Unit Size in Unit Selection Speech Synthesis

S P Kishore and Alan W Black

Language Technologies Research Center
International Institute of Information Technology, Hyderabad
and *ISRI, Carnegie Mellon Univesity*
kishore@iiit.net
Language Technologies Institute, Carnegie Mellon University
awb@cs.cmu.edu

Abstract

In this paper, we address the issue of choice of unit size in unit selection speech synthesis. We discuss the development of a Hindi speech synthesizer and our experiments with different choices of units: syllable, diphone, phone and half phone. Perceptual tests conducted to evaluate the quality of the synthesizers with different unit size indicate that the syllable synthesizer performs better than the phone, diphone and half phone synthesizers, and the half phone synthesizer performs better than diphone and phone synthesizers.

1. Background

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 details of the development of a Hindi speech synthesizer using unit selection techniques and in particular address the issue of choice of unit size in unit selection synthesis.

2. Synthesis Framework

This work is done within the FestVox voice building framework [1], which offers general tools for building unit selection synthesizers in new languages. The unit selection paradigm is a cluster based technique where units of the same type (phones, diphones, syllables or whatever) are clustered based on their acoustic differences [2]. The clusters are then indexed based on high level features such as phonetic and prosodic context. Voices generated by this system may be run in the Festival Speech Synthesis System [3].

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 will exploit in this paper.

3. Hindi Synthesis

The basic units of the writing system in Indian languages are *characters* which are an orthographic representation of speech sounds. A character in Indian language scripts is close to a syllable and can be typically of the following form: C, V, CV, VC, CCV 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. In Hindi, there are five vowels, five long vowels, two diphthongs, four semivowels, and 31 consonants. There are a few more vowels and consonants existing in Hindi, but we did not consider them as they are rarely used in the current times.

3.1. Letter to Sound Rules

The scripts of Indian languages are phonetic in nature. There is more or less one to one correspondence between what is written and what is spoken. However, in Hindi the *inherent vowel* (short /a/) associated with a consonant is not pronounced depending on the context. This is referred to as Inherent Vowel Suppression (IVS) or schwa deletion. For example, the word *kamala* [lotus] is mapped to a sequence of consonant and vowel sounds */k/ /a/ /m/ /a/ /l/*, ignoring the vowel associated with /l/.

A set of heuristic rules to detect IVS of a consonant character are noted below. These rules have been derived by observing a few hundred Hindi words, and the rule set may not be a complete description of the phenomenon.

1 No two successive characters undergo IVS.

2 Characters present in the first position of a word, never undergo IVS. IVS occurs only to the characters present in middle and final positions.

3 For characters in final position, the inherent vowel (/a/) is always suppressed.

4 For characters in word middle position, IVS occurs if the next character in the word is not the last character or the next character has a vowel other than /a/.

3.2. Syllabification Rules

In Hindi, words could be composed of basic characters (example *samay* [time]), as well as complex clusters of C*VC* (example *san'sthaa* [organization]). For the latter cases, there is need to come up with rules to break the word into syllables. We derived certain simplistic rules for syllabification i.e. rules for grouping clusters of C*VC* based on heuristic analysis of several words in Telugu and Hindi languages.

- When nasals such as */n'/*, half pronounced /m/ or /n/ sound, (refer to Figure 1 where Hindi characters are represented in ITRANS-3, a transliteration scheme) succeed a vowel immediately, they would be treated as a part of the vowel and also the same syllable. For example, */n'/* in *san'sthaa* will be a part of syllable containing */sa/*.
- When there are three or more consonants between two consecutive vowels, the first consonant would be a part of the coda of the previous syllable while the remaining consonants would be onset of the next syllable. Applying these rules to *san'skrit* [sanskrit], the obtained syllable sequence would be */san's/ /krit/*.
- When there are exactly two consonants between two vowels, the first consonant would be part of coda of previous syllable and the second would be onset of the next syllable. For example, *dharti* [earth] would be split as */dhar/ /ti/*. Exceptions for this rule are the following cases.
  - When the second consonant is a member of the set { */r/ /s/ /sh/ /shh/* }, both the consonants would be a part of onset of the next syllable. For example, *yaatra* [tour] would be split as */yaa/ /tra/*.

3.3. Hindi Speech Database

To build a unit selection speech synthesizer in Hindi our first task was to define the phoneme set; then construct a set of prompts that best covers the language. We generated a prompt-list covering most of the high frequency syllables in Hindi. 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 [4]. 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. These selected sentences were recorded by a female speaker, and a speech corpus of about 96 minutes was generated. The recording was done in a quiet room with a noise canceling microphone using the recording facilities of a typical multimedia computer system. The speech database was labeled at the phone level and the label boundaries were hand-corrected.

