International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 21 (2018) pp. 15011-15015
© Research India Publications. http://www.ripublication.com

# Unit Selection to Improve Naturalness in Speech Synthesis

**Dr. K.V.N.Sunitha**[1] **P. Sunitha Devi**[2]

*[1]Principal, BVRIT College of Engineering for women, Bachupally, Hyderabad, Telangana, India.*

*[2]Assistant Professor, Computer Science and Engineering Department,*
*G.Narayanamma Institute of Technology & Science, Shaikpet, Hyderabad, Telangana, India.*

## Abstract

Speech synthesis is a process that can generate human-like speech for any text input to imitate human speakers. The objective of a text to speech system is to convert an arbitrary text into its corresponding speech. Generating speech that is close to natural human speech is always a challenging issue for synthesis systems. Naturalness depends on the kind of available database, and the algorithms that choose the appropriate speech units from the database. The issue lies on what should be the unit of speech to be stored in database. This paper projects the work carried out in identifying most appropriate speech unit towards improving naturalness. As Telugu language is syllabic by nature, analysis is carried out on 5 million words size corpus to identify the required syllable units that can cover the major vocabulary in Telugu language.

**Keywords:** phone, syllable, coverage, speech synthesis.

## INTRODUCTION

The most natural mode of human communication is speech, and is the driving force underlying several significant advances in speech technology. A text-to-speech (TTS) system converts normal language text into speech; this can be achieved by concatenating recorded speech units stored in a database [1]. TTS systems differ in the size of the stored speech units; a system that stores phones or diphones gives the largest output range, but the output speech may not be natural. For natural sounding speech synthesis, it is essential that the text processing component produce an appropriate sequence of orthographic units corresponding to an arbitrary input text [2, 3]. The difficulty of conversion is highly language depended and includes many problems. For English and most of the other languages the conversion is much more complicated. A very large set of different rules and their exceptions is needed to produce correct pronunciation and prosody for synthesized speech. Most of the Indian languages are phonetic in nature [4]; the conversion is quite simple because written text almost corresponds to its pronunciation. Conversion can be divided in three main phases, text preprocessing, creation of linguistic data for correct pronunciation, and the analysis of prosodic features for correct intonation, stress, and duration. Current state-of-art TTS system in English and other well-researched languages use such rich set of linguistic resources such as word-sense disambiguation, morphological analyzer, Part-of-Speech tagging, letter-to-sound rules, syllabification, stress-patterns in one form or the other to build a text processing component of a TTS system. However for minority languages (which are not well researched or do not have enough linguistic resources), it involves several complexities starting from accumulation of text corpora in digital and processable format. Linguistic components are not available in such rich fashion for all languages of the world. In practical world, minority languages including some of the Indian languages do not have that luxury of assuming some or any of the linguistic components.

## FEATURES OF TELUGU LANGUAGE

Telugu is a South-Central Dravidian language predominantly spoken in the South Indian state of Andhra Pradesh and Telangana, where it is an official language. One of the four classical languages of India, Telugu ranks third by the number of native speakers in India (74 million), thirteenth in the Ethnologue list of most-spoken languages worldwide. There are 23 official languages of India, and all of them except English and Urdu share a common phonetic base, i.e., they share a common set of speech sounds. While all of these languages share a common phonetic base, some of the languages such as Hindi, Marathi and Nepali also share a common script known as Devanagari. The property that makes these languages separate can be attributed to the phonotactics in each of these languages rather than the scripts and speech sounds. Phonotactics is the permissible combinations of phones that can co-occur in a language. Telugu language is phonetic in nature, i.e. there is a one to one correspondence between what we write and what we speak. The letter to sound rules required to map Telugu letters to sound is straight forward.

## GNITS Text Corpora

It is important that the chosen text data covers all the common words, phrases and syllables of a language. GNITS Text Corpora is a set of phonetically rich sentences which consists of DoE-CIIL corpus and newspaper articles which is nearly 5 Million word corpus for Telugu language. The text corpus contains sentences of various articles related sociology, history, poetry, and many other areas. All these words are phonetized so that the distribution of basic speech units – phones, biphones, triphone, etc can be analyzed.48 phonemes: 10 vowels, 2 diphthongs and 33 consonants and 3 variations of anuswaras broadly represent the standard Telugu language phones. The number of phonemes, both vowels and consonants in Telugu, is a controversial issue. There are slightly different phonemic systems for Telugu distinguishing the social dialects into standard and non-standard, educated and uneducated, formal and informal, native and non-native. Different notations like WX, IT3 and ROMAN are in use for representing the Telugu text in digital format. The drawbacks in transliteration schemes include

i) More than one notation for representing the same sound. For example, we may use either 'aa' or 'A' for representing long form of 'a'

ii) Not easily predictable notation like w for ‘t’ and ‘x’ for ‘d’

To avoid the drawbacks mentioned in the above notations, different codes are used where ever there is a possibility of such confusion.The text transcribed using these notations may not be straight forward to read, to overcome this problem KNS notation is used [5].

