frequent words, syllable & phoneme speech units.

This document comprises of brief introduction of different modules in the Hindi Text to Speech Synthesis system, different units in speech database, Hindi Syllable Types & their Distribution. It also describes the corpora development methodology including Selection of Textual content, Recording, Annotation and Structural Storage of Corpora detail.

1. Introduction

The Hindi Text to speech system can be divided into different module:

- Preprocessing module
- Phonetic data Unit searching & selection Synthesis module
- DSP Module

1.1 Preprocessing module

The Preprocessing module consists of different sub modules: Normalizer, Chunker, G2P converter, Schwa deletion, syllabification & segmentation. The Normalizer identifies and expands the string patterns of abbreviations, acronyms, telephone numbers, special formats like e-mail address, brackets, and quotation marks into its pronounceable form. The chunker module identifies the chunks for Temporal & Monetary Values.

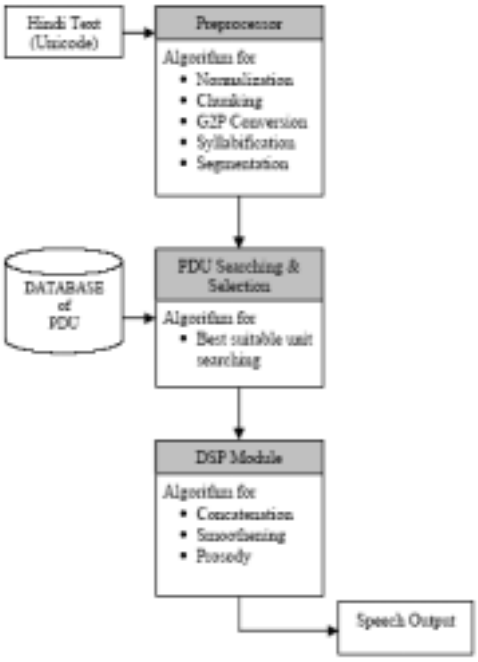

Figure1: Block diagram for TTS

The Phrase identifier identifies the phrase in the sentences by just looking into the phrase lexicons. The G2P converter coverts the strings into its orthographic representation. Then a module for automatic schwa deletion and segmentation segments the word into

its syllable units of $C_0^3VC_0^3$ form. The overall output of the preprocessor is tagged sentences. This output goes to the Phonetic data Unit searching & selection module and then subsequently to the DSP Module.

1.2 Phonetic Data Unit searching & selection module

Phonetic Unit database is categorized into phrases, words, syllables & diphone database. The units are selected from the Phonetic Units database. In this case the appropriate phonetic material was selected from the GyanNidhi corpus. This selected material was recorded priorily and then the recorded speech was carefully labeled using the semiautomatic annotation tool. The output in XML format is divided into the constituents phonetic data units like phrases, words, syllables, and diphones in a database structure.

The Phonetic Data Unit searching & selection algorithm compares the input phonetic description with the speech database information and selects the most appropriate unit depending on the context of input text.

1.3 DSP Module

After the selection of units from the database, they are concatenated. The endpoints of the amplitude are matched to prevent the sudden amplitude jumps in the concatenated units. Appropriate pauses are also imposed to maintain the flow of the sentence. Then the pitch-smoothening algorithm is applied to the synthesized speech. The overall quality of the generated speech will be controlled by the prosody methods.

2. Units of Speech Database

Multi-form units like syllables, phonemes and most frequent words are constituents of the database. In the speech synthesis system being discussed here, syllable is chosen as the main unit for generating synthesized voice. Sounds for which syllables are not covered in the database, phonemes are used as supplementary units.

Commonly occurring syllables (pure and nasalized) are classified as in Table 1. The distribution of these syllables in large corpus is presented in Table 2. Here c, v, and V stand for consonant, vowel and nasalized vowel respectively. It shows that in Hindi 'cv' type syllables have highest and very large percentage of occurrence.

2.1 Hindi Syllable Types & their Distribution

All Indian languages scripts are phonetic in nature and thus share a common phonetic base. The basic unit of Indian languages scripts is characters, which are orthographic representation of speech sounds.

To extract phonetically rich sentences [2] we have considered syllable as the basic unit and used schwa deletion and syllabification rules to syllabify the text. A syllable is a unit of speech that is made up of one or more phones (single sounds or "phonetic segments") and in turn makes up words. It influences the rhythm of a language, its prosody, its poetic meter, its stress patterns, etc.

