Text Normalization and Unit Selection for a Memory Based Non Uniform Unit Selection TTS in Malayalam

Gokul P., Neethu Thomas, Crisil Thomas and Dr. Deepa P. Gopinath
College of Engineering Trivandrum
Trivandrum - 16
gokulpramesh@gmail.com, neetuthomas259@gmail.com
crisil971@gmail.com, deepapgopinath@gmail.com

Abstract

Text to speech synthesis system intended for any language, converts the given text in that language to corresponding speech. The major challenge in TTS system is to generate artificial speech which appears to be natural and intelligible. This is essential for visually impaired people to properly understand and comprehend the generated speech. This paper discuss about text normalization and unit selection for a memory based non-uniform unit selection concatenative speech synthesizer for Malayalam language.

1 Introduction

Text to speech synthesis systems(TTS) help users to interact with computer through speech. System with speech interaction is advantageous for physically challenged, especially visually impaired. Important challenge in speech synthesis is generating synthesized speech which is both intelligible and natural.

TTS systems synthesize speech by articulatory synthesis, formant synthesis or concatenative speech synthesis techniques, of which concatenative speech synthesis excels in performance. In concatenative speech synthesis, segments of speech waveform that are cut from recorded speech and stored in an inventory are concatenated to generate synthesized speech. Non-uniform unit selection involves selection of appropriate units of variable length for synthesizing speech. Non-uniform unit selection is the most popular technique in the current scenario. The selection of appropriate speech waveform for concatenation is the major concern within concatenative synthesis.

TTS systems attempts to synthesize speech which has the qualities of natural speech generated by humans. The best way to improve the quality of synthesized speech is to mimic the way humans generate speech. The process of mimicking human functionality requires a model of how humans store the language within their brain and how they retrieve appropriate units for speech production. A memory based model for Malayalam TTS can be developed based on Memory Prediction Framework, which is a theory of brain function.

This work attempts to implement the front end portions for a memory based Malayalam TTS. The TTS system deals with real world data, hence text preprocessing is an important challenge. The system attempts to mimic the functionality of human brain in generating speech. The system develops a memory model resembling memory organization of linguistic units within brain. The memory based model can be used in unit selection block for non-uniform unit selection concatenative speech synthesis.

Non-uniform unit selection concatenative speech synthesis is not reported for Malayalam language. The use of memory prediction framework for speech synthesis is a novel method in the realm of speech synthesis.

2 Text Normalization

A text to speech synthesis system requires handling of text input from various discourses. Preprocessing is required to convert non standard words of the language into units suitable for speech production. Numbers, abbreviations, acronyms, dates, phone numbers, etc. are examples of non standard words. Thus, one of the main components of a TTS system is the preprocessing part which does text normalization that transforms

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 172–177,
Trivandrum, India. December 2015.

non standard text elements into their expanded form. This requires both linguistic and technical knowledge.

First step in text normalization module is input text tokenization, here the input text is converted into tokens based on the space between words. These tokens are then identified as standard or non standard words (NSW). An inventory of the non standard words must be stored to identify NSW in the given text. NSW includes acronyms, abbreviations, numbers, dates, currency, measurement units, etc. Examples are *12/5/15*, *Dr. prakash*, *12.5 c. m.*, etc. These identified NSW are expanded to standard form. Period at the sentence boundary is identified.

The preprocessed text is used for selecting non uniform units from the database using memory based hierarchical model. The linguistic structure is unique for each language. Hence preprocessing module for Malayalam TTS need to be developed after analysing the linguistic features of language.

3 Memory based model

The memory model is based on Memory Prediction Framework (Hawkins and Blakeslee(2004)). Memory Prediction Framework is a theory of brain function. The organization of language possess a hierarchical tree structure with sentence at top node and syllable at lower node. Letters are combined to form syllables. Syllables are combined to form words. Words are combined with morphemes to form new word. Words and morphemes combine to form sentences or clauses.

