SPEECH SYNTHESIS USING APPROXIMATE MATCHING OF SYLLABLES

E.Veera Raghavendra[†], *B. Yegnanarayana*[†], *Kishore Prahallad*[†‡]

†International Institute of Information Technology - Hyderabad, India.
‡Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
{raghavendra,yegna}@iiit.ac.in, skishore@cs.cmu.edu

ABSTRACT

In this paper we propose a technique for a syllable based speech synthesis system. While syllable based synthesizers produce better sounding speech than diphone and phone, the coverage of all syllables is a non-trivial issue. We address the issue of coverage of syllables through approximating the syllable when the required syllable is not found. To verify our hypothesis, we conducted perceptual studies on manually modified sentences and found that our assumption is valid. Similar approaches have been used in speech synthesis and it shows that such approximation produces intelligible and better quality speech than diphone units.

Index Terms— Speech synthesis, unit size, syllable, approximate matching.

1. INTRODUCTION

Concatenative speech synthesis is based on the concatenation of speech segments. Generally, concatenative synthesis produces the most natural-sounding synthesized speech [1]. This synthesis method uses prerecorded speech units which preserve coarticulation and prosody of the language [2]. The quality of the synthetic speech is thus a direct function of the available units, making unit selection very important. For good quality synthesis, all the units of the language should be present. Moreover, the units should also be generic so that they can be used for unrestricted synthesis.

In the context of Indian languages, the basic units of writing system are characters which are an orthographic representation of speech sounds. A character in Indian language scripts is close to the syllable and can be typically of the following form: V, CV, VC, CCV, CCCV, and CCVC, where C is consonant and V is Vowel. A syllable can be represented as C*VC*, containing at least one vowel and zero, one or more consonants. All Indian language scripts have a common phonetic base, and a universal phoneset consists of about 35 consonants and about 15 vowels. Theoretically possible syllable combinations in an Indian language with V, CV, CCV, CVC, CCVC representation are 680415. For Indian Languages, the syllable units are a much better choice than units like phone, diphone, and half-phone [3]. The reason for using the syllable as a unit is larger the unit lesser the concatenations and reduces the co-articulation effects during the synthesis. Text-to-speech synthesis based on syllables seems to be a good possibility to enhance the quality of synthesized speech with comparison to diphone based synthesizers [3]. The syllable-based approach has to face the problem with a relatively large inventory of the syllables and we cannot cover all the syllables of a language in a speech database.

In order to address the coverage of the syllables, we have hypothesized that approximate matching of the syllable could be used for text-to-speech synthesis [4]. In this paper, we present perceptual experiments and results based on which such an hypothesis (approximate matching of the syllable could be used for speech synthesis) has been proposed.

The rest of the paper organized as follows. Section 2 discusses the approximate matching on manually prepared utterances. Section 3 describes the usage of approximate matching in speech synthesis and Section 4 gives the subjective and objective evaluations for approximate matching in comparison with diphone synthesis.

2. APPROXIMATE MATCHING OF A SYLLABLE

Our perceptual mechanism; ears, eyes, and so forth, are the tools that we use to coordinate our dealings with the world. These mechanisms respond to physical stimuli caused by objects and events in our environments. Our hypothesis is that even though there are some pronunciation mistakes in an utterance, the human ear can understand the sentence without any difficulty. In this work we use the transliteration scheme referred to as IT3 developed by IISc Bangalore and Carnegie Mellon University to represent the Indian language scripts [5]. For example the words in Telugu *praveishin:china, ikkad:a* if pronounced as *praveishin:jina, ikad:a* human listeners do not tend to identify the change given in a context. The modifications are *j* and *k* instead of *ch* and *kk*. In the first example unvoiced unaspirated phone is substituted with voiced unaspirated phone. In the second example left most consonant of the syllable */kka/* has been deleted.

To test how human hearing mechanism works when phones are substituted and deleted, a perceptual study was conducted. We have prepared two sets of 44 utterances. First set (Set A) contains original utterances which were collected

from news bulletin. The second set (Set B) is prepared carefully substituting or deleting with one of the phone in each of the sentence. A native speaker was asked to record both the sets. Native speaker was instructed to record the second set carefully, so that the intended pronunciation mistakes are preserved while recording. Later these samples were played to 15 native Telugu speakers for obtaining mean opinion scores(MOS), i.e, score between 1 (worst) to 5 (best) and ABTest, where original and modified utterances are played in random order and the listener was asked to decide any difference was found between the utterances. MOS and ABTest were combined in the same test.