The duration of the speech data used in this study is about 90 minutes, and it has 620 utterances with 2344 syllables (22960 realizations), 1414 diphones (51282 realizations) and 48 phones (51282 realizations).

4. Unit Size

Earlier work on Indian languages [5] and preliminary experiments with this Hindi database [6] suggested that a syllable based approach to synthesis could lead to more reliable quality. 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. FestVox's default method uses a phone based technique. However because FestVox supports a method of optimal coupling [9], the join points may be moved within the preceding unit, thus with phone-sized units, something more like diphones are actually selected.

Larger units are also possible, from demi-syllables to syllables and larger. [10] tie the phones to words for domain synthesis, although this is not the same as having word-sized units it is in that direction. The choice of unit size is an optimization problem, the larger the units the lesser are the discontinuities in synthesis but it is harder to ensure general coverage. Smaller units make it easier to cover the space of acoustic units but at the cost of more joins.

The choice of unit size 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. As Hindi has a much more regular syllable structure than English we wanted to experiment to find the optimal sized unit for Hindi synthesis.

5. Experiment

In order to investigate the optimal unit size we built synthesizers under four different conditions: **syllable, diphone, phone and half phone**.

The phone synthesizer, the base case, was built with the phone set, letter to sound rules and syllabification rules defined for Indian language.

To build the diphone synthesizer we tagged each phone with its preceding phone, thus units were still actually one phone in length but they are sub-typed based on their previous phone.

For the syllable based synthesizer, we treated the 2344 distinct syllables in the database as "phones" and listed them in our phoneset. These syllable-sized phones were assigned phonetic features based on their combined consonant and vowel part, with the consonant in onset given more preference over the consonant in coda. Thus the units in the inventory became full syllables rather than traditional phonemes. The lexicon parser was appropriately modified to generate these syllable-based phones rather than traditional phone names.

In implementing half phone synthesizer, each vowel was represented by two half phones, while the consonants were full phones. Two phone symbols were defined for each vowel in the phoneset, for example vowel /a/ was represented by /a_1/ and /a_2/. Labels at half phone level were derived by equally dividing the vowel segment into two half phones. The lexicon parser was also modified accordingly, to generate appropriate phone strings.

For perceptual evaluation of these synthesizers, we selected a set of 24 sentences from a Hindi news bulletin. The content of this bulletin was mostly about the political affairs of the world in the middle of March 2003. The syllables and diphones present in these 24 sentences were covered in the corresponding synthesizers. These sentences were synthesized by phone, diphone, syllable and half phone synthesizers and were subjected to the perceptual test of native Hindi speakers. The people who participated in these perceptual tests were working persons and graduate students and none of them had any experience in speech synthesis. 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 are shown in Tables 1-6, with a summary in Table 7. Each row in these tables indicates the evaluation results of a native speaker. An entry such as 8 6 10 in the first row of Table 1 indicates that the listener rated 8 utterances in favor of syllable, 6 utterances in favor of phone and 10 utterances as equally good or bad. The last row in each of these tables summarizes the results present in the corresponding tables.

Table 1: AB Test: Syllable Vs Phone

Test No.	Listener Preference		
	Syllable	Phone	No Preference
1.	**8**	6	10
2.	**5**	4	15
3.	**9**	-	15
4.	**9**	**9**	6
5.	**9**	7	8
	40	26	54

Table 2: AB Test: Syllable Vs Halfphone

Test No.	Listener Preference		
	Syllable	Halfphone	No Preference
1.	2	**4**	18
2.	**9**	3	12
3.	**10**	6	8
4.	**4**	-	20
5.	3	**4**	17
	28	17	75

Table 3: AB Test: Syllable Vs Diphone

Test No.	Listener Preference		
	Syllable	Diphone	No Preference
1.	**13**	8	3
2.	**7**	2	15
3.	**4**	**4**	16
4.	**8**	5	11
5.	**11**	6	7
6.	**13**	5	6
7.	**10**	8	6
8.	**11**	8	5
9.	**11**	6	7
10.	**14**	1	9
11.	**12**	**12**	-
	114	65	85

Table 4: AB Test: Diphone Vs Phone

Test No.	Listener Preference		
	Diphone	Phone	No Preference
1.	7	**8**	9
2.	**4**	**4**	16
3.	3	**4**	17
4.	**8**	6	10
5.	**13**	6	5
	35	28	57

Table 5: AB Test: Diphone Vs Halfphone

Test No.	Listener Preference		
	Diphone	Halfphone	No Preference
1.	**6**	5	13
2.	5	**7**	12
3.	**11**	5	8
4.	1	**5**	18
5.	-	**7**	17
	23	**29**	68

