International Journal of Computer and Communication Technology

Volume 2 | Issue 1 Article 5

January 2011

Vowel classification based approach for Telugu Text-to-Speech System using symbol concatenation

Pamela Chaudhur
Department of CSE, ITER SOA University Bhubaneswar, India, pamela.chaudhury@gmail.com

K Vinod Kumar
Department of CSE, ITER SOA University Bhubaneswar, India, kvinod2208@gmail.com

Recommended Citation

Chaudhur, Pamela and Kumar, K Vinod (2011) "Vowel classification based approach for Telugu Text-to-Speech System using symbol concatenation," *International Journal of Computer and Communication Technology*: Vol. 2 : Iss. 1 , Article 5.
DOI: 10.47893/IJCCT.2011.1065
Available at: https://www.interscience.in/ijcct/vol2/iss1/5

Vowel classification based approach for Telugu Text-to-Speech System using symbol concatenation

Pamela Chaudhur [1], K Vinod Kumar [2]
Department of CSE, ITER
SOA University
Bhubaneswar, India
Email: pamela.chaudhury@gmail.com [1], kvinod2208@gmail.com [2]

Abstract*— *Telugu is one of the oldest languages in India. This paper describes the development of Telugu Text-to-Speech System (TTS) using vowel classification. Vowels are most important class of sound in most Indian languages. The duration of vowel is longer than consonants and is most significant. Here vowels are categorized as starting middle and end according to the position of occurrence in a word. The algorithm developed by us involves analysis of a sentence in terms of words and then symbols involving combination of pure consonants and vowels. Wave files are being merged as per the requirement to generate the modified consonants influenced by deergalu (vowel sign) and yuktaksharas generate the speech from a text. Speech unit database consisting of vowels (starting, middle and end) and consonants is developed. We evaluated our TTS using Mean Opinion Score (MOS) for intelligibility and voice quality with and without using vowel classification from sixty five listeners, and got better results with vowel classification.

Key words: Indian Standard Code for Information Interchange (IISCI); letter to phoneme mapping; Vowel classification; unit sele ction; symbol concatenation.

I. Introduction

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 [1].

For language such as English a pronunciation dictionary of about 125000 words is used along with a letter to sound rules to handle unseen words. Indian languages are phonetic in nature [2], therefore the letter to sound rule is relatively easy. For Telugu there is good correspondence between written text and spoken language. However for some Indian languages such as Hindi, Oriya and Bengali the rules for mapping the letter to phoneme are not so straightforward. Developing Telugu TTS is easier than other Indian languages like Hindi, Oriya or Bengali because Telugu doesn't require Schwa deletion [3]. The inherent "a" associated with a consonant is suppressed depending upon the context in which it is used (like kamala is pronounced as kamal in Hindi). The vowel "a" which inherently occurs in all consonants is called *schwa***.**

II. Language Processing Unit

The objective of the language processing unit or the text processing unit is to process the given input text and produce appropriate sequence of phonemic units. These phonemic units are realized by the speech generation component by selection of a unit from a large speech corpus. For natural sounding speech synthesis, it is essential that the language processing unit produce an appropriate sequence of phonemic units corresponding to an arbitrary input text.

A. *Text-to-phoneme conversion*

Generation of sequence of phonetic units for a given standard word is referred to as letter to phoneme rule or text to phoneme rule. The complexity of these rules and their derivation depends upon the nature of the language. In our Telugu TTS the input is Telugu text in Indian Standard Code for Information Interchange (ISCII). This may be typed in through an ISCII key board or input from a pre stored ISCII file. As an example, the sequence of ISCII codes:

204 199 204 162 194 32 203 165 207 194 167 162 168 209 162

Corresponding to the input text -
(Manamanta bharateeyulam)
(We all are Indians)

మనమంతా భారతీయులం

is parsed into the following sequence of basic units:

Ma, na, ma, n, ta, blank, bha, ra, tee, yu, la, m .