In each table from Table 1-6, first column is the sentence number which was used for perceptual study, second column specifies the average MOS score for original sentence where as third column specifies the average MOS score for modified sentence. Fourth column gives how many subjects out of 15 found that there is no difference between two utterances and the fifth column gives the number of subjects found that there is difference between two utterances. Sixth column gives what is the substitution in the corresponding sentence.

2.1. Phone Substitution

The phone substitution must be handled very carefully. First, we need to study what kind of phones can be replaced. In articulatory phonetics, a *consonant* is a speech sound that is articulated with complete or partial closure of the upper *vocal tract*. Consonants are classified in terms of Place of Articulation (POA) and Manner of Articulation (MOA) [6]. Consonants that have the same place of articulation, such as */k/, /kh/, /g/, /gh/, /ng /* in Telugu, are said to be homogranic.

Table 1. *Perceptual scores for substitution of phones with same POA but different MOA*

	MOS		AB Test			
Sent. No	Set A	Set B	No Diff	Diff	Map	
1	4.07	3.27	6	9	k-kh	UV
2	4.27	3.33	5	10	ch-chh	UnAsp
3	4.37	3.47	7	8	t:-t:h	to
4	3.9	3.25	7	8	t-th	UV
5	4.51	4	**9**	6	p-ph	Asp
6	4.4	4.4	**14**	1	k-g	UV
7	4.47	4.27	**13**	2	ch-j	UnAsp
8	4.2	4.29	**10**	5	t:-d:	to
9	4.29	4.29	**12**	3	t-d	V
10	4.44	3.68	**9**	6	p-b	UnAsp

Table 1 shows perceptual results obtained by substitution of phones with same POA but different MOA. The MOS scores and preference tests (ABTest) in Table 1 show that there is no degradation in intelligibility or change in preference when an unvoiced unaspirated phone is substituted with a voiced unaspirated phone. However, when an unvoiced unaspirated phone is replaced with an unvoiced aspirated phone, the MOS scores and preference scores indicate that the native speakers of Telugu are sensitive to this change. The result can also be attributed to the property that Indian languages have aspirated phones and the native speakers of Indian languages are good in distinguishing aspirated versus unaspirated phones.

Table 2. *Perceptual scores for substitution of phones with different POA and different MOA*

	MOS		AB Test			
Sent. No	Set A	Set B	No Diff	Diff	Map	
11	4.15	2.73	3	12	k-chh	UV
12	4.29	2.67	5	10	ch-th	UnAsp
13	4.3	3.72	**11**	4	t:-chh	to
14	4.55	4.2	**11**	4	t-chh	UV Asp
15	4.47	2.71	3	12	k-j	UV
16	4.21	2.75	6	9	ch-d	UnAsp
17	4.41	2.92	6	9	t:-j	to
18	4.27	2.85	3	12	t-g	V UnAsp

Table 3. *Perceptual scores for substituting one semivowel with other semivowel*

	MOS		AB Test		
Sent. No	Set A	Set B	No Diff	Diff	Map
19	4.49	2.69	3	12	r-l
20	4.4	4.27	**13**	2	l-r
21	4.28	3.72	9	6	l-l:
22	4.35	4.28	**10**	5	l:-l

Table 2 shows perceptual results obtained by substitution of phones with different POA and different MOA. The MOS scores and preference test in Table 2 show that there is significant change in intelligibility or change in preference when an unvoiced unaspirated phone is substituted with an unvoiced aspirated or voiced unaspirated phone from different POA and different MOA. From Table 1 and 2, it signifies that a phone could be substituted with another phone from same POA (different MOA) but not with a phone from different POA (different MOA). Exceptions could be observed in Table 2 for two phones */t:/ and /t/* which could be replaced with phone */chh/*.

The group of phones */y/, /r/, /l/, /l:/, /v/* are called semivowels because of their vowel-like nature. Table 3 shows the substitution between one semivowel to another. The MOS scores and preference tests in Table 3 show that */l/* can be substituted in place of */l:/* and vice-a-versa. Where as in the case of */r/* to */l/*, the results indicate that replacing */l/* with */r/* is acceptable but the reverse (*/r/* with */l/*) is not acceptable.

