Experiments with Unit Selection Speech Databases for Indian Languages

S P Kishore †§, Alan W Black ‡, Rohit Kumar †, and Rajeev Sangal †

† Language Technologies Research Center
International Institute of Information Technology, Hyderabad

‡ Language Technologies Institute, Carnegie Mellon University
§Institute for Software Research International, Carnegie Mellon University

Abstract

This paper presents a brief overview of unit selection speech synthesis and discuss the issues relevant to the development of voices for Indian languages. We discuss a few perceptual experiments conducted on Hindi and Telugu voices.

1 Role of Language Technologies

Most of the Information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so that digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages.

These technologies play a crucial role in multi-lingual societies such as India which has about 1652 dialects/native languages. While Hindi written in Devanagari script, is the official language, the other 17 languages recognized by the constitution of India are: 1) Assamese 2) Tamil 3) Malayalam 4) Gujarati 5) Telugu 6) Oriya 7) Urdu 8) Bengali 9) Sanskrit 10) Kashmiri 11) Sindhi 12) Punjabi 13) Konkani 14) Marathi 15) Manipuri 16) Kannada and 17) Nepali. Seamless integration of speech recognition, machine translation and speech synthesis systems could facilitate the exchange of information between two people speaking two different languages. Our overall goal is to develop speech recognition and speech synthesis systems for most of these languages.

In this paper we discuss the issues related to the development of speech synthesis systems for Indian languages using unit selection techniques. This work is done within the FestVox voice building framework [1], which offers general tools for building unit selection synthesizers in new languages. FestVox offers a language independent method for building synthetic voices, offering mechanisms to abstractly describe phonetic and syllabic structure in the language. It is that flexibility in the language building process that we exploited to build voices for Indian languages. Voices generated by this system may be run in the Festival Speech Synthesis System [2].

2 Speech Synthesis Systems

The objective of a text to speech system is to convert an arbitrary given text into a corresponding spoken waveform. Text processing and speech generation are two main components of a text to speech system. The objective of the text processing component is to process the given input text and produce appropriate sequence of phonemic units. These phonemic units are realized by the speech generation component either by synthesis from parameters or by selection of a unit from a large speech corpus. For natural sounding speech synthesis, it is essential that the text processing component produce an appropriate sequence of phonemic units corresponding to an arbitrary input text.

3 Text Processing Front End

Before we discuss the issues related to text processing, let us briefly understand the nature of the text for which the synthesis systems are built.

3.1 Nature of the Scripts of Indian Languages

The basic units of the writing system in Indian languages are Aksharas, which are an orthographic representation of speech sounds. An Akshara in Indian language scripts is close to a syllable and can be typically of the following form: C, V, CV, CCV, VC and CVC where C is a consonant and V is a vowel. All Indian language scripts have a common phonetic base, and an universal phoneset consists of about 35 consonants and about 18 vowels. The pronunciation of these scripts is almost straight forward. There is more or less one to one correspondence between what is written and what is spoken. However, in languages such as Hindi and Bengali the inherent vowel (short /a/) associated with a consonant is not pronounced depending on the context. It is is referred to as inherent vowel suppression or schwa deletion.

3.2 Format of Input Text

The scripts of Indian languages are stored in digital computers in ISCII, UNICODE and in transliteration schemes of various fonts. The input text could be available in any of these formats. However, the issue of various formats could be conveniently separated from the synthesis engine. As Indian languages have a common phonetic base, the engines could be built for one transliteration scheme which can represent the scripts of all Indian languages. The text processing front end can provide appropriate conversion of various formats of input text into this transliteration scheme.