The basic units of writing system in Indian languages are Aksharas. They are the orthographic representation of speech sounds. For Telugu also we define character set into vowels and consonants. Vowels play a major role in pronunciation of any word. All together we have 14 vowels and 36 consonants. The vowel signs are known as deergalu. Vowels are most interesting class of sound in any language. Their duration in any word is also of most significance. Indian Languages originated from Sanskrit. The synthesis system of vowels is narrated in Vedas. They play a major role in the pronunciation of any word. Each of the vowels are classified as starting, middle and end according to the duration of occurrence in a word.

/a/	అ	short vowel
/aa/	ఆ	long vowel
/i/	ఇ	short vowel
/ee/	ఈ	long vowel
/u/	ఉ	short vowel
/oo/	ఊ	long vowel
/e/	ఎ	short vowel
/ae/	ఏ	long vowel
/ai/	ఐ	short vowel
/o/	ఒ	short vowel
/oa/	ఓ	long vowel

Telugu consists of consonants from /k/ to /h/.Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes they are more difficult to synthesize properly.

III. Signal Processing Unit

Given the sequence of phones the objective of the signal processing unit is to synthesize the acoustic waveform. While the articulatory model suffers from adequate modeling of motions of the articulators, the parametric models require a large number of rules to manifest co articulation and prosody. An alternative solution was to concatenate pre recorded speech segments [4]. For the development of the text-to-speech synthesizer for Telugu we have used concatenation of pre recorded speech units. Connecting prerecorded natural utterances is one of the easiest ways to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods. Current state of art speech synthesizers generate natural sounding speech by using an inventory of large number of speech units. Storage of large number of units and their retrieval in real time is feasible due to cheap memory and computational power.

A. *Concatenation Technique*

One of the most important aspects in concatenative synthesis is to find correct speech unit length. The selection is usually a trade-off between longer and shorter units. With longer units high naturalness, less concatenation points and good control of co articulation are achieved, but the amount of required units increases, there by speech unit database increases. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. With shorter units naturalness reduces, concatenation points increases, and coarticulation is not achieved. In present TTS systems units used are usually words, syllables, phonemes, diphones, and sometimes even triphones.

The approach of using an inventory of speech units is referred to as unit selection approach. We have selected the basic unit for speech as phones[6] (phonemes, diphones and triphones**)** for the Telugu text-to-speech synthesizer. The signal processing unit concatenates at the symbol level. A symbol can be C, V, CV, CCV, VC, VCC, CVV where C is a consonant and V is a vowel. If the symbol is a C or V then a phoneme is concatenated. If C is associated with a single vowel sign then a diphone is selected for concatenation. If the C is associated with a two vowel signs then a triphone is selected for concatenation. In symbol based concatenation conjuncts can be easily uttered and understood by the listeners. Wave files are being merged as per the requirement to generate the modified consonants influenced by deergaalu (vowel signs), and yuktaksharas generate the speech from a text. Pronunciation of tetraphones and pentaphones is yet not possible through synthesizer. Also vowels are categorized into three types depending upon their duration in the word. The vowels classifications are starting, middle and end. During concatenation stage particular vowels are selected from the speech unit database at run time. This results in a better speech quality than using uncategorized vowels.

B. *Generation of Speech Unit Database*

The speech unit database is of primary importance for a good text to speech system. The speech database primarily depends upon the selection of the utterances which has the coverage of all possible units and the recording of the utterances by a good voice talent. Selection of the 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. The natural speech must be recorded so that all used units (phonemes) within all possible contexts (allophones) are included. After this, the units must be labeled or segmented from spoken speech data, and finally, the most appropriate units must be chosen. Gathering the samples from natural speech is very time-consuming.