Table 4 shows the substitution done for fricatives. The perceptual scores in Table 4 indicate */sh/*, */shh/* can be replaced with */shh/*, */s/*, however */s/* cannot be replaced with */shh/* or */sh/*. The phone */sh/, /shh/, /s/* are arranged in the descending sonority levels [7]. The fact that */sh/* or */shh/* can be replaced with */s/* indicates that a less sonority phone can be used as a substitute for a higher sonorant phone. However, vice-a-versa may not lead to good results.

Table 4. *Perceptual scores for substituting one fricative with other fricative*

Sent. No	MOS		AB Test		Map
	Set A	Set B	No Diff	Diff	
23	4.13	4.21	**11**	4	sh-shh
24	4.21	4.35	**14**	1	shh-s
25	4.76	4.03	**13**	2	sh-s
26	4.2	2.89	3	12	s-shh
27	4.43	3.13	4	11	s-sh
28	4.17	4.05	**11**	4	shh-sh

Table 5. *Perceptual scores for substituting one nasal with other nasal*

Sent. No	MOS		AB Test		Map
	Set A	Set B	No Diff	Diff	
29	4.09	4.02	**10**	5	n-nd∼
30	4.45	3.23	**8**	7	n-m
31	4.41	3.43	6	9	nd∼-m
32	4.37	3.38	7	8	m-n

A consonant produced through the nose with the mouth closed is called nasal consonant. Phones */n/, /nd∼/, /m/* belong to nasal consonants. These phones are distinguished by the place at which a total constriction is made. For */m/* the constriction is at the lips, for */n/* the constriction is just backside of the teeth, and for /nd∼/ the constriction is just ahead of the velum. Table 5 shows the substitution done for nasals. Generally, speech signal of /m/ and /n/ looks very similar [8]. Preference tests in Table 5 show that the intelligibility of */m/* and */n/* are interchangeable. The substitution of */n/* with */nd∼/* shows that there is no degradation where as from */nd∼/* to */m/*, the degradation can easily be noted.

2.2. Phone Deletion

Though we apply approximate matching using phone substitution, we may not solve the syllable coverage problem. Because, after substitution of the phone we cannot ensure that the syllable exists in the lexicon. The use of approximate matching using phone deletion could be used to ensure better syllable coverage.

Each syllable may contain consonants before and after the vowel. The duration of the vowel is comparatively very high when compared consonants. Missing of a consonant in the syllable (specially the consonants not immediately preceding/succeeding the vowel) might not lead to serious degradation in intelligibility of an utterance. We wanted to test this hypothesis in designing the approximating nearest syllable using phone deletion. For example, the word */ikkad:a/* if pronounced as */ikad:a/*, and we cannot identify the consonant */k/* missing. Using this hypothesis, we have taken 12 utterances randomly and modified some syllables by removing some consonants of the syllable which were mentioned in the beginning of this section.

12 utterances are used to verify the difference of opinion between original and modified sentences. Perceptual scores in Table 6 shows that majority of the modifications can be ignored except in three cases (*kshha-shha*, *t:t:a-t:a* and *shht:i-t:i*). The difference in the exceptional cases are very low as compared with phone substitution scores. It implies that consonant deletion is acceptable in the syllable units. Syllable is a larger unit and hence when one consonant is removed from the syllable and joined with the consequent units, listener gets the continues flow and ignores the missing phone.

Table 6. *Perceptual scores for Removing consonants of the Syllables*

Sent. No	MOS		AB Test		Map
	Set A	Set B	No Diff	Diff	
33	4.4	3.47	**8**	7	nnai-nai
34	4.32	3.26	6	9	kshha-shha
35	4.35	4.23	**11**	4	rd:u-d:u
36	4.08	4	**10**	5	rnd∼a-nd∼a
37	4.15	3.21	5	10	t:t:a-t:a
38	4.01	3.95	**9**	6	t:t:u-t:u
39	4.06	4	**10**	5	rdyaa-dyaa
40	4.35	4.15	**12**	3	rs-s
41	4.49	4.57	**14**	1	kka-ka
42	4.35	3.57	6	9	shht:i-t:i
43	4.46	4.34	**13**	2	chchi-chi
44	4.39	4.16	**11**	4	vru-ru

3. SPEECH SYNTHESIS USING APPROXIMATE MATCHING OF SYLLABLE

So far we discussed approximate matching with manually recorded utterances. The usefulness of appropriate matching in a text-to-speech system is described below.