## SYLLABIFICATION OF CORPUS

Naturalness in synthetic speech generated by concatenative speech synthesis increases when the number of concatenation points required creating a waveform is minimal. Concatenation performed using the syllable like units results in minimal number of junctures and discontinuity effects across the waveform being generated. Since Indian languages already have a well-defined syllable structure, choosing such a unit increases the quality of synthesis [6].

### Syllabification Rules

There is almost one to one correspondence between what is written and what is spoken in Indian Languages. Each character in Indian language script has a correspondence to a sound of that language. In Indian languages, a consonant character is inherently bound with the vowel sound /a/, and is always pronounced with this vowel [7]. In some occurrences this vowel is not pronounced, and this is referred to as Inherent Vowel Suppression (IVS). This occurs at both word final and word middle positions. While letter to phone rules are straight forward in Indian languages, the syllabification rules are not trivial. There is a need to come up with some rules to break the word into syllables. Syllables can be broadly classified into four classes based on the number of phones they contain i.e. single phone (only vowels - V), biphone (CV or VC), triphone (CVC or CCV), quadphone (CCVC) syllables [8].

We have derived certain simplistic rules for syllabification i.e. rules for grouping clusters of (C+V)* based on heuristic analysis of several words in Telugu language. The rules used for syllabification of GNITS text corpus are listed as follows

1. V : a single vowel can exist as a syllable
2. VV: two consecutive vowels are split into V – V, however in Telugu language two consecutive vowels never occur.
3. VCV: is split into V – CV
4. VCCV: is split into VC – CV
5. VCCCV: is split into VC – CCV, the first vowel is associated with the left consonant and the remaining consonants are associated with the right vowel.

   The syllabification process is explained in the following example

గట్టుపైన

gaTTupaina

| | |
|---|---|
| C***VCCV***CVCV | Rule 4 is applied |
| CVC-C***VCV***CV | Rule 3 is applied |
| CVC-CV-C***VCV*** | Rule 3 is applied |
| CVC-CV-CV-CV | |

gaT-Tu-pai-na

గట్-టు-పై-న

## PROPOSED MODEL FOR IDENTIFYING SYLLABLES

Creating a speech database for a text to speech system requires identifying an optimal set of textual sentences to be recorded from native speakers of the language. These sentences should be minimum in number to save recording effort and should have enough number of occurrences of each type of sound units to cover all the co-articulation effects. Number of syllables in a language is not limited to a finite number; it depends on various aspects like number of phones, for a limited domain or unrestricted words, dialects and context. In this section we explain the method to identify the number of syllables required using the method as given in figure 1.

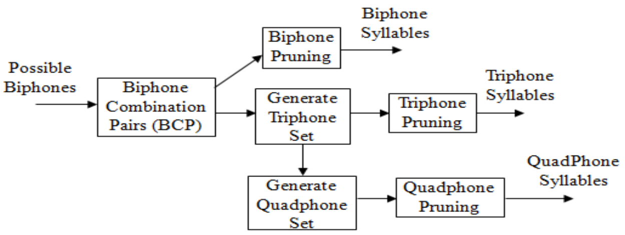
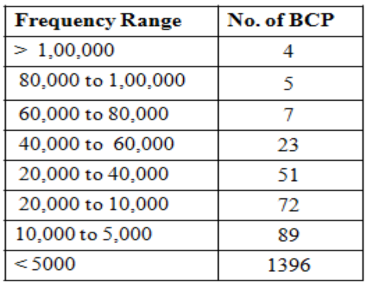


Fig 1: Flowchart for Proposed Model

**Single Phone Syllables**: Vowels are the only single phone units which can be directly considered as syllables. Consonants alone cannot be considered as a syllable, they occur in combination with vowels and consonants.

