SIGNIFICANCE OF VOWEL EPENTHESIS IN TELUGU TEXT-TO-SPEECH SYNTHESIS

Vijayaditya Peddinti, Kishore Prahallad

International Institute of Information Technology, Hyderabad, AP, India.
vijayaditya.p@research.iiit.ac.in, kishore@iiit.ac.in

ABSTRACT

Unit selection synthesis inventories have coverage issues, which lead to missing syllable or diphone units. In the conventional back-off strategy of substituting the missing unit with approximate unit(s), the rules for approximate matching are hard to derive. In this paper we propose a back-off strategy for Telugu TTS systems emulating native speaker intuition. It uses reduced vowel insertion in complex consonant clusters to replace missing units. The inserted vowel identity is determined using a rule-set adapted from L2 (second language) acquisition research in Telugu, reducing the effort required in preparing the rule-set. Subjective evaluations show that the proposed back-off method performs better than the conventional methods.

Index Terms— speech synthesis, back-off, unit-selection

1. INTRODUCTION

Text-to-speech (TTS) synthesis with unit selection is widely used to produce natural sounding speech, with minimal signal processing [1]. TTS systems in Indian languages use syllable sized units for concatenative synthesis, as syllables in these languages have a regular structure [2]. The quality of speech synthesis depends on the syllable coverage of the audio database. Hence the audio database is recorded using an optimally selected text to maximize syllable coverage [3]. However missing syllable units are a frequent occurrence, especially for unrestricted text [4]. Back-off methods are used to tackle these missing units. The back-off methods either substitute the missing syllable with other syllables [5] in the database or synthesize it using smaller sub-word units (SSUs) like diphones. However even missing diphones are a common problem [1]. In these cases, diphone substitution or half-phone back-off is used to synthesize the missing diphones [6].

The design of both these back-off methods is not trivial, as each of them presents it's own set of problems. Phonetic segments within syllables are found to have relatively high coarticulation, compared to those across syllable. The SSU back-off method which forms the missing syllables from smaller sub-word units has to ensure that the units selected have the right coarticulative influences, some of which span across several phones, e.g., influence of lip protrusion of the vowel is seen across many preceding consonants [7]. The half-phone back-off method splices half-phone units, at unstable regions like phone boundaries which is not desirable [8]. In addition to this, the audio databases used in TTS systems are segmented automatically into constituent phone units, from which the SSU (diphone and half-phone) boundaries are derived. The joining of half-phones at these automatically segmented phone boundaries is non-trivial and may not lead to optimal output.

The substitution methods avoid the aforementioned problems, by using other syllable or diphone units in the inventory to replace the missing units. However it takes considerable effort to design the substitution rules, requiring detailed perceptual studies when designing for new languages [5]. Even after these detailed studies, minimal pairs i.e., words which only differ in one phonological element and have distinct meanings, pose a problem when using substitution as a back-off. Hence the application of substitution rules should be done taking care that the meaning of the word does not change. Further diphone substitutions necessitate a change in the neighbouring units, thus creating a complex search problem [1].

In this paper we present a rule-based back-off method motivated from a perceptual and speech production phenomenon, known as vowel epenthesis, to deal with the missing syllable and diphone units in Telugu language. The proposed back-off method emulates native speaker intuition in synthesis of the missing units. Hence it is found to be perceptually acceptable to the native speakers of the language. The rule-base is borrowed from the L2 (second language) acquisition research in Telugu. Further, the proposed back-off strategy is robust to the segmentation errors in automatically segmented databases, helping us ensure the quality of synthesis output.

Telugu, the language of interest in this paper, is one of the official Indian languages. It has phoneset of 15 vowels and 35 consonants. The database used in our experiments covers only 3,000 of more than 10,000 possible syllables in the language, using 12 hours of audio recordings. Thus providing us ample opportunity to study the back-off strategies. The experiments were done using the “clunits” and “Multisyn” unit selection engines in the Festival framework [1].

The paper is organized as follows. Section 2 details the vowel epenthesis phenomenon in Telugu. Section 3 describes the issues in utilizing epenthesis as a potent back-off strategy in Telugu TTS systems, and proposes resolutions for these issues. Conclusions are presented in Section 4.

2. VOWEL EPENTHESIS IN TELUGU

Phonotactics of a language is the permissible combinations of phonemes that can co-occur in that language. Thus phonotactics helps us reduce the number of syllables or diphone units that need to be covered in the unit inventory of a language. However it is difficult to get a complete coverage of even this reduced set of syllables or diphones. Usually about 80% syllable coverage is achieved, even in languages with simple syllabic structure and further increase in coverage is not proportional to the increase in the size of the inventory. In the case of diphones, though it can be ensured that all phonotactically possible diphones occur in the database using carefully constructed sentences, many such diphones would not occur more than once in the database [4].