This work is done within the FestVox voice building framework [9] and the speech database used from [10]. In implementing a syllable synthesizer, we treated 1790, 2757 and 1892 distinct syllables for Telugu, Hindi and Tamil respectively 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 phone. The lexicon parser was appropriately modified to generate these syllable-based phones rather than traditional phone names.

During the runtime the given text is broken into syllables and lexicon is examined for each syllable. Whenever the required syllable is not found in the lexicon the syllable is modified using the phone substitution which is discussed in Section 2.1, and the modified syllable is available in the lexicon. Otherwise approximate matching using phone deletion is applied. Algorithm given in [4] gives the deletion of consonants in the syllable.

4. EVALUATION

To evaluate the syllable based synthesizer which employs approximate matching, we have conduced subjective and objective evaluations in comparison with a diphone based synthesizer. In order to evaluate, 10 utterances were extracted from Telugu, Hindi and Tamil test database. For each utterance, the synthesized speech signal obtained by both methods were randomly presented to 10 listeners. Some of the syllables used in these sentences are approximated using nearest syllable as explained Section 3. Each listener is subjected to MOS i.e score between 1 (worst) to 5 (best) and AB-Test i.e the same sentence synthesized by two different synthesizers is played in random order and the listener is asked to decide which one sounded better. They also had the choice of giving the decision of equality. As a part of objective evaluations Mel-Cepstral Distortion (MCD) [11] are calculated between original and synthesized wave files. Lower the MCD value the better in the speech synthesis.

Table 7. *MOS, MCD and AB-Test scores for Syllable (Syl) and Diphone (DP) voice utterances.*

Test	MOS		MCD		ABTest		
	Syl	DP	Syl	DP	Syl	DP	Similar
Telugu	3.108	2.891	5.873	6.031	50/100	29/100	21/100
Hindi	3.085	2.772	5.226	5.462	49/100	32/100	19/100
Tamil	3.135	2.705	6.726	7.088	50/100	24/100	26/100

The results shown in Table 7 indicate that the syllable based synthesizer employing approximate matching performs better than diphone based synthesizer for Telugu, Hindi and Tamil and also show that our hypothesis of approximate matching is valid. The MOS scores in Table 7 show that approximate matching does not degrade the intelligibility of synthesis in comparison with diphone synthesis. This indicates that approximate matching is a useful technique for developing the syllable based synthesizers for Indian languages without worrying about back-off synthesizers using lower level units.

5. CONCLUSION

In this paper, we have discussed the approximate matching for the syllable and proved that the hypothesis for such approximation is valid. We have built Telugu, Hindi and Tamil synthesizers using approximated syllables. We conducted subjective and objective evaluations to evaluate these synthesizers in comparison with diphone synthesis. The evaluation on the syllable based synthesizer indicate that the approximate matching of syllables is a useful and viable technique to build syllable based synthesizers for Indian languages without requiring any back off synthesizers.

6. REFERENCES

[1] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," *in Proceedings of Eurospeech 97*, vol. 2, pp. 601–604, September 1997.

[2] T. Dutoit, "An introduction to text-to-speech synthesis," *Kluwer Academic Publishers*, 1997.

[3] S. P. Kishore and A. Black, "Unit size in unit selection speech synthesis," *in Proceedings of Eurospeech*, pp. 1317–1320, September 2003.

[4] E. V. Raghavendra, B. Yegnanarayana, A. Black, and K. Prahallad, "Building sleek synthesizers for multilingual screen reader," *in Proceedings of Interspeech*, pp. pp. 1865–1868, September 2008.

[5] P. Lavanya, P. Kishore, and G. Madhavi, "A simple approach for building transliteration editors for indian languages," *Journal of Zhejiang University Science*, vol. 6A, no.11, pp. 1354–1361, October 2005.

[6] Wikipedia, "Place of articulation," *http://en.wikipedia.org/wiki/Place_of_articulation*.

[7] Wikipedia, "Sonority hierarchy," *http://en.wikipedia.org/wiki/Sonority_hierarchy*.

[8] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition.," *Prentice Hall*, p. 30, 1993.

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

[10] E. V. Raghavendra, D. Srinivas, B. Yegnanarayana, A. Black, and K. Prahallad, "Global syllable set for speech synthesis in indian languages," *accepted at IEEE workshop on Spoken Language Technologies.*, December 2008.

[11] T. Toda, A. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech," *in Proceedings of 5th ISCA Speech Synthesis Workshop*, pp. 31–36, June 2004.