Table 6: AB Test: Phone Vs Halfphone

Test No.	Listener Preference		
	Phone	Halfphone	No Preference
1.	**5**	3	16
2.	5	**6**	13
3.	7	**8**	9
4.	**2**	-	22
5.	1	**5**	18
	20	**22**	78

6. Discussion

From Table 1-2 and Table 4-6, we observe that the notion of equality or no preference (referred to as "=" in Table 7) occupies first position. This indicates that the listeners perceived the speech synthesized by different synthesizers as either equally good or equally bad. However, if we look at the choice of unit size, the results shown in Table 1-3, indicate that speech synthesized with syllable sized units is preferred over the speech synthesized with other choices of unit size. The results of Table 4 indicate that diphone performs better than phone while the results of Table 5-6 indicate that half phone performs better than phone and diphone. Table 7 summarizes the results of AB-test in terms of percentages (number of times a unit is favored / total utterances * 100).

It should be noted that the syllables as well as diphones in test sentences were covered by the speech database, though this will not be true in general. However, the prompt-list used for building the speech database was derived from a text corpus which covered a wide range of subjects including literature, dialog, novels, philosophy and short stories, while the 24 sentences used for testing were from a news bulletin describing the global events in the middle of March 2003. The context in which test sentences were derived was not related to the prompt-list used to generate the speech database.

Larger units such as syllables might assimilate prosodic and acoustic information better and have less discontinuities in synthesized speech, resulting in better performance over other units. Units such as diphones have performed better than phone as they preserve the phone-to-phone transitions. However the small differences are due to the joinings moved within the previous units even in the case of phones as a method of optimal coupling.

The smaller units such as half phones involve more number of joinings and could lead to the impression that it produces more discontinuous speech. The results of Table 5-6 indicate that the half phone synthesizers perform better than diphone and phone synthesizers. To join two consecutive units we use optimal coupling [9]. The better performance of half phones could be attributed to its vast coverage and hence the chance of finding an optimal sub segment with required acoustic features would be more.

The choice of larger unit such as syllable seems to be ap-

Table 7: Summary of AB Test (scores are represented in %)

Rank	Syl vs Diph	Syl vs Ph	Syl vs Halfph	Diph vs Ph	Diph vs Halfph	Ph vs HalfPh
I	syl 43%	= 45%	= 63 %	= 47%	= 57%	= 65%
II	= 32%	syl 33%	syl 23 %	diph 29%	halfph 24%	halfph 18%
III	diph 24%	ph 21%	halfph 14%	ph 23%	diph 19%	ph 17%
Sum.	**syl**	**syl**	**syl**	**diph**	**halfph**	**halfph**

propriate choice for syllabic languages such as Hindi and seems to be a better representation for the Indian language scripts. But larger the unit the lesser would be the coverage, which has to be dealt with. Given an arbitrary text, we found that the syllable coverage by this Hindi database was around 84% and the diphone coverage was 88%. With a more careful selection of the prompt-list we believe that it is possible to cover most of the frequently occurring syllables in Hindi, but some back-off method is required too.

7. Conclusions

In this paper, we addressed the issue of choice of unit size in unit selection synthesis. We built the Hindi synthesizer for different choices of unit size: syllable, diphone, phone and half phone. We conducted perceptual tests to evaluate each of these synthesizers in comparison with other. 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 languages such as Hindi. It was also observed that the half phone synthesizer performed better than diphone and phone synthesizers, though not as well as syllable.

8. Acknowledgments

The part of this work carried out at CMU was funded in part by the U.S. National Science Foundation grant "ITR/CIS Evaluation and Personalization of Synthetic Voices" The opinions expressed in this paper do not necessarily reflect those of NSF.

Our special thanks to Chaitanya Krishna of IIIT Hyderabad for conducting the perceptual tests in a short time. We also thank all the people of LTRC and graduate students of IIIT Hyderabad for their participation in the perceptual tests.

9. References

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

[2] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Eurospeech97*, Rhodes, Greece, 1997, vol. 2, pp. 601–604.

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

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

[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] S.P. Kishore, Rohit Kumar, and Rajeev Sangal, "A data-driven synthesis approach for Indian languages using syllable as basic unit," in *Proceedings of International Conference on Natural Language Processing (ICON)*, 2002.

[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] A. Conkie and I. Isard, "Optimal coupling of diphones," in *Proc. ESCA Workshop on Speech Synthesis*, Mohonk, NY., 1994, pp. 119–122.

[10] A. Black and K. Lenzo, "Limited domain synthesis," in *ICSLP2000*, Beijing, China., 2000, vol. II, pp. 411–414.

Figure 1: Characters of Hindi in ITRANS-3 Transliteration Scheme