3.3 Mapping of Non-Standard Words to Standard Words

In practice, an input text such as news article consists of standard words (whose entries could be found in the dictionary) and non-standard words such as initials, digits, symbols and abbreviations. Mapping of non-standard words to a set of standard words depends on the context, and it is a non-trivial problem [3]. For example, digit 120 has to be expanded to /nuut'aa iravai/, "Rs 3005412" to /muppia laqs-alaa aidu velaa, nalagu van'dalaa pannen'd'u ruupayalu/, and "Tel: 3005412" to /phon nambaru, muud'u sunnaa sunnaa, aidu nalagu okati rend'u/. Similarly punctuation characters and their combinations such as :, >, !, -, $, #, %, / which may be encountered in the cases of ratios, percentages, comparisons have to be mapped to a set of standard words according to the context.

Other such situations include initials, company names, street addresses, initials, titles, non-native words such as /bank/ /computer/ etc.

3.4 Standard Words to Phoneme Sequence

Generation of sequence of phonemic units for a given standard word is referred to as letter to sound rules. The complexity of these rules and their derivation depends on the nature of the language. For languages such as English, a pronunciation dictionary of about 125,000 words is used along with a set of letter to sound rules to handle unseen words. For Indian languages such as Telugu, the letter to sound rules are relatively easy due to their phonetic nature, i.e., there is a fairly good correspondence between what is written and what is spoken.

However, for some of the Indian languages such as Hindi and Bengali, the rules for mapping of letter to sound are not so straight forward. In Hindi, the inherent schwa associated with a consonant is suppressed depending on the context. For example, words such as /kamala/ and /dilachaspa/ are pronounced as /kamal/ and /dilchasp/ respectively. Understanding of this phenomenon is important to build a good text processing module and thus to generate natural sounding speech synthesis in Hindi and other such languages. At present, we are using a small set of heuristic rules and are working on machine learning techniques for schwa deletion.

4 Speech Generation Component

Given the sequence of phonemes, the objective of the speech generation component is to synthesize the acoustic waveform. Speech generation has been attempted by articulatory model based techniques and parametric based techniques [4] [5]. While the articulatory models suffer from the lack of adequate modeling of motion of articulators, the parametric models require a large number of rules to manifest coarticulation and prosody. An alternative solution was to concatenate the recorded speech segments. The inventory of these recorded speech segments were limited to a small set of units which have sufficient coarticulation such as diphones and a set of rules were used to manipulate the prosody.

Current state-of-art speech synthesis systems generate natural sounding speech by using an inventory of large number of speech units with differing prosody [6]. Storage of large number of units and their retrieval in real time is feasible due to availability of cheap memory and computation power. The approach of using an inventory of speech units is referred to as unit selection approach. It can also be referred to as data-driven approach or example-based approach for speech synthesis. The issues related to the unit selection speech synthesis system are: 1) Choice of unit size, 2) Generation of speech database and 3) Criteria for selection of a unit.

The objective criteria for selection of a unit depends on how well it matches with the input specification and how well it matches with the other units in the sequence. Costs are associated for mismatch with the input specification and with other units in sequence, and are referred to as target cost and concatenation cost respectively. A unit which minimizes the cost of target and concatenation cost is selected from the speech database.

4.1 Choice of Unit Size

An inventory of larger size of units such as sentences, phrases and words with differing prosody could constitute an ideal speech database for speech generation. However, if the size of the units is large, the coverage of all possible words, phrases, proper nouns, and other foreign words may not be ensured.

Subword units make it easier to cover the space of acoustic units but at the cost of more joins. The choice of subword unit is also related to the language itself. Languages with a very well defined, and a small number of syllables may benefit from a syllable sized unit. There have been various suggestions on unit size for unit selection systems. [7] and other HMM-based techniques are typically using sub-phonetic units: two or three per phoneme. AT&T's NextGen [8], uses half phones. As Indian languages have a much more regular syllable structure than English we wanted to experiment to find the optimal sized unit.