Acoustic quality is maintained by recording voice in noise free environment. The recorded data is then digitized using the Wavsurfer software. The present system employs 16-bit data while the Sampling Rate is taken as 22.05 KHz. (Approximately 20 KHz sampling rate is good enough for maintaining voice quality). Recording of each phoneme is done at a sampling rate of 22.05 kHz. The phoneme database consists of all the consonants and vowels.

I. VOWEL CLASSIFICATION

Vowels are most important class of sound in most Indian languages. Vowels are longer in duration than consonant sounds. Our hypothesis is if vowel sounds can be synthesized perfectly by machine then sound quality achieved would be better [5]. So we have categorized Vowels as starting, middle & end according to the position of occurrence in a word [7]. As the vowels are dominant in the utterance, they are stored for different durations as they occur in the word. Each of vowels are recorded and segmented into starting, middle and end parts .For the Purpose we record each vowel at 22.05 kHz and then segment into three parts. Segmentation is done in such a way that each segment not only represents the vowel but also defines if the vowel belongs to starting, middle or end of a word. There is a distinctive difference between the same vowel when it occurs in the starting, middle and end of a word. So concatenating the right segment of the vowel definitely improves the quality of speech.

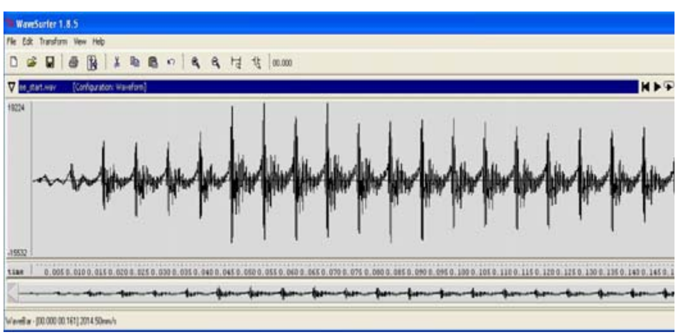

Fig 1: Waveform of vowel ‘ee’ at start of a word like in the word *Vinod*

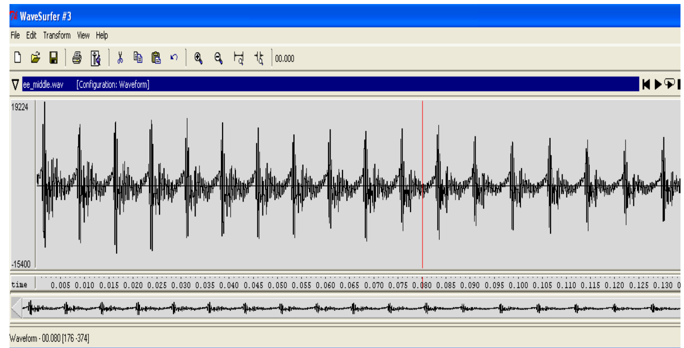

Fig 2: Waveform of vowel ‘ee’ at the middle of a word like in the word *kavita*

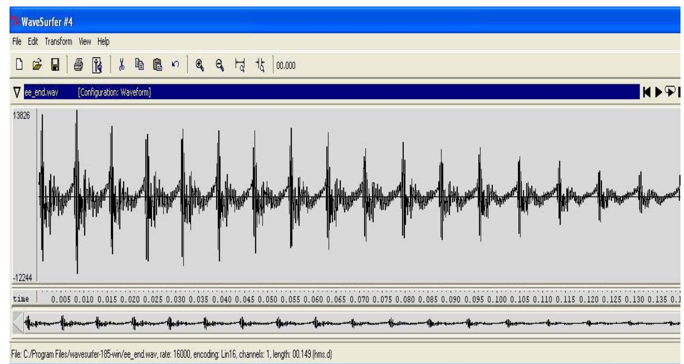

Fig 3: Waveform of vowel ‘ee’ at the end of a word like in the word *kavi*

So the speech unit database consists of three sets of each vowel and three sets of each vowel sign. Each vowel and vowel sign has a specific naming convention. The naming convention helps to retrieve the exact segment of the vowel at run time and there by resulting in a perfect concatenation.