To further compound the problems above, influx of new words into a language, due to borrowings from other languages is a fairly frequent phenomenon [9]. These borrowed words do not necessarily conform to the phonotactics of the synthesis language. Thus, these borrowed words result in missing syllables (or diphones) due to the presence of new consonant clusters disallowed by the phonotactics of the synthesis language. Native speakers of Telugu break such consonant clusters, through vowel insertion, to conform to the phonotactics of Telugu [10]. This phenomenon is known as vowel epenthesis. e.g., The English word "*bulb*" which is pronounced by Telugu speakers trained in English as [balb][1], is pronounced as [bal**ubu**] by native Telugu speakers untrained in English. As the consonant cluster "lb" is new to native Telugu speakers, they perform an insertion of the vowel "u" to break it. Another "u" is also inserted after the word final stop consonant b, as words in Telugu do not end with stop consonants. These inserted vowels are called epenthetic vowels. It is this property of epenthesis that we want to exploit as a back-off strategy in Telugu TTS systems.

3. INCORPORATION OF VOWEL EPENTHESIS IN TELUGU TTS SYSTEMS

The TTS system can emulate the epenthesis phenomenon to break the consonant clusters and since the new syllables conform to the phonotactics of Telugu, they can be found in the unit inventory with a greater probability. It is safe to assume that it would be acceptable for native Telugu speakers to hear a TTS system, which performs epenthesis (insertion) of the vowel "u" when faced with the uni-syllable word [balb], missing in the database, to produce the syllables [ba],[l**u**] and [b**u**]. Thus providing a new back-off strategy for the missing unit problem.

The resolution of the following issues is necessary to successfully incorporate vowel epenthesis as a back-off strategy for Telugu TTS systems :

1. How to determine the identity of the epenthetic vowel?
2. How to use epenthesis to simplify frequent clusters?
3. Is multiple epenthesis within a complex consonant cluster perceptually acceptable?
4. Is the performance of vowel epenthesis based back-off strategy better than other back-off strategies?

The following sub-sections detail these issues and the proposed resolutions for these issues.

3.1. Identity of the epenthetic vowel

If a consonant cluster in a syllable violates the phonotactic constraints of Telugu, it is broken using epenthesis. The identity of this epenthetic vowel is determined by the vowel harmony rule [10]. According to the vowel harmony rule, if the epenthetic vowel is inserted in a word medial consonant cluster, it's identity is dependent on the identity of the vowel following the cluster. If the epenthetic vowel is inserted in a word final consonant cluster, it's identity is determined based on the word final consonant. These rules are tabulated in Tables 1 and 2.

Table 1. Identity of word medial epenthetic vowel in Telugu

Following vowel	Epenthetic vowel
a,a:	a
i,i:,e,e:	i
u,u:,o,o:	u

Table 2. Identity of word final epenthetic vowel in Telugu

Word final consonant	Epenthetic vowel
non-palatal consonant	u
palatal consonant	i

In the case of words with inflexional suffixes, the epenthetic vowels are determined treating the root word and suffix as isolated words. Similarly in the case of compound words the epenthetic vowels are determined separately for each word.

3.2. A reduced vowel for Epenthesis

The TTS system can emulate vowel epenthesis to tackle missing syllables. e.g., The syllable [kloo] when missing, can be replaced with syllables [ku] and [loo] after epenthesis of the vowel [u] in the consonant cluster [kl]. However preliminary experiments showed that tackling all missing syllables using this strategy was not suitable. Some consonant clusters though frequent in current usage (e.g., [ʈɽ]), are not valid according to the phonotactics of Telugu. Breaking such clusters

[1] written in IPA phonetic transcription [11]

with vowel insertion, to tackle missing syllables, was found to be unacceptable, in preliminary experiments. In addition to this, the TTS system has to cater to Telugu speakers with various degrees of fluency in other languages. Informal experiments also showed that Telugu speakers fluent in other languages, like English, from which words are borrowed, do not prefer vowel insertions in the consonant clusters of those words. In order to deal with these issues, we propose the insertion of a reduced[2] form of a vowel, spanning just a few pitch cycles, in place of the complete vowel. We claim that the proposed method can be used to simplify all kinds of consonant clusters in Telugu. The experiments below are designed to test this hypothesis.