4.1.1 Perceptual Response for Different Choice of Unit Size

In order to investigate the optimal unit size we built Hindi synthesizers for four different choices of unit size: **syllable, diphone, phone and half phone**, and conducted a perceptual evaluation of these synthesizers [9]. A set of 24 sentences were selected from a Hindi news bulletin and these sentences were synthesized by phone, diphone, syllable and half phone synthesizers and were subjected to the perceptual test of native Hindi speakers. Each listener was subjected to AB-test, i.e. the same sentence synthesized by two different synthesizers was played in random order and the listener was asked to decide which one sounded better for him/her. They also had the choice of giving the decision of equality.

The results of AB-test conducted on 11 persons in the case of syllable and diphone synthesizers and on 5 persons for the rest of the synthesizers. From the perceptual results, it was observed that the syllable unit performs better than diphone, phone and half phone, and seems to be a better representation for Indian languages. It was also observed that the half phone synthesizer performed better than diphone and phone synthesizers, though not as well as syllable.

4.2 Generation of Unit Selection Speech Databases

There are two issues concerning the generation of unit selection databases. They are: 1) Selection of utterances which has the coverage of all possible units and with all possible prosody and 2) Recording of these utterances by a good voice talent. Selection of utterances is linked with the choice of the unit size. The larger the size of the unit the larger would be the number of utterances for the coverage of the units.

To generate unit selection speech databases, we selected a set of utterances covering most of the high frequency syllables in Hindi and Telugu. A syllable is said to be a high frequency syllable if its frequency (occurrence) count in a given text corpus is relatively high. We used the large text corpus available with frequency count of the syllables in Indian languages [10]. This text corpus contains text collected from various subjects ranging from philosophy to short stories. We selected sentences from this text corpus if it contained at least one unique instance of a high frequency syllable, not present in the previous selected sentences. These sentences were examined by a linguist primarily to break the longer sentences into smaller ones and to make these smaller sentences meaningful and easy to utter. The statistics of the set of utterances selected for Hindi and Telugu are show in Table 1.

Table 1. Syllable coverage and duration of Hindi and Telugu Speech Database

Language	No. Utt	No. Uniq. Syllables	Total Syllables	Dur. of Speech
Hindi	620	2324	22960	90 m
Telugu	1100	3394	32295	125 m

These sentences were recorded by a native female speaker for Hindi and a native male speaker for Telugu. The female speaker had an experience of working in Radio station for reading news articles, while the male speaker went through a couple of practice sessions before final recording. The utterances were read in a natural fashion, but with out any emotions or excitement which otherwise would be associated with the message in a typical news broadcast. The recording was done in a noise free environment using a multi-media desktop and a noise cancellation microphone.

The speech signal was sampled at 16 KHz. The speech database was labeled at the phone level using Festvox tools and the label boundaries were hand-corrected using Emulabel. The syllable level boundaries were derived from these corrected phone level boundaries. The total duration of the speech data recorded for Hindi and Telugu are shown in Table 1.

4.2.1 Perceptual Response for Varying Coverage of Units

Table 2. Syllable coverage and duration of speech databases derived from D6

DataBase	Duration	No. of Sen.	No. of Syl.	Total Syl.
D1	10	100	725	2681
D2	30	300	1548	8032
D3	52	500	2187	13738
D4	76	700	2622	19665
D5	99	900	3019	25450
D6	125	1100	3394	32295

In order to investigate the quality of speech synthesizer and its dependency on the coverage of units in the database, we conducted the following perceptual experiment on Telugu synthesizer. We derived six sets of speech databases with difference in the coverage of syllables as shown in Table 2. Six different synthesizers S1 S2 S3 S4 S5 S6 were built using the databases D1 D2 D3 D4 D5 D6 respectively. These synthesizers were built with syllable as basic unit. If a syllable was not covered in the database, a simple back-off method was used to generate that particular syllable by concatenating a sequence of phones.