In Hindi, words are composed of syllables of type $C_0^3VC_0^3$ e.g. /kʌmʌl/ (Lotus) consists of syllables /kʌ/ and /mʌ/ of type 'cv' and /strɪtv/ (womanhood) consist syllable /strɪ/ of type 'cccv'.

For in-depth analysis, we are also taking the position of the syllable into consideration, at which it appears in a word e.g. in /kam/ (work), /mʌkan/ (house), /ʊsʌka/ (his/her) and /ka/ (of) syllable appears at starting, middle, end and isolated positions in the word respectively.

Table 1: Different types of syllables in Hindi

Pure syllables	
v	vowel
cv	consonant-vowel
vc	vowel-consonant
ccv	consonant-consonant-vowel
cvc	consonant-vowel-consonant
vcc	vowel-consonant-consonant
ccvc	consonant-consonant-vowel-consonant
cvcc	consonant-vowel-consonant-consonant
vccc	vowel-consonant-consonant-consonant
ccvcc	consonant-consonant—vowel-consonant-consonant
cvccc	consonant-vowel-consonant-consonant-consonant
Nasalized syllables	
cV	consonant-nasalized_vowel
ccV	consonant-consonant-nasalized_vowel
cVc	consonant-nasalized_vowel- consonant
ccVc	consonant-consonant-nasalized_vowel- consonant

Distribution of different more frequent syllables in the large corpus is given in the Table 2, which shows that 'cv' and 'cvc' type of syllables constitute almost 80% of the syllables in the text.

Table 2: Distribution of some of syllables in large corpus

Syllable Type	%age of occurrence
cv	48.8
cvc	31.2
cV	6.1
v	3.9
cvcc	2.9
vc	2.7
ccvc	1.2
V	1.2
ccv	0.9
cVc	0.6

The speech database used in the system covers different types of syllables as per the distribution shown in the above table. A large corpus (GyanNidhi) is statistically analyzed using a Statistical Analysis Tool (Vishleshika [2]). *Vishleshika* performs several types of analysis ranging from simple frequency counts to extraction of linguistic features.

The syllables are chosen in four contexts (i.e. start, middle, end and isolated respectively). Here is an example sentence from phonetically rich set of sentences, marked for unique syllables. The syllables are marked with 1, 2, 4, and 8 (i.e. start, middle, end and isolated respectively) which represent the position at which these syllables occurred in the word.

(विष्णु) इतने (दयालु) (न) (बनो), (ज़रा) (सोचो) हम (खिलौने) और (जेयर) (महापौर) के (सुपुर्द) क्यों करें।
4:णु 4:लु 8:न 4:नो 1:ज़ 1:सो 4:चो 1:खि
2:लौं 1:ज़े 2:पौर 2:पु

wɪʃɳu ɪtne dʌyʌlʊ nʌ bʌno dʒʌra sotʃo
hʌm kʰɪlone ɔr dʒewʌr mʌhapɔr ke
supurd kyõ kʌrẽ

2.2 Syllable Boundary Identification

For searching the syllables from speech database, the syllables have to be identified in the text which is to be synthesized in speech form. The syllable boundary identification process involves schwa deletion and syllable identification rules. These rules have been made and implemented .

3. Corpora development methodology

Speech Corpora design for the Text to Speech Synthesis System in Indian Languages involves the following four steps:

- Selection of Textual content
- Recording of Textual content
- Annotation of Speech signal
- Structural Storage of Corpora

3.1. Selection of Textual content

The objective of the text corpora design is to construct a minimum but sufficient speech corpus for concatenative TTS reflecting the speaker's natural voice. For practical use, it is also important that by using the created corpus, words and sentences that appears in the text is synthesized naturally and comprehensively. The Figure 2 display the overall methodology for minimum but sufficient text content selection

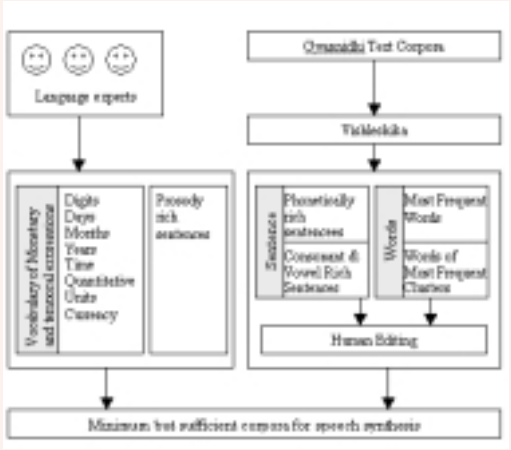

Figure 2 Block diagram for content selection

3.1.1 Vocabulary

To generate high quality synthesized speech for all monetary and temporal expressions with the limited set of words, a vocabulary of unique 250 words was created. The list covered words related to Digits,

Days, Months, Years, Time, Quantitative Units, and Currency etc.