The first experiment was designed to understand if the vowel insertions, of a few pitch cycles, were perceptually acceptable in various kinds of consonant clusters. A set of 12 words containing biconsonantal clusters, both frequent and infrequent, were selected. This set also includes borrowed words. The bi-consonantal clusters contain transitions between consonants of various classes (glides, fricatives, obstruents, lateral approximants, nasals, affricates). These words were synthesized by performing a vowel insertion in the consonant cluster,according to the previously mentioned rule-set (Tables 1 and 2). After synthesis of these words, the number of pitch cycles in the epenthetic vowels were reduced, using pitch-synchronous overlap add (PSOLA) method. Three samples containing 4, 2 and 0 pitch cycles of the epenthetic vowel were generated manually from each synthesized word. These three cases represent various degrees of vowel reduction[2].

Subjects were asked to rate the samples generated, for naturality and intelligibility, using the mean opinion score (MOS) scale. The results of the subjective study are summarized in Table 3. The MOS of the frequent category supports the hypothesis that insertions of even 4 cycles of the epenthetic vowels, do not lead to unacceptable synthesis due to perception of the extra vowel in well recognized and highly frequent consonant clusters. As it can be seen from Table 3, 0 cycle insertions are the most preferred category, with high MOS scores. 0 cycle insertions are equivalent to picking a smaller sub-word unit from the immediate context of the epenthetic vowel, which ensures proper coarticulation within in the resultant syllable. e.g., Syllable [kloo] when missing is effectively being replaced with the smaller sub-word unit [k] having right context of [u] and syllable [loo], when the epenthetic vowel [u] is reduced to 0 pitch cycles.

Usually TTS systems use syllable units with boundaries derived from automatic segmentors. These units have considerable segmentation errors. Due to these erroneous boundaries, it might not be possible to ensure reduction to 0 pitch cycles during the PSOLA based reduction of the epenthetic vowel. This leads to unintended insertion of a few pitch cycles of the vowel. Hence in the rest of the paper we analyse the performance of the proposed method in the worst case scenario, by using 4 pitch cycle insertions of the epenthetic vowel.

Table 3. MOS scores† for various degrees of reduction in the epenthetic vowel

Word Type	MOS		
	4c*	2c	0c
Infrequent	3.80	4.08	4.29
Frequent	4.09	4.30	4.56

$*$ c = pitch cycles
† No. of trials in each MOS test = 8 subjects X 12 words

Table 4. AB tests† for various degrees of reduction in the epenthetic vowel

	4c* vs 2c	4c vs 0c	2c vs 0c
Prefer 4c	5%	10 %	-
Prefer 2c	17%	-	10%
Prefer 0c	-	30%	20%
No Preference	78%	60%	70%

$*$ c = pitch cycles
† No. of trials in each AB test = 5 subjects X 12 words

AB tests were conducted to see if subjects noticed considerable difference among the samples with 0, 2 and 4 pitch cycles of the epenthetic vowel, in the words with bi-consonantal clusters. The results are tabulated in Table 4. It can be observed that the extent of insertion is perceptually indiscernible in most cases.

The above results show that vowel insertions of few cycles are acceptable in both frequent and infrequent bi-consonantal clusters. Thus the proposed method enables the use of epenthesis as a back-off strategy for missing syllables with both types of clusters.

3.3. Epenthesis in tri and quadra consonantal clusters

In the previous experiment perceptual acceptability of the vowel insertion in bi-consonantal clusters was tested. However the main application of the backoff method is in simplifying syllables with complex clusters containing 3 or 4 consonants. Hence an experiment was designed to study multiple vowel insertions in words with tri and quadra consonantal clusters i.e., a cluster $C_1C_2C_3V$ is simplified with multiple vowel insertions as $C_1V_{ins}C_2V_{ins}C_3V$ where V_{ins} represents the reduced epenthetic vowel and C represents consonants. Each of the consonant-consonant junctures has four pitch cycles of vowel insertion (to represent worst case, as discussed previously). 10 words with complex clusters were synthesized using the proposed method, after being fitted into a sentence. The clusters in these words are broken by multiple epenthesis of reduced vowels. Subjects were asked to rate the samples generated, for naturality and intelligibility, using the MOS scale. The MOS score averaged over 10 subjects was 4.1. Thus proving the applicability of the method for simplifying even complex clusters.

[2] Please note that the word reduction is not used here in the phonetics sense, and simply means reducing the number of pitch cycles in the vowel.

The results of the above subjective tests, support the hypothesis that vowel insertion can be used to deal with all types of missing syllables, by creating simplified syllables and subsequently doing a PSOLA based reduction of the pitch cycles in the epenthetic vowels.