A set of five sentences taken from a news paper were synthesized using these different synthesizers. A typical multimedia PC with desktop speakers was used to play these utterances to five native male speakers of Telugu in a calm laboratory environment. The test subjects were around 20 to 25 years of age, and they did not have any prior experience in `speech synthesis experiments. They were asked to give a score between 0-5 (0 for bad quality and 5 for high quality), for each of the five utterances. The mean of the scores given by each individual subject for each of the synthesis systems S1 S2 S3 S4 S5 and S6 is shown in Table 3. The mean and variance of the scores obtained across the subjects for each of the synthesis systems is show in Table 4.

Table 3. Mean of the scores given by each subject for each synthesis system.

Subject → System ↓	P1	P2	P3	P4	P5
S1	2.0	2.8	0.6	2.4	1.2
S2	2.6	3.4	1.8	2.8	2.6
S3	2.8	3.2	2.2	3.4	3.0
S4	3.4	3.6	3.0	3.6	3.2
S5	3.6	3.8	3.2	3.4	3.6
S6	3.6	4.0	3.6	3.4	3.2

Table 4. Mean and Variance of the scores obtained across the subjects.

System	Mean Score	Variance
S1	1.80	0.80
S2	2.64	0.32
S3	2.92	0.21
S4	3.36	0.06
S5	3.52	0.05
S6	3.56	0.08

The results of Table 3 indicate the response of each individual subject for each synthesis system, while the results in Table 4 indicate the overall response of the subjects for each of the synthesis systems. We can observe that mean score increases with the increase in the syllable coverage. Another interesting point could be noted from the variance of the scores. We can see that when the coverage of the speech database is small, the variance in the scores given by all subjects is 0.8 for S1, and it decreases with the increase in the syllable coverage to as low as 0.08 for S6. This shows that the response of each individual vary quite high when the speech quality is bad, and a more consistent response could be obtained when the synthesis quality is high.

5 Evaluation of Hindi Speech Synthesis System

To evaluate the speech synthesis system, we synthesized arbitrary texts available from a major Hindi news portal. We developed a text processing front end to read Hindi text in Unicode, and to handle most of the non-standard words such as date, currency, digits, address abbreviations etc. The schwa-deletion was performed by adding a few more rules to the set specified in [9].

A set of 200 sentences were synthesized and nine native speakers of Hindi were asked to evaluate the quality of the Hindi synthesizer. The synthesizer was built using half phones so that the general coverage of all units is ensured. Each subject was asked to evaluate about 40 sentences randomly chosen from these 200 sentences in the following steps.

- The synthesized wave file was played to the subject once or twice only, without the sentence being displayed.
- The subject was asked to rate the naturalness of the synthesized speech waveform between 0-5 (0 for bad quality and 5 for good quality)
- The sentence was displayed and the subject was asked to click on the words which were Not Sounding Natural (NSN) to him/her. Here the subject was allowed to listen to the synthesized speech any number of times.

The mean of the scores given by each subject and the number of NSN words found by him/her is shown in Table 5.

From the last column of Table 5, we can see that mean score across all the subjects is 3.21 (variance is 0.14) and the percentage of NSN words across all subjects is 9.74%. The full list of NSN words were analyzed and the following observations were made: 1) Around 30% of NSN words were loan words from English language. Ex: cricketer, software, loan, record, yorkshire, shroder etc. 2) Not all proper nouns were pronounced correctly. 3) The rules used for schwa deletion were not successfully deleting schwa for all words 4) The text processing component was not handling all the combinations of punctuation characters with digits and letters. Ex: word! (word), 77" etc.

Table 5. Mean of the scores and the number of errors words given by each subject.

Subject	No. of Sen.	Mean Score	No. of NSN Words Total Words	Error in %
H1	42	3.14	119/1397	8.52%
H2	42	3.26	115/1420	8.10%
H3	45	3.18	120/1336	8.98 %
H4	40	2.75	193/1156	16.7 %
H5	39	3.33	90/1126	7.99 %
H6	39	3.95	87/1315	6.62 %
H7	40	3.30	113/1349	8.38 %
H8	38	2.61	162/1126	14.3 %
H9	41	3.39	113/1188	9.51 %
Overall		**3.21**	**1112/11413**	**9.74%**

These results indicate that that unit selection databases do require additional phonetic coverage for proper nouns and loan words which could be occur more frequently. The studies also highlight need for good text processing component for Indian language synthesizers.