II. EVALUATION

The TTS developed has a good voice quality. We evaluated the TTS using Mean Opinion Score .We evaluated our TTS for intelligibility and voice quality with and without using vowel classification from sixty five listeners. The first two charts show the performance without the vowel classification and the last two charts with the vowel classification.

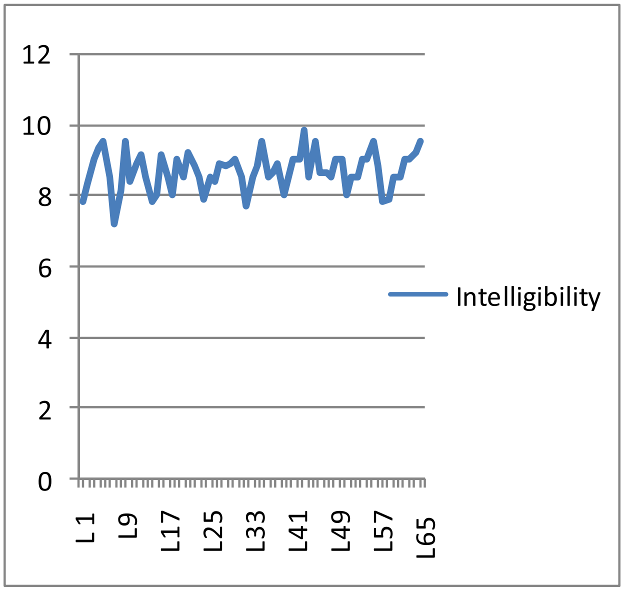

Fig 4: MOS of TTS intelligibility without vowel classification

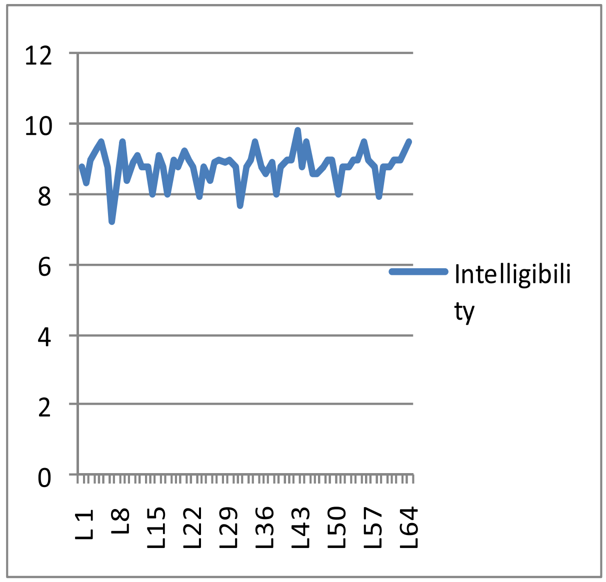

Fig 6: MOS of TTS intelligibility with vowel classification

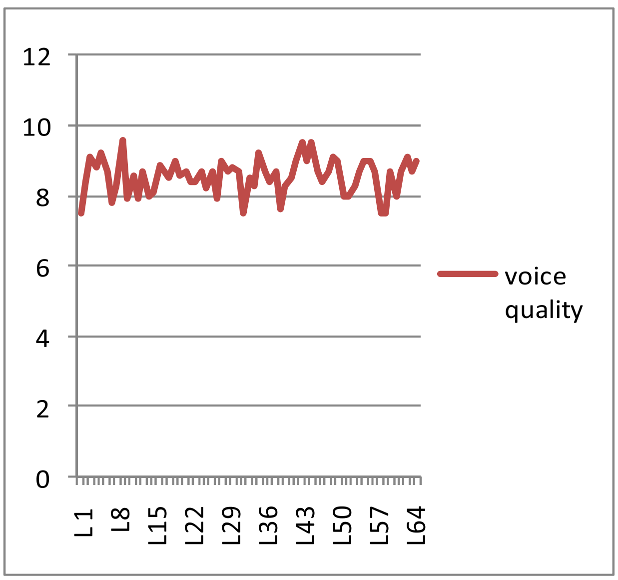

Fig 5: MOS of TTS voice quality without vowel classification

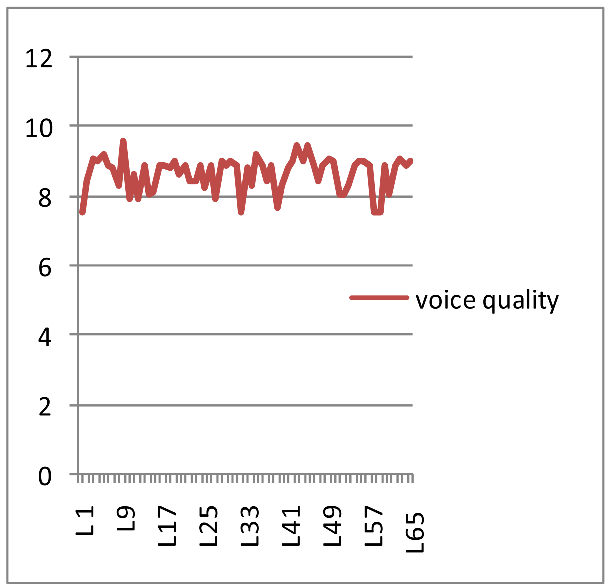

Fig 7: MOS of TTS voice quality with vowel classification

The MOS for intelligibility is 8.67 and for voice quality is 8.55 on 10 point scale without vowel classification.

MOS result for intelligibility is 8.80 and for voice quality is 8.64 on 10 point scale with vowel classification.

III. Conclusion And Future Work

The Text- to-speech synthesizer for Telugu has been successfully designed and developed using concatenation of prerecorded speech units. Classification of vowels based on position is relatively simple technique and it improves the quality of speech to a great extent. Symbol based concatenation is better than concatenation techniques like diphone based concatenation because the speech unit database is smaller thereby reducing the memory requirements. Also the conjuncts can be better uttered. Vowel classification aims at synthesizing vowels in a near to perfect way depending upon