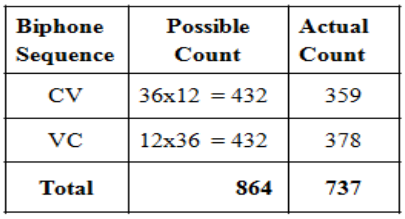
**Biphone Combination Pairs (BCP):** Possible biphone combinations that can be generated is 2304 (48 X 48) in pairs of <CV>, <VC>, <CC>, <VV>. The frequencies of these biphone combinations are calculated and it is found that 1647 such Biphone Combination Pair exist with a huge difference in their frequencies as shown in Table 1.



| Frequency Range | No. of BCP |
|---|---|
| > 1,00,000 | 4 |
| 80,000 to 1,00,000 | 5 |
| 60,000 to 80,000 | 7 |
| 40,000 to 60,000 | 23 |
| 20,000 to 40,000 | 51 |
| 20,000 to 10,000 | 72 |
| 10,000 to 5,000 | 89 |
| <5000 | 1396 |

Table 1: Frequencies of BCP

**Biphone Pruning**: We can directly discard <VV>and <CC> pairs as they do not appear as syllables in Telugu language and few <CV> and <VC> pairs also occur very rarely and are least significant, only 737 such biphone syllables occur as given in Table 2.



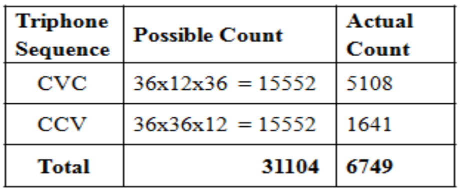
| Biphone Sequence | Possible Count | Actual Count |
|---|---|---|
| CV | 36x12 = 432 | 359 |
| VC | 12x36 = 432 | 378 |
| **Total** | **864** | **737** |

Table 2: Biphone Syllables

**Generate Triphone Set:** The triphone set consists of triplets <CVC> and <CCV> which form syllables of the language. The number of triphones that are formed are huge in number but most of them are not valid. To avoid generation of unwanted triplets the following steps are used

1. From the list of BCP select all <CC> and <CV> pairs.
2. Create triplets <CCV> and <CVC> if *C-V* and *V-C* pair exists in BCP i.e for each <CC> pair include <CCV> if *C-V* exists in BCP and similarly for each <CV> pair include <CVC> if *V-C* exists in BCP.

**Triphone Pruning:** All the generated triplets occur as syllables in Telugu language, in Table 3 we can see that most of the triphones are unwanted and are discarded after applying pruning.



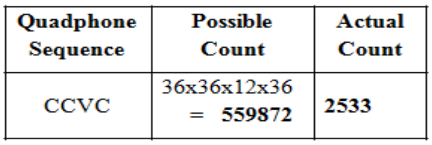
| Triphone Sequence | Possible Count | Actual Count |
|---|---|---|
| CVC | 36x12x36 = 15552 | 5108 |
| CCV | 36x36x12 = 15552 | 1641 |
| **Total** | **31104** | **6749** |

Table 3: Triphone Syllables

**Generate Quadphone Set:** The quadphone set consists of <CCVC> which form syllables of the language and this set can be generated using the following steps

1. From the list of BCP select all <CC> pairs.
2. For each selected <CC> pair create triplet <CCV> if *C-V* pair exists in BCP.
3. For each <CCV> include the quadphone <CCVC> if *V-C* pair exists in BCP.

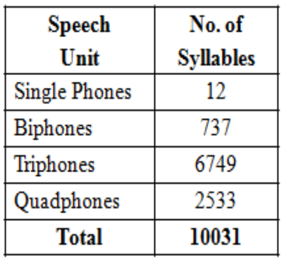
**Quadphone Pruning:** Most of the sequences do not occur as syllables in Telugu language. Apply pruning to discard the unwanted pairs and only 2533 quadphones occur as syllables as given in Table 4.



| Quadphone Sequence | Possible Count | Actual Count |
|---|---|---|
| CCVC | 36x36x12x36 = **559872** | **2533** |

Table 4 : Quadphone Syllables

Corpus design is a critical concern for building rich annotated corpora used for TTS systems, which require huge amount of speech data to train data driven models to produce synthetic speech. Data collection involves cost for recording, annotating etc., and as more data is collected the cost increases. The method proposed in this paper identifies the syllables of each phone unit type required to be stored as speech units in the training set and are listed in the Table 5.