3.1.2 Most Frequent Words

A significant portion of the text is covered by most frequently used words. Hence these databases contribute towards better quality-synthesized speech. In the present module 1000 most frequent words from Gyannidhi corpus [3] have been selected which covers approximately 50% of Speech.

In addition most frequently clustered words also have been taken into its original form thus covering major portion of the total clustered words.

3.1.3 Phonetically rich sentences

ype monosyllables in maximum numbers. These syllables may occur at starting, middle or at end positions in the word or it may occur in isolation. A typical set of Phonetically rich sentences created using Vishleshika[1,2] from GyanNidhi Corpus [3] consist of 500 sentences. These sentences were examined by linguists to break the longer sentences into smaller ones and to make these smaller sentences meaningful so that they are easy to read & examine the sentences.

3.1.4 Consonant & Vowel (Phoneme) Rich Sentences

Acoustic properties of a phoneme depend significantly on the adjoining phonemes. So there is a need to have phoneme sounds in all possible bigram contexts. A typical set of Consonant & Vowel Rich Sentences created using Vishleshika from GyanNidhi Corpus consist of 450 sentences.

3.1.5 Prosody rich sentences

The type of sentences influences the prosodic patterns. Therefore a set of about 1000 prosody rich sentences was created with the help of linguists. This set contains sentences reflecting anger, joy, and sadness, question type sentences, negative, command, exclamations etc.

3.1.6 Domain specific text

The text related to the domain of News, Scientific Text, Plays was selected from different sources from which different sets of phonetically rich sentences were also made.

3.2 Recording of Textual content

Professional speakers (Male & Female) were used for recording to maintain constant pitch and prevent stress phenomenon. The recording environment was noise free & echo cancelled studio and the speech samples were recorded at a sampling rate of 44.1kHz (16 bit) in stereo mode.

3.3 Annotation of Speech signal

Annotation is performed to cover descriptive or analytic notations applied to raw language data including some additional information. The recorded corpora thus needs to be annotated properly depending on the syllable units. The speech units are tagged in a hierarchical manner consisting of sentence, word, syllable and phoneme levels.

The labeler tool generates an XML file that consists of structured information regarding the tagging. Unicode support in the tool enables annotation in Indian script.

3.4 Structural Storage of Corpora

The Corpora is stored in a structural form in order to maintain uniformity of the information stored. This will aid in maintenance of the corpora for future enhancement. The signal data files and annotated data are stored separately in the following directory structure

- META_DATA.
- TEXT_DATA
- RAW_SPEECH.
- ANNOT_SPEECH.
- SPEECH_DATABASE. .
- TOOLS.

3.4.1 META_DATA:

It contains all metadata information files.

Metadata information

The term Meta data for speech recordings refers not to the recorded speech data itself, but to data about these recorded data. Meta data consists of categorized, machine –readable data that may be used to classify the speech data contained in speech data. As the

diversity of speech corpora is growing quickly, it is not usually possible to access the speech data itself to check, but it is always possible to access Meta data since it is formal description of the underlying speech corpus.

This present corpora of Indian Languages have three type of Meta data which is stored in XML Format in meta data directory.

- Recording information Meta data
- Speaker profile Meta Data
- Content Type Meta data

In Recording information Meta data, all type of recording information about the actual recording is logged. e.g. Recording Date, Recording Time, Recording environment, Recording Setup, Technical Specifications e.g. Sampling Rate, Stereo/Mono, Number of Channel Signal File Format, Recording software Sample size 16 bit etc. If the speech corpus contains only recording under exactly the same conditions, only one recording meta information for the complete corpus is stored.

In Speaker profile Meta Data, the information like Name, Age, Sex, Address, Occupation, Mother Tongue etc is stored.