their position in a word. The beauty of the TTS is that with very minor changes in the software the Text to Speech system can be developed for most of the Indian languages. Generated speech is without prosody. Adding prosody is the subject of our future work.

References

[1] Anand Arokia Raj, Tanuja Sarkar, Satish Chandra Pammi, Santhosh Yuvaraj, Mohit Bansal, Kishore Prahallad, Alan W Black "Text Processing for Text-to-Speech Systems in Indian Languages", 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007.

[2] Anil Kumar Singh "A Computational Phonetic Model for Indian Language Scripts" online proceedings of Constraints on Spelling Changes: Fifth International Workshop on Writing Systems. Nijmegen, The Netherlands, October, 2006.

[3] Monojit Choudhury, "Rule Based Grapheme to Phoneme Mapping for Hindi Speech Synthesis" 90th Indian Science Congress of ISCA, (Abstract published), Bangalore, India, 2003.

[4] Anupam Basu, Debasish Sen , Shiraj Sen and Soumen Chakraborty "An Indian Language Speech Synthesizer –Techniques and Applications" National Systems Conference, Indian Institute of Technology, Kharagpur, december 17-19, 2003

[5] Susan Choge, M.Phil, "Understanding Kiswahili Vowels". The Journal of Pan African Studies, vol.2, no.8, March 2009

[6] Hunt A.J. and Black A.W., "Unit selection in a concatenative speech synthesis system for a large speech database," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing,, pp. 373–376,1996.

[7] Carlson, R., & Nord, L. "Vowel dynamics in a text-to-speech system - some considerations". In Proceedings Eurospeech '93 (pp. 1911-1914). Berlin, 1993.