| Speech Unit | No. of Syllables |
|---|---|
| Single Phones | 12 |
| Biphones | 737 |
| Triphones | 6749 |
| Quadphones | 2533 |
| **Total** | **10031** |

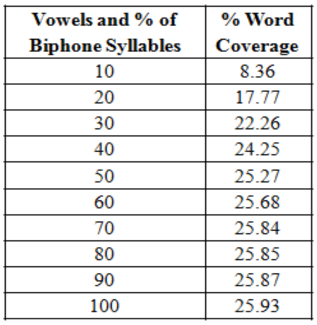
Table 5: Total Syllables

## WORD COVERAGE ANALYSIS

Telugu text corpus of 5 million words which contains sentences that are phonetically rich and various articles related sociology, history, poetry, and many other areas. All these words are converted into KNS notation and are syllabified according to the syllabification rules. The coverage analysis is in two ways i) vowels and variation of biphone units and ii) combination of all speech units.

### Vowels and Biphone Units

For coverage analysis all the vowels and variation in biphone syllable units (10 % to 100%) is taken and the number of words covered in the total corpus is given in Table 6 and its corresponding graph is given in figure 2.



| Vowels and % of Biphone Syllables | % Word Coverage |
|---|---|
| 10 | 8.36 |
| 20 | 17.77 |
| 30 | 22.26 |
| 40 | 24.25 |
| 50 | 25.27 |
| 60 | 25.68 |
| 70 | 25.84 |
| 80 | 25.85 |
| 90 | 25.87 |
| 100 | 25.93 |

Table 6 : Word Coverage with variation in % of Biphone Syllables

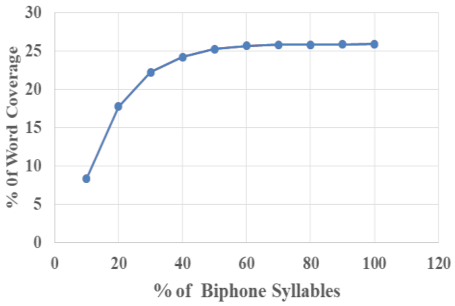


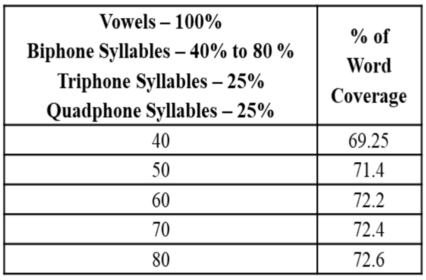
Fig 2 : Word Coverage with variation of Biphone Syllables

The following observations can be made regarding the coverage of speech corpus using the biphone syllable units.

- The set of all vowels and biphone syllables covers 25.93% of the total corpus.
- Top 50% of the biphone syllables along with vowels is covering 25.27% of the total corpus, this shows that syllables below 50% are of less importance.

**Combination of different units:** Analysis is carried out in different combinations to identify the most promising syllables.

a) A set of speech units comprising of vowels -100%, biphone syllables ranging from – 40% to 80%, triphone syllables – 25% and quadphone syllables – 25% is taken and word coverage is given in Table 7 and its corresponding graph is given in figure 3.



| Vowels – 100%<br>Biphone Syllables – 40% to 80 %<br>Triphone Syllables – 25%<br>Quadphone Syllables – 25% | % of Word Coverage |
|---|---|
| 40 | 69.25 |
| 50 | 71.4 |
| 60 | 72.2 |
| 70 | 72.4 |
| 80 | 72.6 |

Table 7 : Word Coverage with variation of Syllables

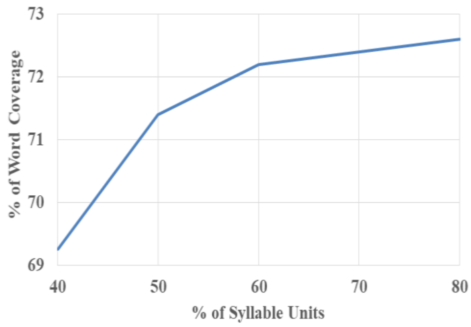


Fig 3 : Word Coverage with Variation of Syllables

Observations:

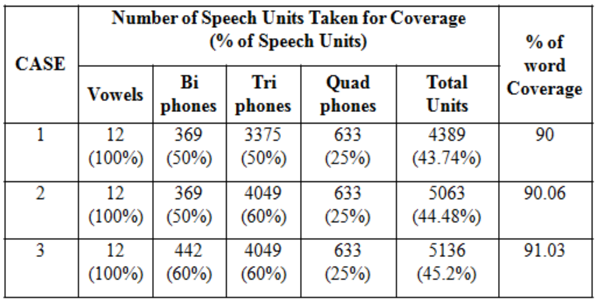
- From the Table 7 we can observe that inclusion of biphone syllables below top 50% is not increasing the word coverage.
- The top 25% of triphones and quadphones are contributing more to the coverage.
- The variation in the percentage of biphone units is not making a prominent change in coverage; we can say that the top 50% of biphones are promising.

b) Analysis is carried out for different combinations to demonstrate how the variation in speech units is affecting the word coverage. Word coverage analysis for the 3 cases is given in the Table 8.

➢ Case 1: Set contains all the vowels, top 50% of biphone, 50% of triphones and 25% of quadphones.

➢ Case 2: Set contains all the vowels, top 50% of biphone, 60% of triphones and 25% of quadphones.

➢ Case 3: Set contains all the vowels, top 60% of biphone, 60% of triphones and 25% of quadphones.



| CASE | Number of Speech Units Taken for Coverage (% of Speech Units) | | | | | % of word Coverage |
|---|---|---|---|---|---|---|
| | Vowels | Bi phones | Tri phones | Quad phones | Total Units | |
| 1 | 12<br>(100%) | 369<br>(50%) | 3375<br>(50%) | 633<br>(25%) | 4389<br>(43.74%) | 90 |
| 2 | 12<br>(100%) | 369<br>(50%) | 4049<br>(60%) | 633<br>(25%) | 5063<br>(44.48%) | 90.06 |
| 3 | 12<br>(100%) | 442<br>(60%) | 4049<br>(60%) | 633<br>(25%) | 5136<br>(45.2%) | 91.03 |

Table 8 : Word Coverage for all 3 Cases

**Observations:**

- It is observed that triphone syllables play a major role as they are present in most of the words.
- Only vowels and biphone syllables coverage is 25.93% but when top 50% triphone and 25% quadphone syllables are taken the coverage is 90% as shown in case - 1.

- In case - 2 when the top 60% triphone syllables are taken the coverage percentage did not change much this shows that the top 50% triphone syllables are prominent.
- In case - 3 the top 60% of biphone and triphone syllables are taken to analyze the coverage percentage, but the results were not so impressive.

## CONCLUSIONS

Naturalness in concatenative TTS systems depends on the type of speech unit stored and the length of each unit. Concatenation using syllables as unit has low discontinuity and the more the length of each unit minimizes the number of concatenation points. Experimental analysis carried out on 5 million word corpus to identify the syllables required to be stored as speech units. From the analysis it is observed that triphone syllables contribute more towards naturalness and are prominent in word coverage.

## REFERENCES

[1]. X.Huang, A. Acero and H.W.Hon, “ Spoken Languages Processing A Guide to Theory, Algorithm and System Development”, New Jersy, Prentice Hall, 2001.

[2] Speech communications Human & Machine by Douglas O’Shaughnessy

[3] Speech and Language Processing by Daniel Jurafsky & James H.Martin

[4] Peri Bhaskara Rao, “Salient phonetic features of Indian languages in speech technology” Sadhana vol. 36, part 5, October 2011, pp. 587–599 © Indian Academy of Sciences

[5]. K.V.N. Sunitha and A.Sharada , “KNS Phoneme Set – A new Telugu Phoneme Set for Telugu Speech Processing Technology” ICRTC 2013 (ISBN No.:978-93-80965-65-9) 4-5 Oct 2013, SRM University, NCR Campus

[6]. M. Nageshwara Rao,S. Thomas, T. Nagarajan and Hema A. Murthy Text-to-speech synthesis using syllable like units, proceedings of National Conference on Communication (NCC) 2005, pp. 227-280, IIT Kharagpur, India, Jan 2005.

[7]. K.V.N.Sunitha and P.Sunitha Devi, “Text Normalization for Telugu Text to Speech synthesis”, in the proceedings of International Journal of computers and Technology, 2013, pages 2241 – 2249.

[8]. K.V.N.Sunitha ,N.Kalyani and N.Sreekanth, “Minimum Data Set Based on Syllable Position for Telugu Speech Systems” in the proceedings of International Journal of Advanced Research in Computer Science and Software Engineering IJARCSSE Vol 3,Issue 10, Oct 2013.