6 Applications of Speech Synthesis Systems

The applications of speech synthesis systems are wide ranging including human-computer interaction, telecommunications, talking books, language education and aid to handicapped persons. Below are a few applications developed at Language Technologies Research Center, IIIT Hyderabad using Hindi and Telugu voices.

6.1 Talking Tourist Aid

Talking tourist aid is built using a limited domain synthesizer to operate on handheld devices. This application allows a non-native Hindi/Telugu person to express his queries/concern related to city, travel, accommodation etc., in the language of the native speaker. It uses the limited domain synthesizer built for this purpose to speak out the queries selected by the user.

6.2 News Reader

The news reader extracts the new items from major news portals, and does a lot of text processing to clean up the text, before giving it to the speech generation component. The current version speaks the text from a Hindi news portal.

6.3 Screen Reader for Visually Impaired

The application of screen reader is helpful for visually handicapped persons. It aids the person to navigate through the screen, and reads out the news items and articles available in Indian language scripts. Apart from the wide range of text processing issues, it also needs to address the issue of seamless navigating across various applications. A preliminary version of this application using Hindi voice has been released for use.

7 Conclusion

In this paper, we discussed the issues relevant to the development of unit selection speech systems for Indian languages. We studied the quality of speech synthesis systems for varying coverage of units in the speech database by observing the scores (response) given by the subjects. It was observed that when the coverage of units is small, the synthesizer is likely to produce a low quality speech, and there would be high variance among the scores given by different subjects. As the coverage of units increases, it increases the quality of the synthesizer and there would be less variance in the scores given by different subjects.

We also conducted an evaluation on Hindi speech synthesis system including its text processing end. The subjects were asked to identify the words which are not sounding natural. The observations on this list of words indicated the effect of loan words should be considered while building the speech corpus, and also highlighted the need for good text processing front end for Indian language synthesizers.

Acknowledgements

We would like to thank Dr. Raj Reddy for the interactions we had with him and for his valuable suggestions and inputs. We also would like to thank Chaitanya Krishna of IIIT Hyderabad for conducting these perceptual experiments and all the persons who participated in these experiments.

References

[1] Black and K. Lenzo, "Building voices in the Festival speech synthesis system," http://festvox.org/bsv/, 2000.

[2] Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," http://festvox.org/festival, 1998.

[3] R. Sproat, Alan W Black, S. Chen, S. Kumar, M. Oste ndorf, and C. Richards, "Normalization of non-standard words," Computer Speech and Language, vol. 15, no. 3, pp. 287 – 333, 2001.

[4] D.H. Klatt, "Review of text-to-speech conversion for english," pp. 737 – 793, 1987.

[5] B. Yegnanarayana, S. Rajendran, V.R. Ramachandran, and A.S. Madhukumar, "Significance of knowledge sources for a text-to-speech system for Indian languages," Sadhana, pp. 147 – 169, 1994.

[6] Alan W Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," 1997, pp. 601 – 604.

[7] R. Donovan and P. Woodland, "Improvements in an HMM-based speech synthesiser," in Eurospeech95, Madrid, Spain, 1995, vol. 1, pp. 573 – 576.

[8] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in Joint Meeting of ASA, EAA, and DAGA, Berlin, Germany, 1999, pp. 18 – 24.

[9] S.P. Kishore and Alan W Black, "Unit size in unit selection speech synthesis," in Proceedings of Eurospeech 2003, Geneva, Switerzland, 2003.

[10] Bharati, Akshar, Sushma Bendre, and Rajeev Sangal, "Some observations on corpora of some Indian languages," in Knowledge-Based Computer Systems,Tata McGraw-Hill, 1998.