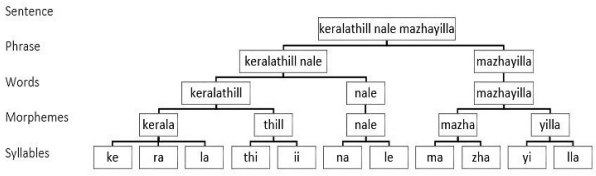

Figure 1: Hierarchical structure in language

Memory model exploits all properties of memory organization of language within brain. The model attempts to build a hierarchical text database that contains tokens of linguistic units at various levels from sentence to syllables. The unit selection algorithm selects the appropriate unit for concatenation from memory model for concatenative speech synthesis. The development of memory model requires a thorough study of morphological and syllable structure of Malayalam language.

4 Morphological and Syllable structure of Malayalam

Morphological variations for words occur in Malayalam due to Inflections, Derivations and Word compounding(Shanavas(2015)).

As a general rule, the syllable structure in Malayalam can be schematized as (C)(C)(C)V(V)(C), where parenthesis indicates optionality(Mohanan(1989)). C stands for consonants and V for vowel. The basic syllable structure in Malyalam consist of a vowel and consonant. The syllabification rules are(Asher(1997))

- All intervocalic single and geminate consonants are assigned to following syllable
- Sequences of two homorganic oral plosives in which first unaspirated and second aspirated are treated in the same way as geminate consonant
- In other sequences of two consonants where both belong to P class(stops, oral, nasal), the two segments are assigned to different syllables.
- Where L(liquids or glides) as first segment followed by P or F(Fricatives), it is assigned to preceding syllable.
- Two-consonant clusters of the type PL,FL,LL are all assigned to following syllable with one exception, where in L_1L_2, L_1 is $/y/, /y/$ goes to preceding syllable.

In phonology, an allophone is one of a set of multiple possible spoken sounds used to pronounce a single phoneme. The allophonic variations has its impact mostly on syllables. If these variations are not considered, the synthesized speech will sound differently from original speech. A set of rules is to developed to account for this variability in Malayalam language.

5 Non- Uniform Unit Selection

Concatenative speech synthesis involves selecting optimal units from database and concatenation of these units to synthesize speech. Non-uniform unit selection involves selecting units which vary in length. The most appropriate units have to selected from the inventory to improve the quality

of synthesized speech. The non-uniform unit selection algorithm formulated in this paper exploits the properties of brain in generating speech.

The unit selection algorithm maintains a top-down approach within the memory model. As we move down the hierarchy, sentences get unfolded into memory of sequences of phrases. In the next layer down, each phrase is unfolded into a memory sequence of words and so on. The best optimal unit will be selected while moving through the hierarchy. The algorithm is prioritized to select the longest unit if present in memory. In case of unavailability of longer units, the algorithm moves down the hierarchy to obtain shorter units.

6 Zipf's Law

The distribution of words in a corpus of any natural language is non uniform. The rank of a word (in terms of its frequency) in a given corpus of natural language utterances is approximately inversely proportional to its actual frequency, and so produces a hyperbolic distribution according to Zipf's law(K.(1949)). If the words of a sample text are ordered by decreasing frequency, the frequency of the kth word P(k), is given by

$$P(k) \propto k^{-\alpha} \qquad 1 < \alpha < 2 \qquad (1)$$

The memory based model replicates the memory organization within brain. Human brain memorizes only frequently repeating linguistic units. Such a system with limited size and efficiency is possible only if Malayalam language has certain units that would repeat more frequently compared with others. This assumption is evaluated using Zipf's law.

7 Methodology

7.1 Text Normalization

Non-standard words are identified and are then categorized into NSW with letters, numbers or combination of both. The NSW are converted to their standard form in the text normalization module.

In Malayalam language, conversion of number to text is not straight forward. For example in English the number "12,500" can be expanded to "twelve thousand and five hundred" by using a simple algorithm, because words like twelve, thousand, five, hundred, etc, are repeated in the text representation of numerals. But in Malayalam this does not occur, the same number is expanded in malayalam as "pantiiraayiratti anjnjuuR+". Thus here for Malayalam TTS complex algorithm is required for number conversion.