3.4. Comparison with conventional back-off methods

The proposed back-off method was compared with the conventional back-off strategy of synthesis using smaller subword units (SSUs). An AB test was conducted to identify the preferred method. The test consisted of 10 subjects judging 10 sentences with words containing complex clusters, synthesized by the candidate methods. Results summarized in Table 5 show a preference for the proposed method.

Table 5. Results of AB test † comparing proposed back-off method with SSU back-off technique

Prefer proposed method	59%
Prefer conventional back-off	17%
No Preference	24%

† No. of trials = 10 subjects X 10 sentences

The proposed method is also superior to the unit substitution method in two aspects. Firstly, when epenthesis is used for back-off the new phone sequence with the reduced vowel(s) does not result in a minimal pair with the original phone sequence of the word. Thus reducing the overhead of tracking minimal pairs. Secondly epenthesis is already well researched in several languages and this reduces the effort needed in creating the rule base for the proposed back-off method.

The epenthetis phenomenon can also be exploited in missing diphone synthesis, using half-phones. As mentioned in Section 1, half-phone joins are done at phone boundaries, which are derived automatically. Optimal coupling necessary to make these joins robust to segmentation errors [8], cannot be done during half-phone back-off. Thus half-phone back-off involves the splicing of half-phone units at phone boundaries, with a small portion of the neighbouring phone due to segmentation errors, which in turn leads to traces of undesirable phone(s) insertions at the joins. It can be ensured that these insertions are the epenthetic vowels, by picking the half-phone units with the immediate context of the epenthetic vowel from the inventory, wherever possible. The splicing of such half-phones at the phone boundaries produces a diphone, which would have at most few pitch cycles of the epenthetic vowel at the join. These minor insertions have been shown to be perceptually acceptable to the native speakers.

4. CONCLUSION

In this paper we proposed a rule-based back-off strategy motivated by the vowel epenthesis phenomenon, to synthesize missing syllable and diphone units. Epenthesis of reduced vowels was proposed, to the enable the application of back-off strategy to all missing syllables. The rule-set used in the proposed method is adopted from L2 acquisition research in Telugu, thus making the rule design a minimal effort process. AB tests conducted to compare the performance of the proposed method with conventional back-off methods, showed a preference (59%) for the proposed method. The proposed back-off method is robust to segmentation errors of automatically segmented databases. The proposed method is not limited to just Telugu, as epenthesis phenomenon is widely observed in several other languages. However the rules for identifying the epenthetic phone (usually vowel) vary according to the language.

5. ACKNOWLEDGEMENT

The authors would like to thank Prof. Peri Bhaskararao, for useful discussions and suggestions on this work.

6. REFERENCES

[1] R Clark, K Richmond, and S King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[2] S P Kishore and Alan W Black, "Unit Size in Unit Selection Speech Synthesis," *in Proceedings of Eurospeech*, pp. 1317–1320, 2003.

[3] JPH Van Santen and AL Buchsbaum, "Methods for optimal text selection," *in Proceedings of Eurospeech '97*, pp. 2–5, 1997.

[4] B. Möbius, "Rare events and closed domains: Two delicate concepts in speech synthesis," *International Journal of Speech Technology*, vol. 6, pp. 57–71, 2003.

[5] E. Veera Raghavendra, B. Yegnanarayana, and Kishore Prahallad, "Speech synthesis using approximate matching of syllables," in *Proceedings of IEEE SLT Workshop*, 2008, pp. 37–40.

[6] JA Louw and M Davel, "Halfphones: A Backoff Mechanism for Diphone Unit Selection Synthesis," *in Proceedings of 17th Annual Symposium of the Patt. Recog. Assoc. of South Africa*, vol. 29, 2006.

[7] R.D. Kent and F.D. Minifie, "Coarticulation in recent speech production models," *Journal of Phonetics*, vol. 5, no. 2, pp. 115—-133, 1977.

[8] Alan W. Black and Paul Taylor, "Automatically clustering similar units for unit selection in speech synthesis.," *in Proceedings of Eurospeech*, vol. 2, pp. 601–604, 1997.

[9] Theodora. Bynon, Tim Bowden, and Bill. Bunbury, *Historical linguistics / Theodora Bynon*, Cambridge University Press, Cambridge ; New York :, 1977.

[10] G. Uma Maheshwar Rao, "A nonlinear analysis of syllable structure and vowel harmony in Telugu," *PILC Journal of Dravidic Studies*, , no. 6, pp. 55 – 84, 1996.

[11] Peter Lagedfoged, "Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet," 2000.