In content type meta data, information like Content Language, Content type, Content domain, Content source is recorded.

3.4.2 TEXT_DATA

It contains the text files of the recorded speech.

3.4.3 RAW_SPEECH:

It contains recorded speech files corresponding to each text file.

3.4.4 ANNOT_SPEECH:

It contains annotated speech in XML Format. The following figure-3 displays the four tiers of sentences, words, syllables & phoneme along with some information about the speaker .Each PDU have information about its position (Tkey) in the word, and the surrounding contextual units (Cleft, Cright).

3.4.5 SPEECH_DATABASE:

The XML Files that accompany the speech data and other information are stored into the SQL Tables of different phonetic units e.g. sentence words, syllables etc. to facilitate fast search Figure-3

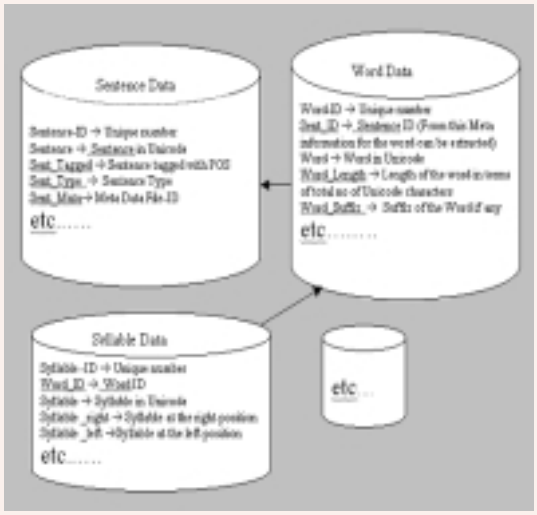

Figure 3 Speech Database

3.4.6 TOOLS:

It contains software & tools like

- Software used to annotate & segment the data,
- Software to converts XML into databases
- Software to converter Database table to XM format etc.

4. Conclusion

Most Frequent Words, Vocabulary of digits, Time, day, months, year, units, & the initial set of 500 Hindi sentences designed using this methodology has been recorded. They have been segmented & labeled into different speech units as required for development of

a Hindi speech synthesis system. Further work is continuing in this direction.

5. Acknowledgement

We would like to thank Dr. Om Vikas, Sr Director, MC&IT and Shri R K Verma, Executive Director, Centre for Development of Advanced Computing, Noida for constant support and conducive environment for this work. Special thanks are due to Prof K K Goswami for providing linguistic support.

We are grateful to Sh. C K Joseph Controller (Admn) CDAC Noida and Sh R M Goel Controller (Finance) for providing necessary tools and infrastructure required for conducting this research work.

5. References

[1] Sunita Arora, et al (2002) Vishleshika: *Statistical Text Analyzer for Hindi and other Indian languages*, TIFR, Mumbai

[2] Karunesh Arora, Sunita Arora, Kapil Verma, S S Agrawal, *"Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages"*INTERSPEECH2004-ICSLP Jeju, Korea

[3] Karunesh Kr. Arora, Sunita Arora, Vijay Gugnani,V N Shukla, Dr S S Agrawal , *GyanNidhi: A Parallel Corpus for Indian Languages including Nepali* ITPC, 2003

[4] Akemi Lida, Nick Campbell. *A database design for a concatenative speech synthesis system for the disabled*. In Proceedings of ISCA 4th International Workshop on Speech Synthesis, pp.188-194

[5] Alan W. Black, Kevin A. Lenzo. *Optimal Data Selection for Unit Selection Synthesis.* In 4th ESCA Workshop on Speech Synthesis, Scotland., 2001.

[6] Esther Klabbers, Karlheinz Stober, Raymond Vendhuis, Petra Wagner, Stefan Breuer. *Speech Synthesis Development Made Easy: The Bonn Open Synthesis System*. In: Proc. Eurospeech'2001, Aalborg, Denmark, Vol. 1, pp. 521-524

[7] P. Stathopoulou-Zois. *The UOP Text-to-Speech System for Greek Speech Synthesis.* In Proceedings of ISCA 4th International Workshop on Speech Synthesis, paper 208.

[8] Baris Bozkurt, Michel Bagein, Thierry Dutoit. *Demo System for NU-MBROLA concatenator.* Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis, pp. 155-156, Blair Atholl, Scotland, 2001.