NSW with numbers is further classified into various subcategories which includes phone number, cardinal number, date and time. This is because a number "123" is pronounced as "nuuRRi irupatti muunn+", where as a phone number "9961..." is pronounced as "onpatu onpatu aaR+ onn+...". Due to this difference in the output text their classification is necessary, to generate the required output.

Finally acronyms and abbreviations are expanded and period at sentence boundary is identified eliminating all other periods denoting acronyms, abbreviations or ellipsis.

7.2 Identification of Linguistic Constructs for Building Memory model

The frequently repeating sentences, phrases, words, morphemes and syllables has to be identified for development of memory model. Sentences and words in each domain is identified by frequency analysis of text data in that domain. Phrases are obtained by ngram modelling of words.

N-grams are contiguous sequence of n items. The ngram modelling gives an idea of collocations which can be modelled for prediction of next word.

eg: nale mazha peyyum ennu thonnunnu

bigrams:(nale mazha), (mazha peyyum), (peyyum ennu), (ennu thonnunnu)

trigrams: (nale mazha peyyum),(mazha peyyum ennu),(peyyum ennu thonnunnu)

The morphemes are obtained after analysing the morphological structure of Malayalam language. The words containing morphemes are splitted into root word and morpheme. Both root word and morpheme are stored within memory model. In order to generate a new word we could combine root and morpheme from another word, if both are available in memory model. This would avoid the necessity syllable concatenation even if the required word is not present in memory.

The syllables are identified using automatic syllabification algorithm developed based rules for syllable formation in Malayalam(Asher(1997)). During syllabification the allophonic variations are also taken into account.

8 Results and Discussions

8.1 Text Normalization

Acronym expansion, abbreviation identification etc, was tested on the text corpus (including articles, newspapers) consisting of 90,000 sentences. Number to text conversion algorithm for Malayalam was successfully implemented.

A complex algorithm is implemented for number conversions considering the exceptions in case of Malayalam language. Initially, the algorithm classifies, the digits in the number based on their places in the string and separate conversion algorithm is adopted for digit at each place. For first and second place, the conversion is straight forward, even if the number of digit increases (eg : 12 is pronounced as ”pantranTu” in 312 and in 1012), but complexity increases for numbers with higher places. Hence, rules are formulated for higher places considering exceptions, for example, if digit 1 is present at the third place in a four digit number it is pronounced as ”orunnuuRRi” (eg : ”aayiratti orunnuuRRi onn+” for 1101) and as ”nuuRRi” (eg : ”nuuRRi onn+” for 101) when at same place in a three digit number. And for numbers having fifth place the dictionary consisting of exceptions (such as ”patinayyaayiram” for 15000) is used.

Acronym expansion and abbreviation identification was done by storing commonly occurring acronym. Period defining sentence boundary was identified after removing period defining acronym, abbreviation and ellipsis.

8.2 Verification of Possibility of memory model

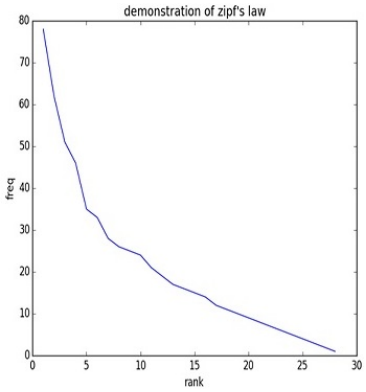

Figure 2: Zipf’s law for Malayalam

The evaluation of Zipf’s law is done over text database obtained from various online sites. Zipf’s law is evaluated independently for each domain(news, literature, education, travel etc..) and for each writer. The hyperbolic graph obtained is found to follow the Zipf’s law.

It is also observed that, not only words but phrases and whole sentences also follow Zipf’s law. The significance of Zipf’s law is that, a major portion of any speech corpus can be represented by the most frequently occurring constructs. Hence a memory model can be developed by storing frequently occurring constructs.

8.3 Development of Memory model

The memory model is developed by analysing 1000 Malayalam sentences taken from a collection of speech data prepared by IIIT-Hyderabad for various Indian languages(Kishore Prahallad(2012)). The development of memory model involved identification of frequently occurring constructs like sentences, phrases, words, morphemes, syllables. The memory model is a 8 layer structure with longer units at top. The model consisted of 10 sentences, 18 trigrams, 70 bigrams, 156 words, 78 morphemes and 670 syllables. The unit selection algorithm best suitable longest available unit for concatenation to synthesize speech.

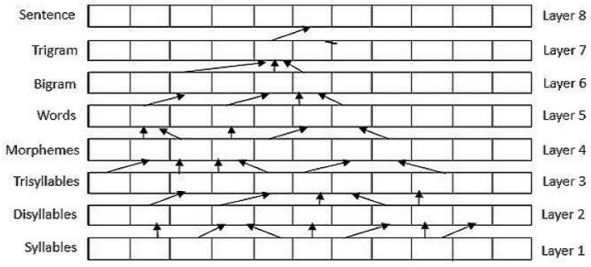

Figure 3: Memory model for Malayalam

9 Letter to sound rules

The rules to account for allophonic variation in Malayalam are[1]:

1. Allophonic variations for vowels
   (a) Case /u/
   If not word initial or word final and if preceding vowel (vowel in the preceding phoneme) is not /u/ then replace /u/

[1]The letter to sound rules were framed in a two day workshop of linguists, language experts, computational linguists and speech processing experts organized by SCERT (State Council for Education Research and Training), at Calicut University as a part of the programme to develop speech enabled systems for visually impaired.

with raised and retracted form of schwa eg:uTuppu

(b) Case /a/

If not in word final syllable and if succeeding consonant is palatal or alveolar or if preceding consonant is voiced stop or /ya/, /ra/, /Ra/, /la/ then replace /a/ with a special form of /a/ which is more similar to /e/ eg:balam

(c) Case /i/

If not word initial or word final then replace /e/ with raised and fronted form of shwa.

2. Palatalization of geminate velar plosives

If preceding phoneme is /i/, /e/ or /ya/ then replace /kka/ with its palatalized version(/k'k'/) eg:adik'k'uka

3. Case alveolar /n/

If not word initial and if preceding consonant or succeeding consonant is dental then replace with dental n eg:sandhi

Exceptions : perunnal, varumnal, malanad, somanadhan, karinizhal,vananira

4. Voicing of intervocalic plosives

If preceding phoneme is nasal sonorant then replace /ka, ca, ta, pa/ with a special form of /ga, ja, da, ba/ replace /Ta/ with /Da/

else if preceding phoneme is nonnasal sonorant and if succeeding phoneme is a vowel then replace /ka, cha, ta, pa/ with a special form of /ga, ja, da, ba/ replace /da/ with /Da/ eg: makan, apakadam

5. Consonant cluster with /h/

When in a consonant cluster with a nasal, /h/ is not pronunced, instead the consonant is geminated.

If succeeding phoneme or succeeding phoneme is nasal then replace /h/ with that nasal eg:brahmam, chihnam

6. Case /w/

(a) If preceding phoneme is a consonant then replace /w/ with a special form, which is labiodental eg:varam

(b) If preceding phoneme is anuswaram then replace /w/ with a special form of /w/ eg:swayamvaram

7. If preceding phoneme is a consonant and is succeeded by /a/ then replace /ya/ with a special form of /e/. eg:vyasanam

8. If not word initial or word final and if succeeding phoneme is not /m/ then replace /t/ with /l/ else if succeeding phoneme is /m/ then replace /t/ with /$l^p m$/ or /$t^m p$/ eg:athmavu

9. Post nasal stops converted to nasals

If not word initial or word final if succeeding phoneme is a consonant then replace it with corresponding nasal sound

10. If preceding phoneme is a consonant and is succeeded by /a/ then replace /ya/ with a special form of /e/ eg: nanni

11. If not word initial or word final and if succeeding phoneme is /k/ vargam then replace it with corresponding nasal eg:bhangi, sangeetham

12. For /n/ if preceding phoneme is /p/ then replace /p/ with a special form /p^t/ eg:swapnam

13. If preceding phoneme is /g/ then replace /p/ with a special form /g^b/. eg:yugmam

9.1 Evaluation

Front end developed consisting of text normalization and unit selection block was incorporated with waveform generation module and listening test was conducted. Listening tests involve preparing several samples of synthesized output from TTS system, randomizing the system sentence combinations and asking listeners to score each output audio. For DMOS evaluation based on semantically unpredictable sentences (SUS), 5 sentences were chosen and the evaluation was done with 6 subjects(Viswanathan and Viswanathan(2005)). The DMOS obtained must be unbiased. To ensure this the same listeners were not asked to participate in different tests. For DMOS we play randomly the natural sentences and synthesized sentences. An original file (sentence x) was played followed by a synthesised sentence (sentence y). All these sentences had different text. Headphones of reasonable quality was used for evaluation. Further, each listener was made to listen to a different set of sentences. The average DMOS score obtained is 3.7.

Sentence	DMOS
111000	3
naale skool avadiyaan+	3.5
aayiraM kollaM pazhakamulla nadiyaan+	4
keiralatile oru jillayaan+ tiruvanantapuraM	4
keiralappiravi navambeR 1	4

Table 1: Evaluation Results

10 Conclusion

Text to speech synthesis system finds its application for visually impaired people. Naturalness and intelligibility of the synthesized speech are necessary for a text to speech synthesis system.

The text processing module was developed for identifying the non standard words in Malayalam and to generate normalized text. Units selection for concatenation was performed using these normalized text.

To make the synthesized speech to appear similar to natural speech, we developed a model that mimics human brain in generating speech. Human brain maintains a hierarchical memory organization of language containing linguistic units. Brain uses non-uniform units stored within memory to generate speech. The memory based model also employs non-uniform unit selection algorithm.

The memory based model is developed by obtaining frequently occurring linguistic units in Malayalam Language. The evaluation of Zipf's law for Malayalam provided base for assumption that a few frequently repeating constructs could represent an entire language.

The model was incorporated to the unit selection block of non-uniform unit selection speech synthesizer. The average mean opinion score of 3.7 shows that the synthesized speech appears to be natural and intelligible. The memory model could be fine tuned by incremental learning by using a large text database.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

R Asher. Malayalam, 1997. Routledge.

Ramon Ferrer Cancho and Ricard V. Sole. Least effort and the origins of scaling in human language. *Proc. National Academy of Sciences of the United States of America*, 100:788–791, 2003.

Jeff Hawkins and Sandra Blakeslee. On intelligence, 2004. Times Books.

Haowen Jiang. Malayalam: a grammatical sketch and a text. *Department of Linguistics, Rice University*, 2010.

R D Johnston. Beyond intelligibility: the performance of text-to-speech synthesizers. *BT Technological Journal*, 1(2):100–111, 1996.

Zipf G. K. *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley, 1949.

Simon King. Degradation mos and word error rate for text to speech synthesis systems. *Private Communication*.

E.Naresh Kumar Kishore Prahallad. The iiit-h indic speech databases. *INTERSPEECH*, 2012.

Tara Mohanan. Syllable structure in malayalam. *Linguistic enquiry*, 20(4):589–625, 1989.

SK Saranya. Morphological analyzer for malayalam verbs. *Unpublished M. Tech Thesis, Amrita School of Engineering, Coimbatore*, 2008.

SA Shanavas. Structure of a computational laxicon of malayalam. 2015.

Catriona Tullo and James R Hurford. Modelling zipfian distributions in language,. *in Proc. of Language Evolution and Computation Work- shop at ESSLLI*, pages 62–75, 2003.

Mahesh Viswanathan and Madhubalan Viswanathan. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos)scale. *Computer Speech and Language*, 19(1):55–83, 2005.